Foreign Language Teachers' Attitudes Toward Written Recall Protocol as a Practice of Reading Comprehension Assessment

Kenneth J. Boyte*
Cabrillo College, Middlebury Institute of International Studies

Abstract

As part of an international effort to develop theory and best practices for teaching languages, the U.S. military has, since the American Revolution, been a leading supporter of literacy education to improve the job performance of soldiers. One important aspect of literacy education today—which continues to be a priority for government agencies, private industry, and public school teachers—involves the development of tools to more accurately measure reading skills. This study highlights an alternative assessment framework known as immediate written recall protocols, currently being used by at least one U.S. government facility dedicated to training military linguists and known for implementing pedagogical innovations. The study explored the beliefs and assessment practices of foreign language teachers at this school regarding their use of traditional item types and immediate written recall protocols, which require students to produce written responses to summarize main ideas and to identify details in texts immediately after reading. Using a questionnaire and a follow-up procedure, this mixed-methods study found that properly trained foreign language instructors believe that immediate written recall protocols are superior to traditional item types because the alternative assessment framework can provide insight into comprehension breakdowns and thus more directly inform corrective instruction.

Keywords: Reading comprehension, immediate written recall protocols, diagnostic assessment

Introduction

Immediate written recall protocols, a diagnostic assessment grounded in the cognitive theory of constructivism (Bartlett, 1932; Spiro, 1980; Spivey, 1989), provide an alternative to traditional measurements of reading comprehension, which have often included multiple-choice, true-false, and cloze-completion item types (Fletcher, 2006; Kamil, 1984; Oller, 1979). Unlike such item types popular in the psychometric tradition of discrete-point tests (Galton, 1879; Goodman, 1968, 1988; Smith, 1971), immediate written recall protocols require students to produce written language to summarize the main ideas and to identify the details of texts immediately following reading. These written responses then can be analyzed to identify information gaps and communication breakdowns, which in turn informs corrective instruction (Bernhardt, 1983, 1991, 2000, 2011). The procedures for using immediate written recall protocols are similar to those used in the first recorded test of reading comprehension reported in 1884. In that experiment of psychology, "Adults read a 10-line paragraph during a fixed time period, after which they wrote down everything they could remember" (Venezsky, 1984, p. 13).

Although defining the unobservable psychological trait of reading comprehension has been difficult and remains elusive today (Perfetti, Landi, & Oakhill, 2005; RAND, 2002; Sterziik & Fraser, 2012), determining best practices for accurately assessing reading skills continues to be a priority for educators (National Reading Panel, 2000), private industry (Lindhour & Ale, 2009), and government agencies (RAND, 2002). For military linguists

* *Email: kenboyte@gmail.com. Tel.: 831-402-8480. Address: 12765 Via Linda, Castroville, CA 95012 U.S.A.*

on the battlefield, national security is at stake (Kincaid, Fishbourne, Rogers, & Chissom, 1975). In terms of the strategic military importance of literacy education and reading comprehension, DuBay (2004) reported, "General George Washington first addressed concerns about the reading skills of fighters during the Revolutionary War… Since then, the U.S. armed services have invested more in studying workplace literacy than any other organization" (p. 3).

Because misunderstanding texts could result in severe consequences in military operations, reading comprehension should be taken very seriously in the foreign language classrooms that prepare military personnel for their jobs. Responding to this concern, the U.S. Department of Education instructed the RAND Reading Study Group to investigate best practices for teaching and assessing reading comprehension. Consistent with the recommendations of the RAND report published in 2002, the assessment framework known as immediate written recall protocols has been proposed to address government concerns about inadequate literacy education and best practices for measuring reading skills. However, implementation of the alternative assessment framework has not been widespread, likely resulting from a lack of knowledge about immediate written recall protocols and the general ideas held by foreign language teachers about best practices for assessing reading comprehension (Bernhardt, 1991, 2011).

The purpose of this study is to explore the beliefs and practices of foreign language teachers regarding their use of immediate written recall protocols and traditional item types for assessing reading comprehension. In the study, I will demonstrate that the assessment practices and attitudes of respondents regarding immediate written recall protocols and traditional item types are similar to those of foreign language teachers reported in the literature. Further, I show that participation in the research inquiry had an "awareness-raising" impact on some respondents, who reported changes in their beliefs about immediate written recall protocols. The next section of the paper presents a literature review of research on reading comprehension, which began at the end of the 19th century with the birth of the field of psychology and continued to the present, driven in part by the expanding technological demands of education, industry, and national security (DuBay, 2004).

Reading Comprehension

In 1879 Sir Francis Galton of England introduced and defined the term *psychometrics* as "the art of imposing measurement and number upon operations of the mind" (Barrett, 2003). That same year, the scientific study of reading began when Wilhelm Wundt established the world's first laboratory of experimental psychology in Leipzig, Germany (Venezsky, 1984). In the first recorded study of reading comprehension reported in 1884, "Adults read a 10-line paragraph during a fixed period, after which they wrote down everything they could remember" (Venezsky, 1984, p. 13). Later, the British psychologist Sir Frederic Bartlett (1932) reported using written recall protocols during 20 years of investigations into the role of memory in reading comprehension. This work provided the foundation for the development of a constructivist theory of learning and reading comprehension, which gained popularity in the 1970s (Spivey, 1989). Prior to this time, multiple-choice items, a favorite item type of the Audio-Lingual method (Aitken, 1976), were initially introduced by the U.S. military during World War I to rapidly and objectively process and classify large numbers of people. In addition to supporting mass standardized testing procedures, the response format of the item type "lends itself to quantification" (Wigdor & Green, 1991, p. 19). Multiple-choice items can also be characterized as selected-response (Downing, 2006), compared to constructed-response items—often a single word, sentence, or paragraph —to assess writing (Livingston, 2009). One type of constructed-response item, the cloze item, introduced by Taylor (1953, 1956), requires readers to restore words that have been either systematically or randomly deleted from texts. Although debate has persisted about the constructs measured by cloze items, the item type remains popular with teachers because of its "ease of construction, administration, and scoring" (McKamey, 2006, p. 115).

Nevertheless, despite popularity with teachers and test administrators, because the processes involved in reading comprehension are invisible and cannot be directly measured, multiple-choice, cloze, and other

traditional test items generally only serve as indirect measurements, from which inferences about reading comprehension are made (Wolf, 1993); particularly problematic for educators, these inferences have often been misleading (RAND, 2002). Michell (2000) explained the following:

> The attributes that psychometricians aspire to measure are not directly observable (i.e., claims made about them can only [at present] be tested by first observing something else and making inferences). What psychometricians observe are the responses made to test items. Intellectual abilities, personality traits, and social attitudes are theoretical attributes proposed to explain such responses, amongst other things. Typically, test scores are frequencies of some kind, and the hypothesized relations between these theoretical attributes and test scores are taken to be quantitative relations. (p. 648)

In addition to this assumption that reading comprehension and other unobservable psychological constructs can be quantified and measured, the field of psychometrics has been closely associated with a tradition of discrete-point tests, which itself has been based on the false assumption that students who can answer a discrete set of test items can demonstrate language proficiency (Bernhardt, 2011). Aitken (1976) reported, "The essence of discrete point fallacy … is the incorrect assumption that a test of many isolated and separate points of grammar or lexicon is a test of language in any realistic sense" (p. 9). A related issue is that test scores do not have inherent meanings but must be interpreted in relation to the scores of a group of test-takers or a defined assessment standard (Wigdor & Green, 1991). Thus, the lack of correlation, between statistically inflated and/or deflated test scores and the demonstrated language skills of students, has been an underlying source of false inferences about language proficiency derived from traditional assessments (RAND, 2002). The societal consequences of such false assumptions and inferences, and the subsequent misdiagnosis of comprehension skills, are far-reaching (National Reading Panel, 2000; RAND, 2002). Industrial workers who cannot read manuals or warning signs may get hurt (Lindhour & Ale, 2009); likewise, military linguists who make translation errors or are otherwise linguistically unperceptive may be a threat to national security (Kincaid et al., 1975).

