A META-ANALYSIS OF SECOND LANGUAGE CLOZE TESTING RESEARCH

YUKIKO WATANABE & DENNIS KOYAMA

University of Hawai'i at Mānoa

INTRODUCTION

Cloze procedure first appeared in 1953 when Wilson Taylor researched its effectiveness as a procedure for estimating the readability of textbooks for school children in the United States. A decade later, research began to appear on the usefulness of cloze for testing the reading proficiency of native speakers of English (e.g., Bormuth, 1965, 1967; Crawford, 1970; Gallant, 1965; or Ruddell, 1964). In the 1960s and 1970s, a number of studies focusing on second language learners emerged and the usefulness of cloze procedure as a measure of overall ESL proficiency was examined (see Alderson 1978; Oller 1979; Cohen 1980 for summaries of this early ESL research). Since then, the cloze procedure has gained popularity and is a commonly used test in many language programs.

The literature on cloze testing now spans more than half a century of research and covers a diverse range of topics. Researchers have manipulated an assortment of variables including, but not limited to: deletion patterns (for example, every *n*th word, or rational deletion), starting point of deletions (for example, after the first sentence, or prescriptively at the *n*th word), scoring methods (such as exact answer or acceptable answer), text difficulty (using a readability index, or the source of text; see, Brown, 1993), and the number of items—to name just five of the many variables that have been researched.

With such a myriad of research topics both within and among studies, it not surprising that the field of cloze test research has produced varying results. One noteworthy topic is that of reliability. The reliability estimates for various cloze tests, over the years, have spanned the reliability spectrum from both weak 0.31 (Brown, 1982) to very strong 0.95 (Brown 1978) estimates. Many cloze studies (e.g., Alderson, 1979a, 1979b 1980; Bachman, 1985; Brown, 1980, 1983, 1984, 1988b, 1989, 1993, 1994, 1998, 2002; Brown, Yamashiro, & Ogane, 1999, 2001; Conrad, 1970; Darnell, 1970; Hinofotis, 1980; Irvine, Atai, & Oller, 1974; Jonz, 1976; Mullen,

Second Language Studies, *26*(2), Spring 2008, pp. 103-133.

1979; Oller 1972a, 1972b; Oller & Inal, 1971; Pike 1973; Revard, 1990; Stubbs & Tucker, 1974) have investigated how different approaches to creating, scoring, and interpreting, cloze tests could be used to maximize their reliability, validity, or both. This trend shows that much research on cloze procedures has focused on cloze procedures as a test, whether a diagnostic test, a placement test, a proficiency test, or otherwise. The lively research brewing on the crafting of cloze tests, caused some researchers to argue about the constructs that cloze items were primarily tapping. Some researchers argued that cloze tests assessed student abilities to handle clause or sentence level grammar (e.g., Alderson, 1979a; Markham, 1985; Porter, 1983), while other researchers were arguing that cloze items measure at the intersentential level (e.g., Bachman, 1985, Brown, 1983, 1994; Chavez-Oller, Chihara, Weaver, & Oller, 1985; Chihara, Oller, Weaver, & Chavez-Oller, 1977; and Jonz, 1987, 1990).

Use of Meta-analysis for Informing Test Design

As mentioned above, cloze test studies have produced a variety of contradictory results. One way to untangle these contradictions is to systematically analyze the characteristics of the primary studies and identify moderating or confounding variables; this approach is often called research-synthesis[1]. The accumulated information (e.g., effect sizes, descriptive statistics, and reliability estimates) from primary studies will provide an extensive picture of the relationships among variables—and in some cases—provide direction and the magnitude of the effects across studies (i.e., meta-analysis). Given the popularity of the use of cloze tests for various purposes, it is important for test designers to be aware of different ways to craft and tailor cloze tests, as well as the causes of biased test construction[2] that may lead to spurious results and irresponsible decisions.

Numerous studies have accumulated over the last forty years in which second language testing researchers have examined various cloze test characteristics, including what cloze tests are measuring, under what conditions, and for what type of learners. In order to understand the

[1] Meta-analysis and research-synthesis have been widely used across many fields, including but not limited to medicine, education, psychology, and more recently in second language acquisition research (see Norris & Ortega, 2006 for a comprehensive overview and various meta-analytic and research-synthesis studies on language learning and teaching).

[2] The Association of Language Testers in Europe (1990) succinctly defines test bias concerns as follows: "A test or item can be considered to be biased if one particular section of the candidate population is advantaged or disadvantaged by some feature of the test or item which is not relevant to what is being measured" (p. 136).

accumulated knowledge of cloze procedures in second language testing research, the following research questions were posed:

1. What are the test and learner characteristics in the primary cloze test studies to date?
2. How does the deletion pattern affect the reliability of the cloze results?
3. How do various scoring methods (exact, acceptable, and clozentropy) affect the reliability of the cloze tests?
4. What is the strength of relationship between acceptable and exact scoring methods?

The following section describes the study identification procedures, the study inclusion and exclusion criteria, and the coding system. The findings from the above research questions are presented in two sections. The first section summarizes the results of the research synthesis, identifying the characteristics of the study, the test, and the participants, while the second section delineates the results of the meta-analysis of 33 studies.

METHOD

The Literature Search

We screened three literature databases, Education Resources Information Center (ERIC), Linguistic and Language Behavior Abstracts (LLBA), and PsychInfo, to identify empirical studies on cloze procedures since 1953, the year of Taylor's seminal article. The following search terms were utilized: "Cloze procedure" (descriptor) AND "language test" NOT "dyslex*, disab*." Since the cloze procedure is often utilized for testing learning disability (e.g., dyslexia), we excluded studies that has "dyslex*" and "diab*" as keywords to avoid introducing additional learner variables. The database search identified a total of 114 studies including 16 dissertations (See Table 1). To assure an exhaustive search, we also manually searched *Language Testing Journal*, locating two more studies. An additional 96 studies were discovered through footnote chasing, resulting in total of 212 studies.

Table 1

Literature Database Search

	LLBA	ERIC	PsychInfo	Total without overlap
(Cloze procedure) AND (language test) NOT (dyslex, disab)	61 (3)	20 (0)	39 (13)	114 (16)

Note. ( ) indicates the number of dissertations in addition to the journal articles.

The Inclusion and Exclusion Criteria

The following four criteria were set to determine whether the retrieved studies qualified for inclusion in the research synthesis and meta-analysis:

1. The study is an empirical study published between 1957 and March, 2007.
2. The study participants are adult English as a second or foreign language learners. The scope of the current study is strictly limited to English as a target language, since English cloze tests are most studied among other languages. Because it is not clear whether cognitive development affect cloze test results, we decided to only include adult learners and avoid populations under 18 years of age.
3. The study used at least one cloze test. Here, we define cloze test as a test that has certain words deleted from a passage. We did not include gap-filling test that has only one sentence as a stem, such as section two in the computer-based TOEFL (also see example below).

   Because of the storm and rough waves, it would be foolish to go out sailing today in a small ____.

   a) automobile b) house c) boat d) beast

   (Perkins, & Linnville, 1987, p. 128)
4. The study adequately described the cloze test employed in the study, so test characteristics can be coded.