Complicating this measurement dilemma, defining reading comprehension has also been difficult; the construct has been described as "multidimensional" (Carlson, Seipel, & McMaster, 2014), "sociocultural" (Roebuck, 1998), and "psycholinguistic" (Goodman, 1968, 1988); it involves bottom-up (Gough, 1972), top-down (Goodman, 1968, 1988), and integrative processes (Glynn, 1983; Kintsch & van Dijk, 1978; Rumelhart, 1990) occurring "as the reader builds one or more mental representations of a text message" (Perfetti & Adlof, 2012, p. 1). Duke and Carlisle (2011) defined *comprehension* as "the act of constructing meaning with oral or written text" (p. 200).

Similarly, cognitive models of the reading process that became popular in the 1970s described reading "as mediated via processes in working memory, a capacity system limited both in terms of quantity of ideas stored and the duration of storage" (Fraser, 2007, p. 373). For example, LeBerge and Samuels' (1974) study presented a model of reading comprehension based on information processing theory, which described the workings of the mind as dependent upon the limited capacity of memory. According to this model, which views the mind as functioning like a computer, reading comprehension is believed to include two main processes: decoding and comprehending (Grabe & Stoller, 2002; Pikulski & Chard, 2003). In contrast to automatic processes of reading, which "are carried out rapidly, effortlessly, and accurately demanding little attention and cognitive resources," controlled processes such as decoding are believed to be slower, more deliberate, and more resource-intensive (Fraser, 2007, p. 372).

Alternatively, the model of reading comprehension advanced by schema theory (Bartlett, 1932) explained comprehension as the relating of "textual material to one's own knowledge," which has been described as the mapping of inputs onto existing concepts via both bottom-up and top-down processes, where "schemata are hierarchically organized, from most general at the top to most specific at the bottom" (Carrell & Eisterhold, 1983, pp. 356-357). The lack of experience or prior knowledge, conceptualized as content schemata, has been

one source of difficulty for L2 readers, particularly relating to unfamiliarity with cultural and historical information (Fisher & Frey, no date; Reisman & Wineburg, 2008; Stahl et al., 1991). Referring to the construction of meaning arising from an interaction between schemata and text, Clarke and Silberstein (1977) reported, "More information is contributed by the reader than by the print on the page" (pp. 136-137). For example, Carlson et al., (2014) reported that the product of reading comprehension, which is believed to involve the tracking of causal relationships presented in text, is "what the reader learned or stored in memory from the text after reading" (p. 41). Sterzik and Fraser (2012) explained, "Overall, text-based comprehension requires students to remember propositions (i.e., ideas) and to attach them to new propositions as they read" (p. 108). Of the five basic structures in expository texts identified by Meyer and Freedle (1984), the most difficult for readers appears to be cause/effect, compared to description, sequence, problem/solution, and compare/contrast. Since the 1970s, the assessment of such aspects of reading comprehension has involved determining learners' abilities to identify main ideas and details in texts using multiple-choice, cloze, and other traditional item types, as well as recall protocols (Akhondi & Malayeri, 2011), which Bachman and Palmer (1996) described as "an extended production response… [ranging] from two sentences or utterances to virtually free writing…" (p. 54).

Written Recall Protocols

Since the 1980s, Bernhardt (1983, 1991, 2000, 2011) and other scholars (e.g., Bernhardt & Deville, 1991; Bintz, 2000; Chang, 2006; Gass & Mackey, 2000; Hayes & Flower, 1980; Johnston, 1983; Leaver, 2013; Miller & Kintsch, 1980; Wells, 1986) have proposed the use of immediate written recall protocols as an alternative to item types generally used in the psychometric tradition of discrete-point tests (Galton, 1879; Goodman, 1968, 1988; Smith, 1971). Criticisms of multiple-choice and other traditional item types have included that "being able to complete the conventional comprehension tasks does not always mean that the students 'understand' a passage" (Bernhardt, 2011, p. 103). Also critical of discrete-point tests, the RAND Reading Study Group (2002) identified false inferences based on traditional item types as the major problem in assessing language competence generally and reading comprehension specifically.

One purported advantage of immediate written recall protocols over traditional item types is that the alternative assessment framework does not interfere with the processes involved in reading comprehension because no leading questions are asked. Wilkinson and Son (2011) reported that the simple act of asking a question changes meaning and alters comprehension. In addition to being less intrusive, immediate written recall protocols are believed to provide a more accurate framework for the assessment of comprehension skills because they can "show where a lack of grammar is interfering with the communication between text and reader, while not focusing a reader's attention on linguistic elements in texts" (Bernhardt, 1991, p. 200). Hayes and Flower (1980) also observed that immediate written recall protocols provide insight into readers' analytical processes, and Johnston (1983) characterized immediate written recall protocols as "the most straightforward assessment of the result of the text-reader interaction" (p. 54).

Grounded in constructivism (Bartlett, 1932), a theory that postulates that readers build "a mental representation from textual cues by organizing, selecting, and connecting content" (Spivey, 1989), immediate written recall protocols may also help researchers study some of the cognitive processes involved in reading, which is viewed as an active process that involves an integration of bottom-up and top-down processing (Kintsch & van Dijk, 1978; Rumelhart, 1990; van Dijk & Kintsch, 1983). Proponents of the theory point to the modifications, regroupings, and simplifications of texts produced by students during the recall process as evidence of the productive nature of reading. Constructivist theory also postulates a strong relationship between reading and writing (Spivey, 1989).

Immediate written recall protocols may also support learning resulting from the cognitive connections believed to be made while summarizing texts and otherwise responding. Although not endorsing immediate written recall protocols specifically, Fisher, Frey, and Lapp (2009) reported that "[s]ummarizing improves students' reading comprehension of fiction and nonfiction alike as it helps the reader construct an overall

understanding of a text" (p. 24). Citing Aebersold and Field (1997), Hedgcock and Ferris (2009) also noted that "[s]ummary-writing is a good review and comprehension check tool" (p. 105). A current example of an immediate written recall protocol developed by Bernhardt and Leaver (no date) to assess students' reading comprehension skills includes the following procedure: "Student reads target language text until they feel comfortable. Then text is removed and student writes down in complete English sentences everything they recall." Describing the procedures involved in immediate written recall protocols, Brantmeier (2006) also reported, "In the written recall task students are asked to read a text, and, without looking back at the text, write down everything they can remember" (p. 2). The productive nature of the task, which integrates reading and writing, is consistent with the theory of constructivism that began with Bartlett (1932) and became popular in the 1970s (Miller, 1973; Spiro, 1980; Spivey, 1989). Although other researchers had used written recall protocols before, Bartlett (1932) is generally credited with first advocating a constructivist explanation for the modifications, regroupings, and simplifications of texts that occur during recall (Spivey, 1989). Other practical and theoretical strengths of immediate written recall protocols over traditional item types can be inferred from Spivey (1989), who reported, "Current reading comprehension tests, typically composed of passages to read and multiple-choice questions to be answered, are clearly inadequate when one examines the task and the texts from a constructivist perspective."

Despite purported benefits, immediate written recall protocols have not been widely used in North America (Bernhardt, 1991, 2011), where language teachers have been generally unaware of the alternative assessment frameworks proposed in the 1970s by Rumelhart (1990) and Kintsch (1974). Even among the language teachers in the United States familiar with the assessment framework, immediate recall protocols have been perceived to involve time-consuming procedures for setting up and scoring (Alderson, 2000; Deville & Chalhoub-Deville, 1993). Even Bernhardt (2011) reported that the matrices for qualitatively and quantitatively scoring "can take many hours to construct" (p. 104). The procedure involves pasting propositions (i.e., ideas) identified in the text into an Excel spreadsheet and ranking each proposition on a scale of 1-4 in terms of importance (least to most important) in relation to the text's meaning. After students have responded, scoring follows, which involves matching the reader's recall to a rubric of propositions (Bernhardt, 2011). Regarding the time-consuming scoring procedures, Bernhardt (2011) noted that research is underway to develop a valid and reliable framework for scoring student responses holistically, rather "than by counting all propositions" (p. 106). Another criticism of immediate written recall protocols has been that the assessment framework relies too much on memory. For example, Koda (2005) reported that "with its strong reliance on memory, free recall makes it difficult to distinguish recalled elements in the text from those retrieved from knowledge bases" (p. 257). This criticism, however, has been challenged by a growing consensus among researchers that memory is essential in reading comprehension (Clark & Silberstein, 1977; Fraser, 2007; Lutz & Huitt, 2003).