Studies were excluded from the analysis, if any of the following criteria were met:

1. The study was published as conference proceedings or in-house publication (working paper), or was an unpublished manuscript (master's thesis or dissertation).
2. The study employed Analysis of Variance, and did not report any descriptive statistics.
3. The cloze test was used just to determine learner characteristics (e.g., grouping learners into high and low proficiency groups), and was not the main focus of the study.

Based on the inclusion and exclusion criteria, 33 studies remained for the research-synthesis and meta-analysis. After determining the number of studies to be included in the meta-analysis, we set a coding scheme for extracting study information (research questions and study setting), learner characteristics (number of participants, L1, and language proficiency), test characteristics (text length, number of items, response type, deletion pattern, starting point of deletion, and scoring criteria), and test results (descriptive statistics and reliability estimates). Table 2 summarizes the coding scheme we utilized. The interrater reliability for two raters on four independently read research studies was 93.5%. Any inconsistencies and discrepancies were resolved though discussions and the end agreement percent was 100%.

Table 2

Coding Scheme

Study information	Learner characteristics	Test characteristics	Quantitative data
- Author(s) & publication year - Research question(s) - Study setting (EFL or ESL)	- Number of participants - First language - Age - English proficiency (researchers' description on proficiency)	- Text length (total # of words) - Text difficulty (Flesch-Kincaid, etc.) - Response type (open-ended or multiple choice) - Total number of items per passage. - Deletion pattern (n^{th} word deletion, rational deletion) - Starting point of deletion - Scoring criteria (exact, acceptable, clozentropy)	- Descriptive statistics (mean, standard deviation, min, max) - Reliability estimate (split half Spearman-Brown coefficient, KR-20, KR-21, Cronbach's alpha)

RESULTS AND DISCUSSION

Research Synthesis on Study Characteristics

Learner characteristics. To identify the proficiency levels stated by the researchers, we adapted the four proficiency categories suggested by Thomas (1994, 2006): impressionistic judgement (IJ), institutional status (IS), in-house assessment (IH), and standardized test (ST). In addition to these four classifications of proficiency, we added two others[3], experience (EX), and self-assessment (SA). Following Thomas's guidelines, impressionistic judgement means the author has suggested an second language (L2) ability based on the author's impression, for

[3] For studies that did not clearly articulate learner proficiency, we labelled them as 'no mention' (NO).

example, “The examinees were intermediate-level”. This label of ‘intermediate-level’ is not usually qualified by any concrete evidence and is therefore considered impressionistic. Institutional status is similar to impressionistic judgement in that there is no evidence that indicates learner proficiency level, rather an association of ability and institutional status is made by categorizing students, for example, “The students in this study were English majors”. This sort of labelling is often combined with IJs, such as, “The examinees in this study were second year English majors and had an intermediate-level of English”. In-house assessment was used to identify studies that offered proficiency assessments based on institutional or in-class proficiency exams, for example, “The students took a placement exam that consisted of an essay test.” The standardized test classification was used when a researcher noted students’ abilities on widely used standardized exam, such as, “The students had a range of 300 to 500 on the TOEFL exam.” Students’ language proficiency was categorized as EX when the study interpreted their learning history through the students’ experiences, for example, “All the examinees had spent at least 2 years abroad.” Self-assessment was used when researchers collect information from the students through interviews or questionnaires. Students may have had to answer a question item on a survey like, “At which level would you rank your English ability: advanced, intermediate, or beginning.”

Table 3 describes and summarizes the learner characteristics. This description includes details about the number of participant subgroups studied, study settings, (ESL, EFL), L1s, proficiency assessments of the subjects, and the number of participants. Across the 33 studies analyzed in this meta-synthesis, 17 studies were in ESL settings and 15 studies in EFL settings. One study (Oller & Inal, 1971) was conducted both in ESL and EFL settings. In total, there were 44 independent distinct subgroups across the 33 studies. By *distinct*, we mean studies that collected data at clearly different times and administrations (e.g., fall term students, spring term students, etc.), or studies that collected data on wholly separate populations in different countries, institutions, or locations (e.g., data collected in Japan compared with data collected in Peru). Seven studies had multiple sub-groups within one study. Within the 44 sub-groups, there were 23 sub-groups of ESL learners and 21 sub-groups of EFL learners.

The most common L1 classification was “various” with ten studies followed by: Japanese with five sub-groups, Chinese and Iranian with three sub-groups each, and Spanish had two sub-groups. Other L1s (in alphabetical order) included: Farsi, German, Indo-European, other S.E.

Asian, Turkish, and Vietnamese. The participant size per sub-group varied greatly, ranging from as little as six to as many as 14,613. The average participant group size was 647.86, with a majority of the sub-groups (*n* = 28) with 101 to 500 participants, followed by studies with 100 people or less (*n* = 10), and finally, studies with more than 500 participants (*n* = 6).

Table 3

Learner Characteristics

Context	*N*	Participant subgroups	*N*[4]	L1[a]	*N*	Proficiency[b]	*n*	Participants	*N*
ESL	17	Studies with 1 subgroup	26	various	10	NO	8	101-500	28
EFL	15	Studies with over 1 subgroup	7	Japanese	5	IH	5	20 (*Min*)-100	10
ESL/EFL	1			Chinese	3	ST	3	500-14613 (*Max*)	6
				Iranian	3	NO/EX	3		
				Spanish	2	IS	2		
				Other[C]	6	EX	2		
						IS/ST	2	(Average)	(647.86)
						IS/EX/ST	1		
						IH/EX	1		
						IH/EX/SA	1		
						IH/IJ	1		
						IJ	1		
						IJ/EX	1		

[a] Number of L1 subgroups across all studies that indicated by the authors.
The following L1s were reported once: Farsi, German, Indo-European, Other S.E. Asian, Turkish, Vietnamese.
[b] Proficiency abbreviations are: institutional status (IS), in-house assessment (IH), standardized test (ST), impressionistic judgment (IJ), experience (EX), self-assessment (SA), no mention (NO).

Test characteristics. In addition to the characteristics of students, the characteristics of each independent cloze test were noted. In particular, the number of studies that reported test level characteristics such as: response type (e.g., open-ended or multiple-choice), deletion pattern (e.g., n^{th} deletion, or rational), scoring method (exact, acceptable, or clozentropy), text length (i.e., total number of words in the test), total items (i.e., the number of cloze items), and readability indexes (e.g., Flesch, Flesch-Kincaid, Gunny Fog, or Fry).

There are essentially two response types for cloze tests, open-ended (OE) and multiple-choice (MC). These response types are inextricable linked to scoring methods. For an open-ended cloze test, there are three ways to score the test: exact scoring, (EX; i.e., the response must

[4] *N* indicates the number of studies, whereas *n* represents distinct sub-groups within a study. For example, if three classrooms take the same cloze test, these are not distinct groups for research purposes and are counted as one group with an aggregated population size. Conversely, a study that collects cloze test results from separate administrations (i.e., to different students, such as an entrance exam across two separate terms) is considered a multi-group study.

match the original phrasing of the given blank), acceptable scoring (AC; i.e., the response must be grammatically and/or semantically acceptable response, often judged by a native speaker), and clozentropy scoring, which is a scoring system developed by Darnell (1970). For clozentropy scoring (CE), a corpus of answers is developed and the frequency of each answer on the corpus is calculated. Then, this information is cross-checked by administering the same cloze test to native speakers and compiling those responses. The answers are, finally, weighted according to a logarithm of the frequency of each response. Multiple-choice responses, however, necessarily must be EX answer method, as it makes little sense to construct a test where the responses include a list of acceptable answers. Notably, though, the matter of deletion is divorced from the response type, and therefore is not affected by the method of scoring.

The primary forms of deletion are n^{th} deletion pattern and the rational deletion. The n^{th} deletion is a pseudo-random form of deletion wherein the test creator selects a number say, six, and eliminates every sixth word from the passage. This means, however, the lower the number, the greater the likelihood that more than one word will be deleted from a single sentence. However, some authors account for this issue and consider other mitigating factors when deleting a word. For example, Kobayashi (2002) used an n^{th} word deletion, but avoided deleting proper nouns and numbers. If the 13^{th} word was a proper noun or number, the previous word (i.e., the 12^{th} word) was selected for deletion. The last form of deletion, rational deletion, is patterned targeting specific linguistic forms or words. For example, Oller and Inal (1971) deleted every other preposition in the construction of their test.

Table 4 shows the count of test characteristics for independent tests across all of the studies (k = 157). [5] The most frequently used response type was open-ended with 139 tokens, and next was multiple-choice (k = 13). One unique study by Hinofotis and Snow (1980) had two response types within each of two independent tests. Their article indicates that two test forms (Form A and Form B) had 50 items for each passage and each form was constructed so that in Form A "the first 25 items were MC and the second 25 were open-ended. On Form B the order was reversed" (Hinofotis & Snow, 1980, p. 130). This means one fifty-item passage had two response types. The most common scoring method was EX with 75 tests, next was the combined methods of EX and AC (k = 50) which had 50 independent tests, and the third most common

[5] k denotes the number of independent test.

format of scoring was AC (k = 15). Thirteen authors chose not to state the method of scoring, and one study (Brown, 1993) used all three methods of scoring, AC, EX, and CE.

There were 15 different styles of deletion, and 12th word deletion was the most prevalent with 55 independent tests utilizing this format. The second most common deletion pattern, 7th word, had 25 cases, followed by rational deletion (20 cases). One unique study, Brown (1988), used a range of deletion patterns where words were deleted in such a way that blanks were never closer than every fifth word and no farther apart than every ninth word. Another study (Jonz, 1976) did not state the deletion pattern.

Table 4

Tally of Test Characteristics across Independent Tests

Response type	*k*	Scoring	*k*	Deletion pattern	*k*
OE	143	EX	75	12th[a]	55
MC	13	EX/AC[b]	50	7th	25
OE & MC	2	AC	15	Rational	20
		No mention	12	10th	13
		EX/CE	3	11th	8
		AC/CE/EX	1	13th	8
				6th	7
				8th	6
				15th	3
				5th	3
				9th	3
				16th	1
				18th	1
				4th	1
				5w - 9w	1
				No mention	1

[a]Brown (1993) had 50 independent tests using 12th deletion, and one study (Bachman, 1982) noted the deletion pattern was on average 12th.
[b]Brown (1993) had 50 independent tests using this system of scoring.

A closer examination of the studies reporting on the make-up of the test is summarized in Table 5. This table identifies the total number of tests that state the length of text and total number of items on the test. Descriptive statistics particular to these characteristics is also given. There were a total of 69 tests that gave the length of the cloze passage (or the length of the passage was calculated from the test in the appendix of the article, as in the case of Chapelle (1988) and Chapelle and Abraham (1990)), additionally, 137 tests reported the number of total cloze items. The range of text length was 125 words to 750 words with an average of

approximately 374 words per passage, and the standard deviation was 126.53 words. Of the 137 studies that reported the number of cloze items, the range was 15 to 80, and the average number of items was 34.34 with a standard deviation of roughly ten cloze test items. Although this study set out to list the common readability indexes across the studies, very few authors offered a readability estimate of their cloze passage. However, of the studies that did give a readability index, the most common readability given was the Flesch (n = 4), followed by the Fry index (n = 3). The Gunning-Fog and the Flesch-Kincaid were noted only once. One study by Brown, Yamashiro, and Ogane (2001) rated the passage as coming from an introductory reading text, and another study (Mullen 1979) described the level given in the manual from which the text was extracted as approximately 7th and 12th grades.

Table 5

Tally of Text Characteristics across Independent Tests

	Number of test reporting	*Range*	*M*	*SD*
Text length	69	125 - 750	373.71	126.53
Total items	137	15 - 80	34.34	10.28

Reliability

The concept of reliability estimates as a measure of a given tests consistency in measuring a particular construct (or, multiple constructs) has a long history. The most common formulas for estimating test reliability are: K-R20, K-R21, Spearman-Brown, and Cronbach alpha. Table 6 shows the break down of reliability estimates given. By far the most common reliability estimate given was K-R20 with ten studies using it solely, followed by three studies that used K-R21 and three other studies that used both Spearman-Brown (S-B) and Cronbach (α) estimates of test reliability. Stand-alone estimates of reliability were limited to two studies that used α, and another study that used S-B only. Five other studies used various combinations of reliability to give the reader a broader perspective of the test's reliability. For example, one study used both K-R21 and α while another study reported all four forms of test reliability estimates. The most frequently used estimates were as follows: K-R20, (used alone or with another estimate in 13 studies), Spearman-Brown and Cronbach α (which were used alone or with other estimates in seven studies), and K-R21 (used alone or with another estimate in four studies).

Table 6

Reliability Estimates Used by Studies

Type(s) of reliability specified [a]	*N*
K-R20	11
Cronbach alpha (α)	3
S-B & K-R20	3
Spearman-Brown (S-B) & Cronbach's alpha	2
K-R21	2
K-R21 & Cronbach's alpha	2
S-B, K-R20, Split-half	1

[a] One unique study reported Guttman's lower bounds to reliability estimates, and one study offered no name for the reliability estimate.

Meta-analysis on Cloze Test Reliability Estimates

A total of three statistical analyses were conducted analyzing the reliability estimate by scoring system and by deletion pattern. When performing multiple statistical tests, a family-wise error rate of $p < 0.05$ is maintained within a study to reduce false positives (i.e., Type I error) among a class (family) of tests, under the null hypothesis. Since the current study performed three statistical tests, an *a priori* decision was made to report data for a dependent variable as statistically significant only when multivariate *F* values and the *t*-test value were significant at $p < 0.017$ ($p < 0.05$ divided by three, as dictated by the Bonferonni adjustment).