To the extent that immediate written recall protocols may be a valuable assessment framework, in view of criticism, some researchers (e.g., Young, 1999) have reported using the alternative assessment framework along with traditional item types to assess various aspects of reading comprehension. Many studies conducted by social scientists also have reported using recall protocols to investigate cognitive processes and the mental constructions of texts (Frederiksen, 1975). Other studies have focused on the effects of discourse types on recall (Meyer & Freedle, 1984), the effects of readability levels on recall (Miller & Kintsch, 1980), compared recall protocols to summary tasks (Riley & Lee, 1996), and compared traditional item types (Wolf, 1993). Although written recall protocols have been widely used in social science research as a measure of comprehension, less attention has been given to the assessment practices and attitudes of foreign language teachers regarding the use of immediate written recall protocols and traditional item types for assessing reading skills (Riley & Lee, 1996). However, related attitudinal studies have reported on general educational trends in Southern Asia (Renandya, Lim, Leung, & Jacobs, 1999), as well as the beliefs and practices of teachers and learners regarding various aspects of education, such as the effectiveness of communicative language teaching (Ngoe & Iwashita, 2012). Specifically focused on the beliefs and practices of foreign language teachers regarding their use of immediate written recall protocols and traditional item types, the present mixed-methods study follows Wubshet and Menuta (2015), who

used an informant interview to gather data in a qualitative analysis of the assessment practices of foreign language teachers. To the extent that the team leader who helped recruit participants and distributed/collected data can be considered an informant, this study differed from Wubshet and Menuta (2013) in its research focus, research methodology, and use of data-collection instruments (e.g., a questionnaire and follow-up inquiry).

Research Questions

In light of the literature and to understand language teaching practices, the present study attempts to answer the following research questions:

(1) What procedures do foreign language teachers prefer to use in assessing students' reading comprehension?
(2) Are foreign language teachers using immediate written recall protocols to assess students' reading comprehension? Why or why not?

These research questions are meaningful in view of the important relationship between literacy education and reading comprehension skills needed for education (National Reading Panel, 2000), industry (Lindhour & Ale, 2009), and national security (Kincaid et al., 1975). Describing reading comprehension research unguided by a unified theoretical foundation as "a problem of great social importance," Kintsch and Miller (1984) argue that, "For our society to function, people have to be able to understand what they read" (p. 200). Understanding teachers' practices in assessing language learners' reading comprehension is a step toward enhancing reading instruction.

Methodology

Participants

Of the 28 respondents in the study (Appendix A), 20 were Korean foreign language teachers, whose L1 is Korean, employed at a U.S. government facility where the leadership, since 2013, has recommended the use of immediate written recall protocols to support the teaching and assessment of reading comprehension. In terms of educational background, 11 of the 20 Korean language instructors have master's degrees in TESOL, Applied Linguistics, or Education. The other eight respondents, one of whose L1 is Korean, were graduate students in a local MATESOL program. One of the English-speaking graduate students reported that their L1 is Spanish.

Instruments

The study utilized a questionnaire (Appendix B), defined as "any written instruments that present respondents with a series of questions or statements to which they are to react…" (Brown, 2001, p. 6), consisting of selected-response Likert-scale items paired with constructed-response items to gather information about each item type analyzed. The study also utilized a follow-up procedure (Table 11) consisting of one written question to which respondents provided written responses.

Data Collection and Analysis

A Korean foreign language instructor and team leader at the U.S. government facility helped distribute/collect questionnaires and follow-up data to/from team members and colleagues. Prior to the start of this research project, the proposed study was submitted for IRB approval and exempted from IRB review.

To obtain the widest breadth of data possible about the assessment practices and attitudes of the foreign language teachers regarding each item type analyzed, the questionnaire utilized paired items consisting of both a constructed-response item and a nine-point Likert-scale (Busch, 1995) selected-response item. Both qualitative

(Berkemeyer, 1989; Lazaraton, 1995; Richards, 2003) and quantitative (Turner, 2014) methods were used to analyze the qualitative and quantitative data. Because a Likert scale was used for some items, this data was treated as "interval-like" and statistically measured (Hatch & Lazaraton, 1991). Thus, the study can be characterized as "mixed method" (Nunan & Bailey, 2009) using both non-intervening and intervening measurement procedures (van Lier, 1988). This method is consistent with a framework articulated by Allwright and Bailey (1991), who also favor such a combined method because it allows for a broader collection and analysis of data. In terms of Grotjan's (1987) framework, the research design can be characterized as non-experimental or "exploratory."

Because foreign language was not the focus of this study, it was not treated as a control variable to exclude teachers based on L1 or foreign language(s) taught. However, the questionnaire was used to collect information about the specific foreign language(s) taught by each instructor. The questionnaire also collected demographic data about the respondents' academic backgrounds and total years of experience teaching foreign language(s). However, information that could identify the respondents was not collected. Although some of the respondents who provided data for the follow-up inquiry also completed the questionnaire, the anonymous data from the questionnaire and the follow-up inquiry cannot be linked.

For the follow-up procedure, the Korean team leader noted above, who at the time of the study managed a team of four foreign language teachers, asked team members and colleagues to provide written responses to the written question: "How do you usually assess reading comprehension in your classes?" A change in the methodology relaxed screening requirements to allow one respondent with less than two years of professional teaching experience to participate in the study.

Findings

Immediate Written Recall Protocols

In response to the question, "Have you ever used immediate written recall protocols to assess reading comprehension?", nine of the 28 respondents (32%) reported "yes," 16 (57%) reported "no," and three (11%) reported "don't know." Paired with this selected-response item was the constructed-response item "Why or why not?" Consistent with the generally purported benefits about immediate written recall protocols reported in the literature (Bernhardt, 1983, 1991, 2000, 2011; Bernhardt & Deville, 1991; Bintz, 2000; Chang, 2006; Gass & Mackey, 2000; Hayes & Flower, 1980; Johnston, 1983; Leaver, 2013; Miller & Kintsch, 1980; Wells, 1986), three of the respondents in the present study (11%) reported that they have used immediate written recall protocols to diagnose "students' weaknesses." Others reported using the assessment framework to diagnose "students' needs" and to diagnose students' problems with grammar.

Some criticisms of immediate written recall protocols reported in the literature (Bernhardt, 1983, 1991, 2000, 2011; Bernhardt & Deville, 1991; Bintz, 2000; Chang, 2006; Gass & Mackey, 2000; Hayes & Flower, 1980; Johnston, 1983; Leaver, 2013; Miller & Kintsch, 1980; Wells, 1986) were also expressed by a few of the foreign language teachers in this study. Four (14%) reported that the use of immediate written recall protocols is either time-consuming or requires a considerable amount of time for development, test administration, and grading (Alderson, 2000; Bernhardt, 2011; Deville & Chalhoub-Deville, 1993). Only one respondent reported that the assessment framework is not effective. Another reported being unfamiliar with immediate written recall protocols before participating in this study. More interestingly, as a result of participating in the study and learning about immediate written recall protocols, some respondents' beliefs about the assessment framework appear to be changing. One respondent reported, "Was not interested. Thought it would take too much time. Now I feel it may be useful." Regarding their changes in attitudes and willingness to try using immediate written recall protocols, others reported, "I haven't had the opportunity. But I'm eager to apply that method," "I'd be open to it."

Multiple-Choice Items

Tables 1 and 2 present the response data for the nine-point Likert-scale item related to disagreement or agreement with the statement: "Multiple-Choice items provide a very good measure of reading comprehension." This selected-response item was paired with the constructed-response item "Please explain your response."