Scoring system and reliability. A total of 223 reliability estimates (across 24 studies) associated with each scoring method were analyzed. From the descriptive statistics (see Table 8), the values for exact scoring were more widely dispersed than those for any other methods. This can be attributed to the few studies that used tests with very low reliability. For example, Farhady and Keramati (1996) used a structure-driven deletion cloze test which produced *K-R21* = 0.14, 0.23, and 0.40. Brown's (1983) study also contributed to the large variability in the overall exact scoring method, since his study found nine cloze tests that produced reliability estimate equal to or less than $\alpha = 0.50$.

A one-way analysis of variance (ANOVA) was conducted with reliability as a dependent variable and scoring method (acceptable, exact, and clozentropy) as an independent variable to investigate whether there were any significant differences in reliability among scoring methods. Across 223 cases of reliability estimates extracted from 24 studies, one way ANOVA revealed that the scoring system makes a difference in the reliability results, $F(2, 220) = 16.06$, $p = 0.001$.

A post hoc Tukey HSD test indicated that the exact scoring method produces the lowest reliable among the three scoring methods, and that the acceptable and clozentropy differences are statistically non-significant (see Table 9). However, these results have to be interpreted with caution, since there were a number of cases where clozentropy was less than other two scoring methods. As can be seen in Figure 1, the error associated around the mean in the 95% confidence interval was larger in magnitude for clozentropy than EX and AC scoring methods.

Table 8

Descriptive Statistics of the Reliability Estimates across Scoring Methods

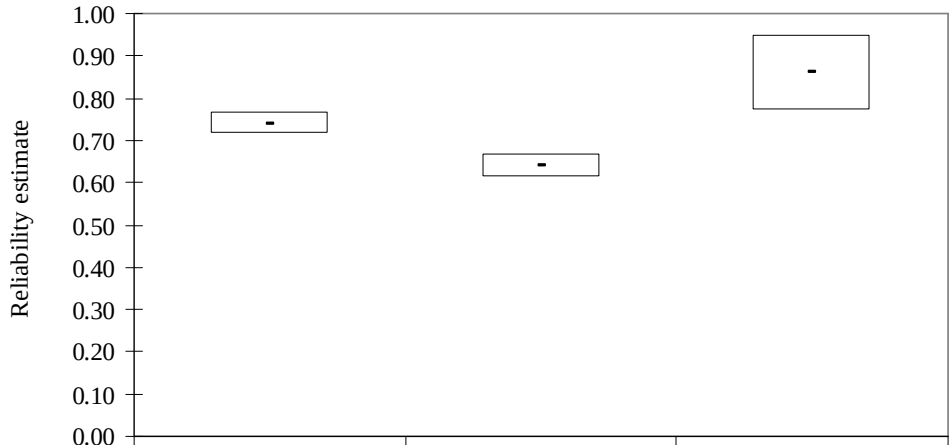
	k	*M*	*SD*	*Min*	*Max*
Acceptable	97	0.74	0.12	0.60	0.97
Exact	122	0.64	0.16	0.14	0.99
Clozentropy	4	0.86	0.06	0.78	0.91

Table 9

Post-hoc Tukey HSD on Scoring Methods

Scoring comparison	Mean difference	*p*
AC-EX	0.10	0.001*
CZ-EX	0.22	0.008*
AC-CZ	-0.12	0.241

*$p < 0.017$



	Acceptable (k = 97)	Exact (k = 122)	Clozentropy (k = 4)
95% C.I. lower bound	0.72	0.61	0.77
- Mean reliability	0.74	0.64	0.86
95% C.I. upper bound	0.77	0.67	0.95

Figure 1. 95% Confidence Interval Around Mean Reliability for Each Scoring Method

A more fine-tuned analysis was performed with studies that compared different scoring systems (N = 12). However, since only two studies (Brown, 1980; Pike, 1979) investigated the reliability difference between clozentropy and other scoring methods, we included only 10 studies that utilized both the exact and acceptable scoring methods when scoring the same cloze test.

The test characteristics (e.g., total number of items, deletion pattern), number of participants for each cloze test, descriptive statistics, and the reliability estimates are summarized in Table 10. The average difference between EX and AC scoring methods was 0.068, the range was from -0.11 to 0.43. A paired t-test was used to determine if the scoring systems made a difference in test's reliability. The analysis revealed that the AC scoring system produced greater reliability than the EX scoring system with statistical significance, $t(37) = 4.10$, $p = 0.001$. This result is reasonable, given that the AC scoring will result in larger means and standard deviation (Overall EX: M = 14.86, SD = 3.78; AC: M = 22.25, SD = 4.61). Across 38 independent tests, the overall relationship between the two scoring methods and reliability was r = 0.793 (p = 0.001). Therefore, 63 percent of the variance in the reliability in one scoring method (EX or AC) can be predicted by the other scoring method (EX or AC). In sum, the results in this section

indicated that AC scoring is a more reliable scoring method than exact scoring, which is in line with the findings of Brown (1980, 1983).

Table 10

Reliability Difference between Exact and Acceptable Scoring

Author	Test ID	# items	Delet. Pattern	*N*	Exact			Acceptable (sem/gram)			Reliability (AC-EX)
					M	*SD*	*Relia-bility*	*M*	*SD*	*Relia-bility*	
Brown (1980)		50	7th	55	15.00	8.54	0.90	25.58	12.45	0.95	-0.04
Brown (1983)		50	7th	66	21.21	4.31	0.61	32.99	4.98	0.67	-0.06
Brown, Yamashiro, & Ogane (2001)	Original	30	11th	40	1.58	2.26	0.76	1.80	2.33	0.74	0.02
	Orig+1R	30	12th	38	2.53	2.27	0.66	2.92	2.45	0.66	0.00
	Orig+2R	30	13th	39	0.72	1.32	0.64	1.64	1.71	0.53	0.11
	Orig-1L	30	10th	38	1.53	1.67	0.56	2.13	1.76	0.48	0.08
	Orig-2L	30	9th	38	3.50	3.22	0.79	4.05	3.32	0.77	0.02
Hinofotis (1980)		50	7th	107	11.90	2.08	0.61	15.30	7.30	0.85	-0.24
Jafarpur (1996)		25	6th	325	8.96	3.60	0.75	15.92	5.60	0.88	-0.13
Jonz (1987)[b]	Reg_fam	50	15th	53	27.28	3.38	0.89	38.66	3.83	0.89	0.00
	Reg_unfam	50	15th	59	23.22	2.42	0.70	38.64	3.39	0.89	-0.19
	Coh_fam[c]	50	Rational	53	31.42	3.95	0.89	41.58	3.07	0.89	0.00
	Coh_unfam[c]	50	Rational	65	16.52	2.98	0.79	34.22	4.40	0.92	-0.13
Klein-Braley (1983)	1	34	9th	23	16.34	3.48	0.62	25.95	3.46	0.63	0.01
	2	46	7th	23	24.74	4.56	0.67	31.44	4.76	0.77	0.10
	3	30	10th	45	14.22	2.50	0.15	21.80	3.15	0.58	0.43
	4	43	8th	45	15.02	3.47	0.46	31.53	4.29	0.60	0.14
	5	35	10th	31	15.97	3.51	0.58	26.58	4.68	0.77	0.19
	6	50	6th	31	21.71	3.98	0.63	40.58	4.69	0.73	0.10
	7	35	10th	29	12.00	2.92	0.41	23.35	3.74	0.59	0.18
	8	50	6th	29	23.62	4.37	0.67	39.41	4.95	0.74	0.07
	9	40	7th	23	20.64	4.21	0.71	31.73	3.94	0.65	-0.06
	10	30	10th	23	16.41	3.31	0.53	23.36	3.17	0.61	0.08
	11	40	7th	53	23.07	4.02	0.74	30.96	4.13	0.70	-0.04
	12	30	10th	53	20.52	2.46	0.41	24.11	2.30	0.50	0.09
Kobayashi (2002)	1	25	13th	63	7.94	2.77	0.62	11.50	3.75	0.72	-0.1
	2	25	13th	66	6.91	3.32	0.72	10.70	4.65	0.80	-0.08
	3	25	13th	61	8.31	4.56	0.82	9.36	4.92	0.84	-0.02
	4	25	13th	65	5.46	3.36	0.73	8.54	4.30	0.79	-0.06
	5	25	13th	63	7.89	3.66	0.75	9.68	4.43	0.80	-0.05
	6	25	13th	66	7.35	3.39	0.73	9.15	4.23	0.80	-0.07
	7	25	13th	61	7.77	3.58	0.73	9.93	4.40	0.80	-0.07
	8	25	13th	65	5.29	2.65	0.62	7.33	3.73	0.77	-0.15
Mullen (1979)	Easy	50	10th	154	---	---	0.83	---	---	0.91	-0.08
	Hard	46	10th	154	---	---	0.73	---	---	0.87	-0.14
Oller (1972)	1	50	7th	132	31.74	6.00	0.99	42.99	6.59	0.96	0.03
	2	50	7th	134	33.85	8.77	0.98	41.52	9.64	0.93	0.05
	3	50	7th	129	22.91	9.17	0.98	34.21	11.49	0.95	0.03

Deletion pattern and reliability. Two hundred twenty-three cases of reliability estimates extracted from 24 studies, were classified based on the deletion pattern of the cloze test. Due to

the small number of cases, fourth to sixth deletion patterns as well as 15th to 18th deletions were treated as one group. Rational deletion and tailored deletion[7] were also separated from random deletion method. Table 11 details the descriptive statistics of the aggregated reliability for each deletion pattern[8]. The most frequently used deletion pattern—across studies that reported reliability—was seventh deletion (N = 10), followed by rational deletion (N = 5). The 12th deletion had the largest number of independent tests with reliability information (k = 54), a majority (k = 50) was contributed by a single study, Brown (1993). The rational deletion had the highest mean reliability (M = 0.80) in opposition to random deletion patterns. The least reliable deletion pattern was the eighth deletion method.

In further analysis, a one-way ANOVA was carried out to assess whether the deletion pattern made a difference in reliability scores. The results revealed an overall statistical significance at $p < 0.017$ level for the effect of deletion pattern on reliability, $F(10, 174) = 4.921$, $p = 0.001$. The post hoc Tukey HSD test demonstrated only few statistically significant differences among deletion patterns at $p < 0.017$ level. The seventh word random deletion test obtained greater reliability than the tenth and twelfth word random deletion tests (see Table 12). Figure 2 depicts the 95% confidence interval around the mean for each deletion pattern.

[7] Tailored items have been deleted due to some tailoring criteria, such as item facility and item difficulty calculations.

[8] The deletion patterns discussed here are deletions noted in research that clearly states reliability estimates per test.

Table 11

Descriptive Statistics of Reliability Scores for Each Deletion Pattern

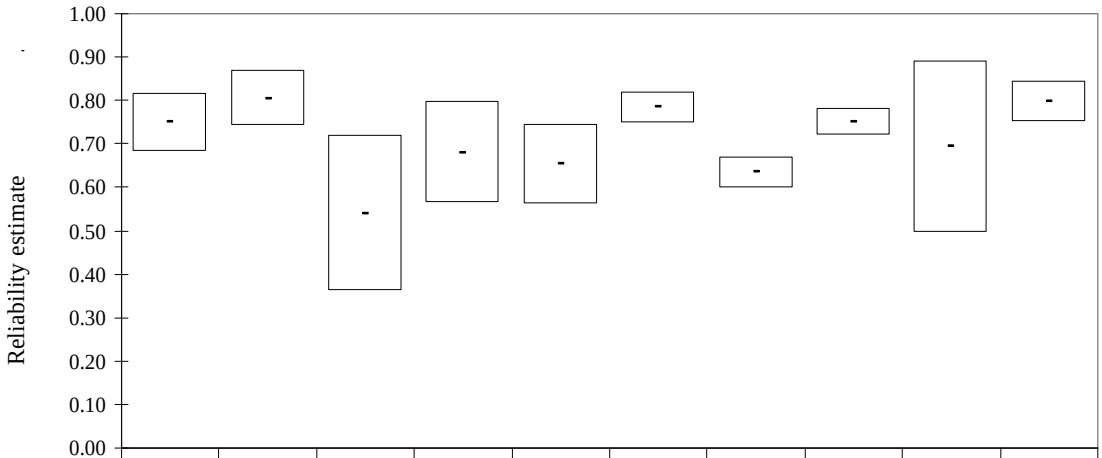
Deletion pattern	*M*	*SD*	*Min*	*Max*
4-6th (k = 9; 4 studies)	0.75	0.09	0.63	0.88
7th (k = 28; 10 studies)	0.80	0.16	0.27	0.99
8th (k = 3; 2 studies)	0.54	0.07	0.46	0.60
9th (k = 5; 3 studies)	0.68	0.09	0.59	0.79
10th (k = 22; 4 studies)	0.65	0.20	0.15	0.91
11th (k = 9; 4 studies)	0.78	0.05	0.74	0.88
12th (k = 54; 3 studies)	0.63	0.13	0.28	0.82
13th (k = 26; 2 studies)	0.75	0.08	0.53	0.84
15-18th (k = 11; 2 studies)	0.69	0.29	0.14	0.92
Rational (k = 15; 5 studies)	0.80	0.08	0.61	0.95

Table 12

Post Hoc Tukey Test on Deletion Pattern (only statistically significant results)

Deletion pattern	Mean Difference	*p*
7th-12th	0.17	0.000*
7th-10th	0.15	0.014*
Rational-12th	-0.16	0.007*

*$p < 0.017$



	4-6th (k= 9; 4 studies)	7th (k = 28; 10 studies)	8th (k = 3; 2 studies)	9th (k = 5; 3 studies)	10th (k = 22; 4 studies)	11th (k = 9; 4 studies)	12th (k = 54; 3 studies)	13th (k = 26; 2 studies)	15th (k = 11; 2 studies)	Rational (k = 15; 5 studies)
95% C.I. upper bound	0.82	0.87	0.72	0.80	0.74	0.82	0.67	0.78	0.89	0.84
Mean reliability	0.75	0.80	0.54	0.68	0.65	0.78	0.63	0.75	0.69	0.80
95% C.I. lower bound	0.68	0.74	0.36	0.56	0.56	0.75	0.60	0.72	0.50	0.75

Figure 2. 95% Confidence Interval Around Mean Reliability for Each Deletion Pattern

DISCUSSION AND CONCLUSION

Through the consistent coding of substantive and methodological features across 33 studies, the research synthesis and meta-analysis provided comprehensive profiles of cloze test research and revealed what test-developers need to consider when creating and implementing a cloze test. This section provides a discussion and suggestions for improving the reporting of research on cloze testing, including: learner and test characteristics and the use reliability estimates.