Table 1
Response Data for Question on Multiple-Choice Items

Question	Number of Responses (Disagree—Agree)								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Multiple-Choice items provide a very good measure of reading comprehension.	1	1	3	4	8	3	8	0	0

Table 2
Descriptive Statistics of Responses to Question on Multiple-Choice Items

Mean	5.231
Median	5.0

Similar to opinions about the item type reported in the literature (Aitken, 1976; Downing, 2006; Oller, 1979; RAND, 2002; Wigdor & Green, 1991), many respondents were critical of multiple-choice items. With a mean of 5.23/9.0 and a median of 5.0/9.0, an analysis of the Likert-scale data revealed that the overall teacher attitudes were somewhat moderate about this item type. Only one respondent indicated strongly disagreeing that multiple-choice items provide a very good measure of reading comprehension (Likert score=1/9), but eight indicated some degree of agreement (Likert score=7/9). Twelve (43%) reported problems with multiple-choice items resulting from poor quality test questions and poor quality distractors. One respondent reported that multiple-choice items "may provide inflated scores of reading comprehension due to background information," and four (14%) reported that students can often guess the correct answer to multiple-choice items. Other respondents also reported that such item types are "limited" in terms of assessing reading comprehension, and that there are "better ways to assess students' overall understanding." Only one respondent reported that multiple-choice items are "objective." However, depending on the quality of the item, one respondent reported that multiple-choice items do a very good job of assessing reading comprehension "because it makes students think." Another reported that such an item type "could be more effective for assessing higher levels of nuance." Multiple-choice items also may support test administration and scoring, according to one respondent.

Written Summaries

To gather data about this item type, respondents were asked to indicate their disagreement or agreement on a nine-point Likert scale with the statement: "Grading students' summaries of written texts is too time-consuming." The constructed-response item, "Please explain your response," was paired with the Likert-scale item; the response data is presented in Tables 3 and 4.

Table 3
Response Data for Question on Written Summaries

Question	Number of Responses (Disagree—Agree)								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Grading students' summaries of written texts is too time-consuming.	1	0	7	2	3	3	9	0	3

Table 4
Descriptive Statistics of Responses to Question on Written Summaries

Mean	5.519
Median	6.0
Standard Deviation	2.190

Consistent with the literature (Alderson, 2000; Bernhardt, 2011; Deville & Chalhoub-Deville, 1993), five of the 28 respondents (18%) reported that grading written summaries is time-consuming (mean 5.52/9.0, median 6.0/9.0). Three reported strong agreement (Likert score=9/9), and nine reported some degree of agreement (Likert score=7/9), with only one reporting strong disagreement (Likert score=1/9). Similarly, one respondent reported that using written summaries is pedagogically "necessary at the stage of foundation." Another reported, "In order to summarize texts, students need to get essential elementary information (Livingston, 2009; RAND, 2002; Riley & Lee, 1996). So, students' summaries would give teachers an idea about how much students comprehend texts." Whether written summaries provide a very good measure of reading comprehension depends on the quality of the rubrics developed for scoring the items (reported by three respondents) and the systematicity and objectivity of the grading process (reported by three other respondents). Four of the respondents (14%) reported that grading and providing feedback to students' written summaries can be difficult because of issues related to handwriting and readability.

True-False Items

Information about the foreign language teachers' attitudes toward true-false test items was obtained by analyzing responses to the statement: "True-False items provide a very good measure of reading comprehension." The response data for the nine-point Likert scale item is presented in Tables 5 and 6. Paired with this item was the constructed-response item, "Please explain your response."

Table 5
Response Data for Question on True-False Items

Question	Number of Responses (Disagree—Agree)								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
True-False items provide a very good measure of reading comprehension.	4	1	3	4	10	2	4	0	0

Table 6
Descriptive Statistics of Responses to Question on True-False Items

Mean	4.444
Median	5.0
Standard Deviation	1.783

Based on an analysis of the data (e.g., mean 4.4/9.0, median 5.0/9.0), true-false items are believed to be the least effective item type investigated in this study. Consistent with the literature (Aitken, 1976; Downing, 2006; Oller, 1979; RAND, 2002; Wigdor & Green, 1991), six of the 28 respondents (21%) reported problems with true-false items related to guessing (e.g., "It's a 50/50 chance to select the correct answer"). Other foreign language teachers reported that the effectiveness of true-false items for assessing reading comprehension depends on the quality of the item, the context of the item, and the specific questions. Although criticized as "very low level" assessment tools, some respondents reported that true-false items do have some assessment value: "They can help to see initial logic/comprehension of a text but don't give a good measure of reading comprehension," and "I don't know if they measure 'very good,' but I think true-false test items can still measure reading comprehension to a certain extent."

Fill-in-the-Blank Items

Tables 7 and 8 present the response data for the nine-point Likert scale item related to the statement: "Fill-in-the-Blank items provide a very good measure of reading comprehension." Paired with this selected-response item was the constructed-response item, "Please explain your response."

Table 7
Response Data for Question on Fill-in-the-Blank Items

Question	Number of Responses								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Fill-in-the-Blank items provide a very good measure of reading comprehension.	0	2	3	3	5	5	9	1	0

Table 8
Descriptive Statistics of Responses to Question on Fill-in-the-Blank Items

Mean	5.286
Median	5.50
Standard Deviation	1.696

Beliefs about fill-in-the-blank items reported in this study also reflected those of other foreign language teachers reported in the literature (Aitken, 1976; Downing, 2006; Oller, 1979; RAND, 2002; Wigdor & Green, 1991). Overall, the item type is thought to be more effective than true-false items for assessing reading comprehension although the teachers' reported beliefs are somewhat moderate (e.g., mean 5.28/9.0, median 5.5/9.0). Based on an analysis of this data, two of the respondents (7%) reported that the effectiveness of fill-in-the-blank items for assessing reading comprehension depends on the quality and focus of the test items. For such items to be effective, two respondents reported that a clear answer key is needed. The placement of blanks is also important, according to two respondents. Although fill-in-the-blank items are criticized for being "low level" and focused only on "surface-level understanding," the item type may have some assessment value. One respondent reported, "Sometimes fill-in-the-blank items are useful to check comprehension because they narrow in on reading for details." Another respondent reported that the item type focuses students' attention on grammatical forms. The usefulness of the item type was further articulated by another respondent, who reported that fill-in-the-blank items require more local thinking and problem-solving.

Cloze Items

Tables 9 and 10 present the response data for the nine-point Likert-scale item related to the statement: "Cloze items provide a very good measure of reading comprehension," which was paired with the constructed-response item, "Please explain your response."

Table 9
Response Data for Question on Cloze Items

Question	Number of Responses								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Cloze items provide a very good measure of reading comprehension.	1	1	4	1	5	5	8	2	1

Table 10
Descriptive Statistics of Responses to Question on Cloze Items

Mean	5.481
Median	6.0
Standard Deviation	1.987

Of all the traditional item types analyzed in this study and compared on the basis of Likert-scale data, cloze items (mean 5.48/9.0, median 6.0/9.0) are believed to be the most effective item type for assessing reading comprehension. In terms of the strengths and weaknesses of cloze items, akin to the beliefs of foreign language teachers reported in the literature (Aitken, 1975; McKamey, 2006; Taylor, 1953, 1956; Williams, 1974), three respondents (11%) reported that cloze items may provide a good measure of comprehension if properly constructed. One respondent reported that an accurate and comprehensive answer key is required for the item type to be effective. Two respondents also reported that cloze items are good for assessing grammar. Others reported that cloze items "might be a good way to test students' understanding in the context of vocabulary," and that cloze items support critical thinking (e.g., "They make students think, evaluate the context of a text"). Respondents also reported that cloze items are good for beginning-level students, and for focusing on "students' accuracy in foreign language learning." Critical of the item type, respondents in the present study reported that only a "limited amount of understanding can be measured" by cloze items and that they "just don't like cloze items." Similarly, Williams (1974) has criticized cloze items on the basis that the item type does not measure the primary processes involved in reading comprehension (e.g., decoding written symbols) but only assesses production (encoding). Alderson (2000), Bachman (1982, 1985), and Shanahan, Kamil, and Tobin (1982) also criticize cloze items for not being able to measure macro-level and higher-order thinking skills.

Follow-Up Inquiry

Table 11 presents the written responses that four foreign language teachers provided to the question: "How do you usually assess reading comprehension in your classes?" Two of the four instructors questioned in the follow-up inquiry appear to be using a form of immediate written recall protocols to assess reading comprehension. For example, respondent two reported, "I ask students to read a text and direct them to write the summary/gist of the reading passage both in Korean and English depending on their level." Respondent four also reported using a similar procedure. Although respondent one reported using a procedure somewhat different from immediate written recall protocols (e.g., focusing on training students to identity the main subject and verbs in complex sentences), respondent one did not report using traditional item types. In fact, only respondent three reported using "comprehend questions" for assessing reading comprehension.