An equal number of studies addressed both the ESL and EFL populations. The identification of the population types (i.e., ESL, EFL) was not as much an issue as the clear identification of the learner proficiency levels. For example any number of issues and concerns can be raised with any of the statements of learner proficiency used across cloze test studies. Issues like, a researcher who may deem their participants as 'intermediate' allows the reader to determine the amount and quality of the students' English abilities. This can be very difficult to do, especially in the EFL based assessments as countries vary in the level and amount of English instruction. This can be likened to judging the intensity a spicy dish on a menu in a restaurant that uses pictures of chilli peppers to indicate the spiciness of the meal. Simply put, learner proficiency statements seem to lack consistency across studies that are interpretable outside of the primary investigators' frames of reference.

In addition to a lack of consistency in identifying learner characteristics, many studies also lacked detailed descriptions of test designs, scoring information, descriptive statistics, and reliability estimates for the tests. Recall that among 158 independent cloze tests the OE response was the most popular construction for cloze tests, and those responses were scored mostly with the EX scoring method (see Table 4). This preference of style and scoring methods may be a result of practicality involved in the ease of test construction and scoring. However, it is noteworthy to highlight that AC scoring can be perceived as a much fairer way to score cloze tests rather than EX scoring methods that only accept answers that were used by the author of the original text. Finally, the method of CE may have pragmatic limitations to its construction and implementation. This pragmatic concern stems from the underlying reason behind utilizing clozentropy, which measures and weighs the linguistic ability of the test-takers to the target norm (i.e., native speakers). The intrinsic focus on native speaker-like answers may call into question

the reasonableness of expecting non-native speakers to produce similar linguistic frequency patterns as to those produced by native speakers.

Ultimately, choosing which scoring system to adopt is the decision of the test administrators. For they are the ones who have to consider the purpose of the test, feasibility of creating and utilizing a corpus of acceptable answers, errors that may accumulate during the scoring procedures, and any other practical implementation issues.

Another important test design feature is the deletion pattern, which may be decided for any number of reasons. Across the 33 studies, the seventh deletion pattern was the most common deletion patterns (N = 14, k = 25), followed by rational deletions with seven studies using rational deletions (k = 20). Just these two deletion patterns account for approximately 29 percent of the studies analyzed and many of these studies did not cite the text length and total number of items. These characteristics are invaluable to other testing researchers (not to mention test construction artists), yet they frequently go unaccounted for. In addition to providing these important test details it will also be necessary to state, when using rational deletion, the range between deletions (i.e., the minimum and maximum distances between cloze items). In addition to the reporting of these test description details, it would also be useful to always append the actual test to the research report.

In this meta-anlaysis, we acknowledge that the aggregation and categorization of reliability across different tests and studies by scoring pattern or test design pattern alone may introduce other possible reasons for the variability of the test scores. Though we need to interpret the results with extreme caution, we found that among scoring methods, clozentropy, was the least used methodology but obtained the largest reliability estimate among the scoring methods. More studies are needed that specifically look at the relationship between clozentropy and other scoring methods within one study, so as to examine whether the differences in reliability are related to other test features.

The deletion pattern comparison in terms of reliability showed that the seventh word deletion and rational deletion cloze tests are most reliable. Despite the popularity of the seventh word deletion pattern, more empirical studies are needed comparing different deletion patterns in order to draw a more certain conclusion as to which deletion pattern is optimal.

Moving onto a discussion of reliability brings to the fore a very important, and fundamental question: What assumptions do researchers make using different reliability estimates (K-R20, K-

R21, Spearman-Brown split-half reliability coefficient, and Cronbach alpha), for estimating cloze test reliabilities? The presence of various reliability estimates across the studies in the meta-analysis studies ignored a need to better understand what is actually going on with reliability estimates as applied to cloze testing.

On the one hand, the K-R20 formula requires the use of three pieces of information, the number of items, the standard deviation, and (perhaps most importantly) the average variance of the items. On the other hand, the K-R21 formula requires similar types of test information, such as the number of test items, the K-R21 also requires the average of the item scores. The key difference, then, between the two formulas lays in the assumptions underpinning them. For the K-R21 one assumption is "that the matrix of inter-item correlations has a rank of one, that these correlations are equal, and that *all items have the same difficulty*" (Kuder & Richardson, 1937, emphasis added)—an assumption that is probably never met in any reasonable way with cloze test were item difficulty values often vary from .00 to 1.00. Thus the difference between K-R20 and K-R21 is clear: K-R 20 calculations estimate reliability with the use of the average of item variances, whereas K-R21 uses the average of item difficulty (Kuder & Richardson, 1937, pp. 158-159). Therefore, the K-R21 formula may be inferior to the K-R20 formula when calculating the reliability estimate for a given cloze test.

The issue of item difficulty is not limited to the discussion of the K-R21 formula. Item difficulty should also be considered if the test's reliability estimates are to be calculated with the split-half Spearman-Brown coefficient because the difficulty of items in each of the halves of the split test will necessarily affect the reliability estimate given by using the split-half method. That is to say, an uneven split of difficult items and less difficult items may cloud the results of the calculation, or boost reliability. As a consequence, extra steps to ensure an equilibrium between test halves is necessary to produce fair and trustworthy estimates of reliability using the Spearman-Brown split-half coefficient. Therefore, unless these extra steps are taken the Spearman-Brown split-half coefficient may be a less desirable estimate of reliability when compared to the K-R20 formula.

Cronbach alpha estimates how well a set of items (or variables) measures a single unidimensional latent construct. When data have a multidimensional structure, Cronbach alpha will usually be low. Since Cronbach alpha utilizes the Pearson product moment correlation coefficient, it has been argued (see Farhardy, 1983) that Cronbach alpha, when used

on cloze test items, violates the assumption of item independence. Cziko (1978) and Brown (1983) found cloze items to be dependent on previous context, thus we suggest that Cronbach alpha may not be a theoretically appropriate reliability estimate to be used for cloze testing.

Considering the discussion on reliability estimates, it may be credible to assert that the K-R20 is, perhaps, the most trustworthy reliability estimate for cloze testing research. Further research and investigation of this claim is necessary, including how item interdependence may affect test reliability.

This research synthesis and meta-analysis highlights the gaps and insights of cloze testing research; in sum, it is our hope that this study sheds light on the enterprise of cloze testing research and test construction.

APPENDIX A

LEARNER CHARACTERISTICS

Author	Setting				Proficiency		
	Context	Subgroup	*N*	L1	Test name	Range	Judgment[a]
Abraham & Chapelle (1992)	ESL		178		TOEFL	500+	ST/IS
Alderson, C. (1979)	ESL		360				NO
Bachman (1982)	ESL	Fall	316	Various			NO/IS
	ESL	Spring	102	Various			
Bachman (1985)	ESL	incoming student	146				NO/IS/EX
		entering full-time study	605				
		continuing study	61				
Bensoussan & Ramaz (1984) Exp2	EFL	University Freshman	691				NO/EX
Experiment 3	EFL	Entrance Examinees	14,613				NO
Experiment 4	EFL	Entrance Examinees	354				NO
Bensoussan (1984)	EFL	Spring	2075				NO/EX
	EFL	Summer	1420				NO/EX
Briere & Hinofotis (1979)	ESL	UCLA	374	Various			NO
	ESL	USC	208				
	ESL	SIU	107				
Brown (1980)	EFL		112				
Brown (1983)	EFL		125	Chinese	TOEFL	390-590	ST
Brown (1988)	EFL		89	Chinese	TOEFL	369-499[b]	ST
Brown (1993)	EFL		2,298	Japanese			NO/EX
Brown, Yamashiro &Ogane (2001)	EFL		144	Japanese			IH/EX/SA
Chapelle (1988)	ESL		66		essay		IS/IH
Chapelle & Abraham (1990)	ESL		201		TOEFL	500+	ST/IS
Farhady & Keramati (1996)	EFL		403	Iranian			IS
Flahive (1980)	ESL		20	Various	TOEFL	437-568	ST/IH
Hinofotis (1980)	ESL		107	Various			IJ/IH
Hinofotis & Snow (1980)	ESL		66	Various			NO
Ilyin, Spurling, & Seymour (1987)	ESL		257				IS/IJ
Irvine, Atai, & Oller (1974)	EFL		159	Farsi			NO
Jafarpur (1996)	EFL		325	Iranian			NO/EX
Jafarpur (1995)	EFL		325	Iranian			NO/IS
Jonz (1976)	ESL		33	Various			NO
Jonz (1987)	ESL	Regular/cohesive & familiar/unfamiliar	230		TOEFL[c] (n = 100)		IS/(ST)[d]
Jonz (1991)	ESL	Familiar/unfamiliar	238[c]		TOEFL[c] (*n* = 158)		IS/(ST)[d]
Klein-Braley (1983)	EFL		204	German			EX, IJ
Kobayashi (2002)	EFL		255	Japanese			NO/EX
Mackay & Williamson (1979)	EFL		22				NO
Mullen (1979)[e]	ESL		154				IS
Oller & Inal (1971)	ESL	Winter 1970	110	Various			ST/IH/IS
	EFL	EFL Turkish	53	Turkish			NO/EX/IH

Oller (1972)	ESL	University students	398	Various	NO
Pike (1979)	EFL	Peru	95	Spanish	SA
	EFL	Chile	143	Spanish	SA
	EFL	Japanese	192	Japanese	SA
Wilson (1980)	ESL		72	Various	IJ

NOTE. [a] Proficiency abbreviations are: Institutional status (IS), in-house assessment (IH), standardized test (ST), impressionistic judgement (IJ), no mention (NO), experience(EX), self-assessment (SA).
[b] This is an estimate provided by the author, who clearly stated these numbers were *estimated*
[c] The number of students who reported TOEFL scores are not available for total *N* size.
[d] The information on standardized test was only available for approximately half of the students.
[e] Mullen (1979) stated TOEFL as a measure of proficiency for 80 participants; however, no specific scores were provided.

APPENDIX B

TEST CHARACTERISTICS

Author	ID	Text length	Read-ability index	Readability value	Response type	Total items	Deletion pattern	Scoring[a]
Abraham & Chapelle (1992)	1				OE	35	11th	AC
	2				OE	35	Rational	AC
	3				MC	35	Rational	EX
Alderson (1979)	1		Teacher	Difficult	OE		6th	EX, AC
	2			Medium	OE		6th	EX, AC
	3			Easy	OE		6th	EX, AC
	4			Difficult	OE		8th	EX, AC
	5			Medium	OE		8th	EX, AC
	6			Easy	OE		8th	EX, AC
	7			Difficult	OE		10th	EX, AC
	8			Medium	OE		10th	EX, AC
	9			Easy	OE		10th	EX, AC
	10			Difficult	OE		12th	EX, AC
	11			Medium	OE		12th	EX, AC
	12			Easy	OE		12th	EX, AC
Bachman (1982)	1	365			OE	30	Ave.12th	AC
Bachman (1985)	1	330			OE	30	Rational	AC
	2	330			OE	30	11th	AC
Bensoussan (1984)	1	313			OE	26	7th	AC
Bensoussan & Ramaz (1984) Experiment 2	1	300			MC	21	Rational	EX
	2	300			MC	28	Rational	EX
	3	300			MC	24	Rational	EX
	4	300			MC	20	Rational	EX
Experiment 3	1	300			MC	15	Rational	EX
	2	300			MC	17	Rational	EX
Experiment 4	1	300			MC	41	Rational	EX
	2	300			MC	28	Rational	EX
	3	300			MC	28	Rational	EX
	4	313			OE	24	Rational	AC
Briere & Hinofotis (1979)	1	427	Flesch	69.3	OE	50	7th	EX
Brown (1980)	1	399			OE	50	7th	EX, AC, CE
	2	399			MC	50	7th	EX
Brown (1983)	1	399	Fry	Abt 8th grade	OE	50	7th	EX, AC
	2	399			OE	51	8th	EX, AC
	3	399			OE	52	9th	EX, AC
Brown (1988)	1	399			OE	50	7th	EX
	2	399			OE	50	5th – 9th	EX
Brown (1993)	1		F-K	9.6	OE	30	12th	EX
	2		F-K	13.5	OE	30	12th	EX
	3		F-K	4.8	OE	30	12th	EX
	4		F-K	7.6	OE	30	12th	EX
	5		F-K	13.9	OE	30	12th	EX
	6		F-K	7	OE	30	12th	EX
	7		F-K	9.9	OE	30	12th	EX
	8		F-K	11.2	OE	30	12th	EX
	9		F-K	15.3	OE	30	12th	EX
	10		F-K	15.2	OE	30	12th	EX
	11		F-K	5	OE	30	12th	EX
	12		F-K	11	OE	30	12th	EX