Table 11
Response Data for Follow-Up Inquiry

Respondent	Response
1	I ask students to identify sentence structures (main subject and verbs) from complex sentences. And I ask … whether they know the meaning of key words/basic words in texts.
2	I ask students to read a text and then direct them to write a summary/gist of the reading passage both in Korean and English, depending on their level.
3	I provide comprehend questions for students and have them answer the questions.
4	(a) I ask students to read a target-language text until they feel comfortable with the material. I then remove the text and ask students to write down in complete English sentences everything they can remember; (b) I collect and analyze the data; (c) I incorporate instructional activities/strategies accordingly.

Discussion

Interpretation of Data

This study has attempted to answer the research questions: (1) What procedures do foreign language teachers prefer to use in assessing students' reading comprehension? and (2) Are foreign language teachers using immediate written recall protocols to assess students' reading comprehension? Why or why not? The results indicate that nine of the 28 respondents (32%) had previously used immediate written recall protocols to assess reading comprehension, which was expected based on trends reported in the literature (Bernhardt, 1991, 2011). In a study by Wubshet and Menuta (2015), none of the foreign language teachers confirmed using any type of alternative assessment, which are reportedly believed to be time-consuming to administer and grade (Alderson, 2000; Bernhardt, 2011; Deville & Chalhoub-Deville, 1993). Respondents in the present study had similar criticisms about immediate written recall protocols. Reflecting other criticism reported in the literature (e.g., Alderson, 2000; Oller, 1979; RAND, 2002), many respondents in the present study additionally doubted the validity of multiple-choice and other traditional item types. Among the main criticisms reported by respondents in the present study were that the correct answers to multiple-choice and true-false items can be guessed without reading related texts, and that traditional item types do not assess higher-order thinking skills. Table 12 presents a ranking of the traditional item types reported by respondents to be the most effective measures of reading comprehension.

Table 12
Perceived "Best" Measurements of Reading Comprehension

Item Type	Mean	Median	Standard Deviation
Cloze	5.481	6.0	1.987
Fill-in-the-Blank	5.286	5.5	1.696
Multiple Choice	5.231	5.0	1.607
True-False	4.444	5.0	1.783

Notice that, like immediate written recall protocols, the top two traditional item types (cloze and fill-in-the-blank items) involve the productive process of writing. Although an overall analysis of the data revealed that most respondents had moderate opinions about immediate written recall protocols and traditional item types (the mean and median of the Likert-scale items included in the questionnaire hovered around 5.0/9.0), looking only at the mathematical averages masks the fact that some teachers hold polarizing opinions about item types. Attempting to understand these differing opinions about item types, an analysis of the demographic data also revealed that the nine respondents who reported using immediate written recall protocols have masters' degrees in TESOL or a related field, indicating to this author the impact of foreign language education on their pedagogical beliefs and assessment practices. The positive effect of educational training also can be inferred from the fact that the nine respondents who reported using immediate written recall protocols all work for a U.S. government facility that has been providing training for foreign language teachers to promote the usage of the alternative assessment framework.

The study additionally revealed that participation in the research had an "awareness-raising" impact on some respondents who reported changes in their beliefs about immediate written recall protocols, which was expected by this author in view of what McCambridge, Witter, and Elbourne (2014) have reported about the 'Hawthorne Effect' (e.g., participating in a research study changes the behaviors of those being studied). Although 16 of the 28 respondents reported never using immediate written recalls, an analysis of the response data revealed that their attitudes and assessment practices can probably be attributed to negatively held opinions about the alternative assessment framework or a general lack of knowledge about immediate written recall protocols (Bernhardt, 1991, 2011). Given that the beliefs and practices of other foreign language teachers may be affected by participating in and learning from a similar research inquiry and assuming that administrators and

policymakers would like more teachers to use immediate written recall protocols in their classrooms, it is the recommendation of this author that this study should be replicated and expanded upon.

Limitations

Although 40 respondents were initially sought for the study, only 28 returned a completed questionnaire. The low rate of participation may have been impacted by the paper format of the questionnaire. A misalignment in the design of the questionnaire, discovered by the author while reviewing the collected data, required the use of a follow-up question to determine the procedures foreign language teachers use in their classrooms to assess students' reading comprehension. Still, because a Likert-scale item was mistakenly not included to collect information about the perceived effectiveness of summary-writing tasks, teachers' attitudes about the item type cannot be directly compared with teachers' attitudes about other item types reflected in the Likert-scale data.

Despite a proofreading error in the questionnaire, which also may have contaminated some of the Likert-scale data, a rich source of confirmation for the reported numerical data was provided by respondents' written explanations to paired questionnaire items, and by hand-written notations on some items. A follow-up question posed to four of the 28 respondents further triangulated the reported interpretation of what was learned about the attitudes and assessment practices of foreign language teachers regarding immediate written recall protocols and traditional test item types. For future research, the author hopes to work more closely with colleagues and an oversight committee, prior to beginning the study, to more carefully review questionnaires, surveys, and other data-collection instruments and to more closely align research instruments with the research questions they are designed to measure.

Pedagogical Implications

In terms of classroom applications, although the actual nature of reading comprehension remains disputed (Carlson et al., 2014; Duke & Carlisle, 2011; Goodman, 1968, 1988; Roebuck, 1998), this study brings attention to the important cognitive processes involved in reading comprehension, as well as the roles that memory, prior experience, and cultural knowledge contribute to reading comprehension (Fisher & Frey, no date; Reisman & Wineburg, 2008; Stahl, et al., 1991). Thus, foreign language teachers can utilize this insight to design lesson plans and curricula that help students build cultural background knowledge, as well as vocabulary and grammar knowledge, through the use of a variety of activities that address the major cognitive processes involved in reading comprehension: decoding and comprehending (Fraser, 2007; Grabe & Stoller, 2002; Pikulski & Chard, 2003) and bottom-up (text-driven) and top-down (knowledge-driven) processes (Carrell & Eisterhold, 1983; Kintsch & van Dijk, 1978; Rumelhart, 1990; van Dijk & Kintsch, 1983). Exposing students to a diverse array of texts is one way to build up top-down cultural knowledge. An example of tasks targeting bottom-up processes could involve vocabulary-building activities focused on word knowledge (Stahl et al., 1991).

Despite criticisms that recall protocols are not valid measures of reading comprehension due to the influence of memory (Koda, 2005), a growing pedagogical consensus that memory is an important component of reading comprehension has been emerging (Clark & Silberstein, 1977; Fraser, 2007; Lutz & Huitt, 2003). Based on this insight, foreign language teachers should help students develop their memories, which in turn may strengthen reading comprehension skills in the same way that summary tasks reportedly support the comprehension of text (Aebersold & Field, 1997; Fisher et al., 2009). Thus, like the assessment practices reported in this study, teachers should provide students with practice in summarizing texts. Similarly, teachers can also help students to track causal relationships by providing classroom activities and homework that focuses on identifying relationships between nouns and verbs (Carlson et al., 2014; Meyer & Freedle, 1984). From the perspective of constructivist theory (Bartlett, 1932; Spivey, 1989), because of the integrative nature of reading comprehension (Riley & Lee, 1996), teachers should also plan learning activities that are integrative and combine both reading and writing tasks (Spivey, 1989). A focus on grammar in a writing activity may reveal relational problems with determining causality that can be addressed with further targeted instruction.

Of all the pedagogical issues associated with immediate written recall protocols, perhaps the most

important concerns scoring procedures, which Wells (1986) reported will continue to limit the deployment of the alternative assessment framework until commercially produced products are available: "At present, in the absence of professionally prepared test passages, analyses, and scoring instruments, it is unreasonable to expect the classroom teacher to use the recall procedure as a large-scale classroom evaluation tool" (p. 178). Although not directly challenging the validity of the item type, many respondents in the present study also complained that scoring procedures are time-consuming (Alderson, 2000; Bernhardt, 2011; Deville & Chalhoub-Deville, 1993). Other concerns about scoring reported by respondents in the present study were related to the validity of poorly constructed test items and scoring instruments, such as rubrics and answer keys. In fact, 12 respondents (43%) reported problems with multiple-choice items regarding poorly constructed distractors and overall item quality. Criticisms about written summaries related to the quality of rubrics and the systematicity and objectivity of scoring. True-False items were criticized on the basis of item quality, item context, and specific test question. Similarly, problems with item quality, answer keys, item focus, and the placement of blanks were reported about fill-in-the-blank items. Respondents additionally reported that cloze items can be problematic depending on the construction of test items and a clear, accurate, and comprehensive answer key. With these findings in mind, it is the recommendation of this author that foreign language teachers should focus their awareness on the related variables identified in this study when developing, administering, scoring, and/or evaluating reading comprehension tests.

Conclusion

Continuing an international conversation about best practices for education and testing dating back to the ancient Greeks (Barrett, 2003; Michell, 1999, 2000), this study has attempted to provide insight into teachers' perceptions of the alternative assessment known as immediate written recall protocols, which proponents believe offers a better measure of reading comprehension than traditional item types (Bernhardt, 1983, 1991, 2000, 2011; Bernhardt & Deville, 1991; Bintz, 2000; Chang, 2006; Gass & Mackey, 2000; Hayes & Flower, 1980; Johnston, 1983; Leaver, 2013; Miller & Kintsch, 1980; Wells, 1986).

With regard to the psychological operations involved in language teaching generally and reading comprehension specifically, this study focused on the assessment practices and attitudes of foreign language instructors at one U.S. military foreign language school that has been using immediate written recall protocols in an attempt to more accurately measure reading comprehension skills. Beyond a report on teachers' assessment practices, this study seemed to have created some positive impact on the teachers themselves. Three of the respondents in this study reported positive changes in their beliefs about the alternative assessment framework as a result of participating in the study. However, prior training provided by the employer, as well as other prior experience studying the subject and prior exposure to reading material about immediate written recall protocols, also may have been part of the 'learning' process leading up to the attitudinal changes reported in this study. Nevertheless, considering that the beliefs and practices of other foreign language teachers could change as a consequence of participating in and learning from a similar research inquiry, this author recommends that this study be replicated and expanded upon. Although classroom teachers in North America have not widely used immediate written recall protocols to assess reading comprehension, this situation could change with education and training—bringing to public attention the importance of reading comprehension and raising awareness about the assessment benefits that immediate written recall protocols may provide beyond the limitations of discrete-point tests and the psychometric model.

In view of the ongoing public debate about best practices for education generally, it is the further opinion of this author that the Pythagorean tradition of using mathematics as the primary tool for discovering and understanding the underlying principles of the natural world should be reconsidered (Barrett, 2003). Whether psychometrics deserves its current prestige or should be considered "a pathology of science" (Michell, 2000), the continuing study of theories of the mind generally and theories of reading comprehension specifically will remain important because research models can impact fundamental aspects of modern life (Rust & Golombok,

2009). As relevant theory and research methodologies continue to develop, I hope that a proper balance can be found between convenience for test administration/ scoring and the utilization of integrative response formats that elicit richer data sources that can be both qualitatively and quantitatively scored to more directly inform instruction. Perhaps such an assessment framework will include a combination of immediate written recall protocols and traditional item types, as well as the ongoing observations of instructors throughout the pedagogical process. More grounded in constructivist theories of the mind and learning, immediate written recall protocols may already be contributing to a shift in educational paradigms, bringing methodologies for assessing reading comprehension back in line with those used in the first scientific experiments in psychology.

References

Aebersold, J. A., & Field (1997). *From reader to reading teacher: Issues and strategies for second Language classrooms.* Cambridge, UK: Cambridge University Press.

Aitken, K. G. (1975). Problems in cloze testing re-examined. *TESL Reporter*, 9, 16-18. Retrieved from: https://ojs.lib.byu.edu/spc/index.php/TESL/article/viewFile/2796/2579

Aitken, K. G. (1976). Discrete structure point testing: Problems and alternatives. *TESL Reporter*, 7-20. Retrieved from: https://ojs.lib.byu.edu/spc/index.php/TESL/article/download/.../2644

Akhondi, M., & Malayeri, F. A. (2011). Assessing reading comprehension of expository text across different response formats. *Iranian Journal of Applied Language Studies, 3*(1), 1-26. Retrieved from: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CB4QFjAA&url=http%3A%2F%2Fijals2.usb.ac.ir%2Fpdf_76_d1d9b0119eb069f45ffa322ab0b19f4e.html&ei=_SHhVLqHPM_coATv9oLgAg&usg=AFQjCNGii1Nl1XBpmAOpHGLK2JPOVC93sw&sig2=5dEIsvRuGjZY4XS371VjUg&bvm=bv.85970519,d.cGU

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

Allwright, D., & Bailey, K. M. (1991). *Focus on the classroom: An introduction to classroom research for language teachers*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*(1), 61-70.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535-555.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Barrett, P. T. (2003). Beyond psychometrics: Measurement, non-quantitative structure, and applied numeric. *Journal of Managerial Psychology*, *3*(18), 421-439. Retrieved from: http://www.pbarrett.net/publications/Barrett_Beyond_Psychometrics_2003_Augmented.pdf

Bartlett, F. (1932). *Remembering*. Cambridge, UK: Cambridge University Press.

Berkemeyer, V. C. (1989). Qualitative analysis of immediate recall protocol data: Some classroom implications. *Die Unterrichtspraxis/Teaching German 22*(2), 131-137.

Bernhardt, E. B. (1983). Testing foreign language reading comprehension: The immediate recall protocol. *Die Unterrichtspraxis/Teaching German 16*(10), 27-33. Retrieved from: http://files.eric.ed.gov/fulltext/ED285420.pdf

Bernhardt, E. B. (1991). *Reading development in a second language*. Norwood, NJ: Ablex.

Bernhardt, E. B. (2000). Second language reading as a case study of reading scholarship in the 20th century. In M. L. Kamil, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research, Vol. III* (pp. 791-811). Manwah, NJ: Lawrence Erlbaum Associates Publishers.

Bernhardt, E. B. (2011). *Understanding advanced second language reading*. New York, NY: Routledge.

Bernhardt, E. B., & DeVille, C. (1991). Testing in foreign language programs and testing programs in foreign language departments. In R. V. Teschner (Ed.), *Issues in language program direction: Assessing foreign language proficiency of undergraduates* (pp. 43-59). Boston, MA: Heinle & Heinle.

Bernhardt, E. B., & Leaver, B. L. (no date). Recall protocol for diagnosing listening and reading comprehension. Faculty Development Division. Monterey, CA: U.S. Army Defense Language Institute.

Bintz, W. P. (2000). Using free writing to assess reading comprehension. *Reading Horizons*, *40*(3), 205-223. Retrieved from: http://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1220&context=reading_horizons

Brantmeier, C. (2006). The effects of language assessment and L2 reading proficiency on advanced readers' recall. *The Reading Matrix*, *6*(1), 1-16. Retrieved from: http://www.readingmatrix.com/articles/brantmeier/article5.pdf

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.

Busch, M. (1995). Using Likert scales in L2 research. *TESOL Quarterly*, *27*(4), 733-736.

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences*, *12*, 40-53.

Carrell, P. I., & Eisterhold, J. (1983). Schema theory and ESL pedagogy. *TESOL Quarterly*, *17*, 553-573.

Chang, Y. F. (2006). On the use of immediate recall protocols as a measure of second language reading comprehension. *Language Testing, 23*(4), 520-543.

Clarke, M. A., & Silberstein, S. (1977). Toward a realization of psycholinguistic principles in the ESL reading class. *Language Learning*, *27*(1), 135-154.

Deville, C., & Chalhoub-Deville, M. (1993). Modified scoring, traditional item analysis, and Sato's caution index used to investigate the reading recall protocol. *Language Testing*, *10*(2), 117-132.

Downing, S. M. (2006). Selected-response format in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 287-302). New York, NY: Routledge.

DuBay, W. H. (2004). *The principle of readability*. Costa Mesa, CA: Impact Information. Retrieved from: http://www.impact-information.com/impactinfo/readability02.pdf

Duke, N. K., & Carlisle, J. (2011). The development of comprehension. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research, Vol. IV* (pp. 199-228). New York, NY: Routledge.

Fisher, D., Frey, N., & Lapp, D. (2009). *In a reading state of mind*. Newark, DE: International Reading Association.

Fisher, D., & Frey, N. (no date). *Background knowledge: The overlooked factor in reading comprehension.* New York, NY: McGraw-Hill Education. Retrieved from: http://mcgrawhillflnetworks.com/pdf/White_Papers/8353_networks_Bckgrnd_Knwld_WhitePaper.pdf

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading*, *10*(3), 323-330.

Fraser, C. A. (2007). Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks. *Modern Language Journal*, *91*, 372-394.

Frederiksen, C. H. (1975). Effects of content-induced processing operations on semantic information acquired from discourse. *Cognitive Psychology, 7, 139-166.*

Galton, F. (1879). Psychometric experiments. *Brain*, *2*, 149-162. Retrieved from: http://galton.org/galton/essays/1870-1879/galton-1879-brain-psychometric- experiments/galton-1879-brain-psychometric-experiments.pdf

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

Glynn, S. M. (April 1983). Cognitive processes involved in text learning. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada. Retrieved from: http://files.eric.ed.gov/fulltext/ED232128.pdf

Goodman, K. (Ed.) (1968). *The psycholinguistic nature of the reading process*. Detroit, MI: Wayne State University Press.

Goodman. K. (1988). The reading process. In P. Carrell, J. Devine, & D. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 11-21). Cambridge, UK: Cambridge University Press.

Gough, P. B. (1972). *One second of reading.* In J. F. Kavanagh & I. G. Mattingley (Eds.), *Language by ear and by eye* (pp. 331-358). Cambridge, MA: MIT Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Harlow, UK: Longman/Pearson Education.

Grotjan, R. (1987). On the methodological basis of introspective methods. In C. Faerch & G. Karper (Eds.). *Introspection in second language research* (p. 59). Clevedon, UK: Multilingual Matters.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle & Heinle Publishers.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 1-30). Hillsdale, NJ: Lawrence Erlbaum.

Hedgcock, J., & Ferris, D. (2009). *Teaching readers of English*. New York, NY: Routledge.

Johnston, P. H. (1983). *Reading comprehension assessment: A cognitive basis*. Newark, DE: International Reading Association.

Kamil, M. (1984). Current traditions of reading research. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 39-62). New York, NY: Longman.

Kincaid, J. P., Fishbourne, R. P. S., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formula (Automated Readability Index, Fog Cog, and Flesch Reading Ease Formula) for Navy enlisted personnel. *Research Branch Report 8-75*. Millington, TN: Naval Technical Training Command.

Kintsch, W. (1974). *The representation of meaning in memory*. New York, NY: John Wiley & Sons.

Kintsch, W., & Miller, J. R. (1984). Readability: A view from cognitive psychology. In J. Flood (Ed.), *Understanding reading comprehension* (pp. 220-232). Newark, DE: International Reading Association.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

Koda, K. (2005). *Insights into second language reading: A cross linguistic approach.* New York, NY: Cambridge University Press.

Lazaraton, A. (1995). Qualitative research in applied linguistics: A progress report. *TESOL Quarterly*, *29*(3), 455-472.

Leaver, B. L. (2013). From the provost's desk: Recall protocols. *Dialog on Language Instruction*, *23*, 178. Retrieved from: http://www.dliflc.edu/wp-content/uploads/2014/04/DLI_Vol_23.pdf

LeBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293-323.

Lindhour, P., & Ale, B. J. M. (2009). Language issues, an underestimated danger in major hazardous control. *Journal of Hazardous Materials*, *172*, 247-255. Retrieved from: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=12&ved=0CCIQFjABOAo&url=http%3A%2F%2Fwww.researchgate.net%2Fprofile%2FPaul_Lindhout%2Fpublication%2F26698675_Language_issues_an_underestimated_danger_in_major_hazard_control%2Flinks%2F0f317539c867223e34000000.pdf&ei=J9_8VJGnItGzoQT9n4LwCQ&usg=AFQjCNEOE0n4T2Sq_FPZK1Tt0tlFCgyKhw&sig2=LlowaxGtxJicyIWfFTGwQw&bvm=bv.87611401,d.cGU

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. *ETS R&D Connections*, *11*. Retrieved from: https://www.ets.org/Media/Research/pdf/RD_Connections11.pdf

Lutz, S., & Huitt, W. (2003). Information processing and memory: Theory and applications. *Education Psychology Interactive*. Valdosta, GA: Valdosta State University. Retrieved from: http://www.edpsycinteractive.org/papers/infoproc.pdf

McCambridge, J., Witter, J., & Elbourne, D. (2014). Systematic review of the Hawthorne Effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267-277. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3969247/

McKamey, T. (2006). Getting closure on the cloze: A validation study of the 'rational deletion' method.

Second Language Studies, *24*(2), 114-164. Retrieved from: http://www.hawaii.edu/sls/wp-content/uploads/2014/09/McKameyTreela.pdf

Meyer, B. J. F., & Freedle, R. O. (1984). Effects of discourse types on recall. *American Educational Research Journal*, *21*(1), 121-143.

Michell, J. B. (1999). Measurement in psychology: *Critical history of methodological concepts.* Cambridge, UK: Cambridge University Press. Retrieved from: https://archivocienciassociales.files.wordpress.com/2015/03/joel-michell-measurement-__in-psychology-a-critical-history-of-a-methodological-concept-1999.pdf

Michell, J. B. (2000). Normal science, pathological science, and psychometrics. *Theory and Psychology*, *10(*5), 639-667. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.202.2742&rep=rep1&type.pdf

Miller, G. A. (Ed.) (1973). *Linguistic communication: Perspectives for research.* Newark, DE: International Reading Association.

Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory, 6*(4), 335-354.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development. Retrieved from: http://www.nichd.nih.gov/publications/pubs/nrp/documents/report.pdf

Ngoc, K. M., & Iwashita, N. (2012). A comparison of learners' and teachers' attitudes toward communicative language teaching at two universities in Vietnam. *University of Sydney Papers in TESOL*, *7*, 25-49. Retrieved from: http://faculty.edfac.usyd.edu.au/projects/usp_in_tesol/pdf/volume07/Article02.pdf

Nunan, D., & Bailey, K. M. (2009). *Exploring second language classroom research: A comprehensive guide.* Boston, MA: Heinle, Cengage Learning.

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

Perfetti, C. A., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading* (pp. 3-20). Plymouth, UK: Rowman & Littlefield Publishers. Retrieved from: http://www.lrdc.pitt.edu/bov/documents/perfetti_reading%20comprehension.pdf

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227-247). Oxford, UK: Blackwell. Retrieved from: http://www.pitt.edu/~perfetti/PDF/Acquisition%20(Oakhill%20chapter).pdf

Pikulski, J., & Chard, D. J. (2003). Fluency: The bridge from decoding to reading comprehension. *Current research in reading/language arts*. Houghton Mifflin. Retrieved from: http://www.eduplace.com/state/author/pik_chard_fluency.pdf

RAND Reading Study Group (2002). Reading for understanding: Toward an R&D program in reading comprehension. Retrieved from: http://www.prgs.edu/content/dam/rand/pubs/monograph_reports/2005/MR1465.pdf

Reisman, A., & Wineburg, S. (2008). Teaching the skill of contextualizing in history. *The Social Studies*, *99*(5), 202-207.

Renandya, W. A., Lim, W. L., Leung, K. W., & Jacobs, G. M. (1999). A survey of English language teaching trends and practices in Southern Asia. *Asian Englishes*, *2*, 37-65. Retrieved from: http://www.academia.edu/6509966/A_Survey_of_English_Language_Teaching_Trends_and_Practices_in_Southeast_Asia

Richards, K. (2003). *Qualitative inquiry in TESOL*. New York, NY: Palgrave Macmillan.

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, *13*(2), 173-189.

Roebuck, R. (1998). *Reading and recall in L1 and L2: A sociocultural approach*. Stanford, CT: Ablex Publishing Company.

Rumelhart, D. E. (1990). Toward an interactive model of reading. In S. Dornie (Ed.), *Attention and performance, VI* (pp. 573-603). New York, NY: Academic Press.

Rust, J., & Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment* (3rd ed.). New York, NY: Routledge.

Shanahan, T., Kamil, M., & Tobin A. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*, 229-255.

Smith, F. (1971). *Understanding reading*. New York, NY: Holt, Rinehart, and Winston.

Spiro, R. J. (1980). *Schema theory and reading comprehension*, Technical Report No. 191. Urbana-Champaign, IL: University of Illinois. Retrieved from: https://www.ideals.illinois.edu/bitstream/handle/2142/17694/ctrstreadtechrepv01980i00191_opt.pdf?sequence=1

Spivey, N. N. (1989). Construing constuctivism: Reading research in the United States. Occasional Paper No. 12. Retrieved from: http://www.nwp.org/cs/public/download/nwp_file/104/OP12.pdf?x-r=pcfile_d

Stahl, S. A., Hare, V. C., Sinatra, R., & Gregory, J. F. (1991). *Technical report No. 526: Defining the role of prior knowledge and vocabulary in reading comprehension: The retiring of number 41.* Urbana-Champaign, IL: University of Illinois. Retrieved from: https://www.ideals.illinois.edu/bitstream/handle/2142/17521/ctrstreadtechrepv01991i00526_opt.pdf?sequence=1

Sterzik, A. M., & Fraser, C. (2012). RC-MAPS: Bridging the comprehension gap in EAP reading. *TESL Canada Journal/Revue TESL Du Canada*, *28*(2), 103-119. Retrieved from: http://www.teslcanadajournal.ca/index.php/tesl/article/viewFile/1103/922

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 414-433.

Taylor, W. L. (1956). Recent developments in the use of 'cloze procedure.' *Journalism Quarterly*, *33*(1), 42-48, 99.

Turner, J. L. (2014). *Using statistics in small-scale language education research*. New York, NY: Routledge.

van Dijk, T. A., & Kintsch, W. (1983*). Strategies of discourse comprehension.* New York, NY: Academic.

van Lier, L. (1988). *The classroom and the language learner: Ethnography and second language classroom research*. London, UK: Longman.

Venezsky, R. L. (1984). The history of reading research. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 3-38). New York, NY: Longman.

Wells, D. R. (1986). The assessment of foreign language reading comprehension: Refining the task. *Die Unterricht/Teaching German*, 178-184. Retrieved from: https://www.jstor.org/stable/3530700?seq=3#page_scan_tab_contents

Wigdor, A., & Green, B., Jr. (Eds.) (1991). *Performance assessment for the workplace, Vol. 1*. Washington, DC: Committee on the Performance of Military Personnel, National Academy Press. Retrieved from: http://www.nap.edu/read/1862/chapter/1

Wilkinson, I. A. G., & Son, E. H. (2011). A dialogic turn in research on learning and teaching to comprehend. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research, Vol. IV* (pp. 359-387). New York, NY: Routledge.

Williams, R. K. (1974). Problems in cloze testing. *TESL Reporter*, 7-9. Retrieved from: https://journals.lib.byu.edu/spc/index.php/**TESL**/article/download/2772/2555

Wolf, D. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal*, *77*, 473-489.

Wubshet, H., & Menuta, F. (2015). Investigating the practice of alternative assessment in English classrooms: The case of selected grade nine English teachers assessment practices. *International Journal of*

Scientific Research in Education, *8*(4), 159-171. Retrieved from: http://www.ijsre.com/Vol.,%208(4)-Wubshet%20&%20Menuta.pdf

Young, D. J. (1999). Linguistic simplification of SL reading material: Effective instructional practice? *Modern Language Journal*, *83*, 350-366.

Appendix A
Respondent Demographics

Foreign Languages Taught	Educational Background in TESOL/Linguistics/ SLA/Education	L1
English	7	7
Korean	11	20
Spanish	1	1

Respondent	FL Taught	# of Years Taught	University Degree	L1
1	Korean	11	MATESOL; BA, English, French	Korean
2	English	6	BA, English, French	English
3	Korean	5	MATESOL; BA, English	Korean
4	Korean	11	MATESOL	Korean
5	Korean	10	MATESOL	Korean
6	Korean	11	MATEFL	Korean
7	Korean	25	MATESOL, BA, Literature	Korean
8	Korean	16+	Applied Linguistics	Korean
9	Korean	20+	Education	Korean
10				
11	Korean	20	MATESOL/SLA	Korean
12	Korean	8	MATESOL	Korean
13	Korean	16	MA	Korean
14	Korean	8	Ph.D.	Korean
15	English/Korean	15+	SLA	Korean
16	Korean	30	Language/Literature	Korean
17	Korean	12	Chinese	Korean
18	Korean/English/ Japanese	10	MA	Korean
19	Korean	12	BA	Korean
20	Korean	12	MA, Journalism	Korean
21	English	2	MATESOL	English
22	Korean	10	MA	Korean
23	English	1	MATESOL; BA, Philosophy	English
24	N/A	N/A	MATESOL	Spanish
25	English	2	MATESOL; BA, International Relations	English
26	None	None	BA, Linguistics	English
27	English	2	MATESOL; BA, English	English
28	English	10	MATESOL; Journalism, Spanish	English

Appendix B
Questionnaire

You have been asked to complete this questionnaire as part of a research project called "Beyond the Psycholinguistic Model: An Analysis of the Attitudes of Foreign Language Teachers Toward Immediate Written Recall Protocols, Multiple-Choice, True-False, and Cloze-Completion Item Types for Assessing Reading Comprehension." The purpose of this questionnaire is to learn about the methods foreign language teachers are using in their classrooms to assess reading comprehension. With this, I hope to discover teacher's beliefs and attitudes about a framework for assessing reading comprehension known as "immediate written recall protocol."

This protocol process requires students to immediately write down everything they remember after reading a text in the target language without looking back at the original text. The "immediate written recall protocol" differs from summaries because in writing a summary, the students may reread the original text.

Your responses are entirely voluntary, and you may refuse to complete any part or all of this questionnaire. This questionnaire is designed to be anonymous, meaning that there should be no way to connect your responses with you. Toward that end, please do not sign your name to the questionnaire or include any information in your responses that makes it easy to identify you. By completing and submitting the questionnaire, you affirm that you are at least 18 years old and that you give your consent to participate in this research. If you have any questions about this research before or after you complete the questionnaire, please contact <AUTHOR>.

Directions: *Please carefully read each of the questions on all three pages of this questionnaire. Then choose the answer you feel most appropriate and provide a written response. If additional space is needed, you may continue your response(s) in the blank space provided on page four.*

By proceeding with the questionnaire, you agree to participate in this study.

1	Have you ever used immediate written recall protocols to assess reading comprehension in your classes?	Yes	No	Don't Know
2	Why or why not?			

3	Multiple-Choice items provide a very good measure of reading comprehension.								
	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree
	1	2	3	4	5	6	7	8	9
4	Please explain your response.								

5	Grading students' summaries of written texts is too time-consuming.								
	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree
	1	2	3	4	5	6	7	8	9
6	Please explain your response.								
7	True-False items provide a very good measure of reading comprehension.								
	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree
	1	2	3	4	5	6	7	8	9
8	Please explain your response.								

9	Fill-in-the-Blank items provide a very good measure of reading comprehension.								
	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree
	1	2	3	4	5	6	7	8	9
10	Please explain your response.								
11	Cloze items provide a very good measure of reading comprehension.								
	Strongly Disagree		Disagree		Neutral		Agree		Strongly Agree
	1	2	3	4	5	6	7	8	9
12	Please explain your response.								
13	What foreign language(s) do you teach?								
14	How many years have you taught foreign language(s)?								
15	In what field is your university degree?								
16	What is your L1?								

Use the space below to continue your responses (if needed):

About the author:
Kenneth J. Boyte is an ESL instructor at Cabrillo College in the San Francisco Bay area and a graduate of the Middlebury Institute of International Studies at Monterey (MA, TESOL). In addition to ESL, he has a background in educational publishing (MA, Journalism, Southern Illinois University; BA, Journalism, Auburn University).

Acknowledgement:
I would like to thank my wife, Sook-Kyoung Boyte, for informing me about immediate written recall protocols, recruiting participants for this study, and distributing/collecting data-collection instruments. Additional thanks to Dr. Kathleen Bailey, Dr. John Hedgcock, Dr. Thor Sawin, and my classmates at the Middlebury Institute of International Studies for their help, as well as the editorial team at TIJ for their guidance throughout the publication process.