	13		F-K	12.1	OE	30	12th	EX
	14		F-K	8.5	OE	30	12th	EX
	15		F-K	12	OE	30	12th	EX
	16		F-K	13	OE	30	12th	EX
	17		F-K	20.4	OE	30	12th	EX
	18		F-K	12.7	OE	30	12th	EX
	19		F-K	10.2	OE	30	12th	EX
	20		F-K	10.8	OE	30	12th	EX
	21		F-K	7.5	OE	30	12th	EX
	22		F-K	10.8	OE	30	12th	EX
	23		F-K	13.9	OE	30	12th	EX
	24		F-K	13.1	OE	30	12th	EX
	25		F-K	10.2	OE	30	12th	EX
	26		F-K	16.6	OE	30	12th	EX
	27		F-K	10	OE	30	12th	EX
	28		F-K	14.4	OE	30	12th	EX
	29		F-K	16	OE	30	12th	EX
	30		F-K	6.5	OE	30	12th	EX
	31		F-K	11.6	OE	30	12th	EX
	32		F-K	9.6	OE	30	12th	EX
	33		F-K	16.3	OE	30	12th	EX
	34		F-K	12.8	OE	30	12th	EX
	35		F-K	4.8	OE	30	12th	EX
	36		F-K	11.3	OE	30	12th	EX
	37		F-K	8.6	OE	30	12th	EX
	38		F-K	12.9	OE	30	12th	EX
	39		F-K	6.7	OE	30	12th	EX
	40		F-K	8.1	OE	30	12th	EX
	41		F-K	14.3	OE	30	12th	EX
	42		F-K	9.1	OE	30	12th	EX
	43		F-K	13.9	OE	30	12th	EX
	44		F-K	13.9	OE	30	12th	EX
	45		F-K	11.1	OE	30	12th	EX
	46		F-K	11.2	OE	30	12th	EX
	47		F-K	11.9	OE	30	12th	EX
	48		F-K	11.2	OE	30	12th	EX
	49		F-K	10.3	OE	30	12th	EX
	50		F-K	21.3	OE	30	12th	EX
Brown, Yamashiro, & Ogane (2001)	1	330	Intro level		OE	30	11th	EX, AC
	2	330	Intro level		OE	30	11th	EX, AC
	3	330	Intro level		OE	30	11th	EX, AC
	4	330	Intro level		OE	30	11th	EX, AC
	5	330	Intro level		OE	30	11th	EX, AC
Chapelle (1988)	1	450			OE	50	7th	AC
Chapelle & Abraham (1990)	1	509			OE	35	11th	
	2	509			OE	35	Rational	
	3	509			MC	35	Rational	EX
Farhady & Keramati (1996)	1	337			OE	80	4th	EX
	2	337			OE	55	5th	EX
	3	337			OE	41	8th	EX
	4	337			OE	35	9th	EX
	5	337			OE	26	12th	EX
	6	337			OE	21	15th	EX
	7	337			OE	20	16th	EX
	8	337			OE	18	18th	EX
	9	337			OE	43	7th	EX
Flahive (1980)	1	400			OE	50	7th	AC
Hinofotis (1980)	1	427			OE	50	7th	EX, AC

Hinofotis & Snow (1980)	1	427			MC & OE	50	7th	EX, AC
	2	427			MC & OE	50	7th	EX, AC
Ilyin, Spurling, & Seymour (1987)	1	393			OE	50	7th	AC
Irvine, Atai, & Oller (1974)	1	394			OE	50	7th	
Jafarpur (1995)	1	125	Flesch	79	OE	25	6th	EX, AC
Jafarpur (1996)	1	193	Flesch	79	OE	25	6th	EX, AC
Jonz (1976)	1	706			MC	33		EX
Jonz (1987)	1	750			OE	50	15th	
	1	750			OE	50	15th	
	2	750			OE	50	Rational	
	2	750			OE	50	Rational	
Jonz (1991)	1				OE	50	7th	AC
	2				OE	50	7th	AC
	3				OE	50	7th	AC
	4				OE	50	7th	AC
Klein-Braley (1983)	1				OE	34	9th	EX, AC
	2				OE	46	7th	EX, AC
	3				OE	30	10th	EX, AC
	4				OE	43	8th	EX, AC
	5				OE	35	10th	EX, AC
	6				OE	50	6th	EX, AC
	7				OE	35	10th	EX, AC
	8				OE	50	6th	EX, AC
	9				OE	40	7th	EX, AC
	10				OE	30	10th	EX, AC
	11				OE	40	7th	EX, AC
	12				OE	30	10th	EX, AC
Kobayashi (2002)	1	357.8[b]	Fry	8.4	OE	25	13th	EX, AC
	2	357.8	Fry	8.4	OE	25	13th	EX, AC
	3	357.8	Fry	8.4	OE	25	13th	EX, AC
	4	357.8	Fry	8.4	OE	25	13th	EX, AC
	5	380.5	Fry	8.2	OE	25	13th	EX, AC
	6	380.5	Fry	8.2	OE	25	13th	EX, AC
	7	380.5	Fry	8.2	OE	25	13th	EX, AC
	8	380.5	Fry	8.2	OE	25	13th	EX, AC
Mackay & Williamson (1979)	1	185			OE	24	7th	
Mullen (1979)	1			7th grade	OE	50	10th	EX, AC
	2			12th grade	OE	46	10th	EX, AC
Oller & Inal (1971)	1				OE	50	Rational	AC
Oller (1972)	1	375	Flesh	100	OE		7th	EX, AC
	2	375	Flesh	77	OE		7th	EX, AC
	3	375	Flesh	69	OE		7th	EX, AC
Pike (1979)	1				OE	25	10th	EX, CE
	2				OE	25	10th	EX, CE
	3				OE	25	10th	EX, CE
Wilson (1980)	1	200[c]			OE		5th	
	2	200			OE		Rational	
	3	200			OE		5th	
	4	200			OE		Rational	

[b] The readability value in Kobayashi's (2002) study was an average of the four cloze test passages taken from the same text.
[c] Approximately 200 words.

REFERENCES

[*Note*. Studies that are included in the Meta-Analysis are indicated with an asterisk (*).]

*Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *Modern Language Journal, 76*(4), 468-479.

Alderson, J. (1978). *A study of the cloze procedure with native and non-native speakers of English.* Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK.

Alderson, J. C. (1979a). Scoring procedures for use on cloze tests. In C. A. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79* (pp. 193-205). Washington, DC: TESOL.

*Alderson, J. C. (1979b). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219-227.

Alderson, J. (1980). Native and non-native speaker performance on cloze tests. *Language Learning, 30,* 59-76.

ALTE. (1990). *Multilingual glossary of language testing terms*. Cambridge: Cambridge University.

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

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

*Bensoussan, M., & Ramaz, R. (1984). Testing EFL reading comprehension using a multiple-choice rational cloze. *Modern Language Journal, 68*(3), 230-239.

Bormuth, J. R. (1965). Validities of grammatical and semantic classifications of cloze test scores. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 283-285). Newark, DE: International Reading Associates.

Bormuth, J. (1967). Comparable cloze and multiple-choice comprehension tests scores. *Journal of Reading, 10*, 291-299.

*Briere, E.,J., & Hinofotis, F.,B. (1979). Cloze test cutoff points for placing students in ESL classes. In E. J. Briere & F. B. Hinofotis (Eds.), Concepts in language testing: Some recent studies (pp. 12-20). Washington D.C.: TESOL.

Brown, J. D. (1978). *Correlational study of four methods for scoring cloze tests*. Unpublished master's thesis, University of California Los Angeles, CA.

*Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal*, *64*, 311-317.

Brown, J. D. (1982). Language testing: Norm-referenced. *Zhongshan Daxue English Studies, 1(*1), 192-204.

*Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 237-250). Rowley, MA: Newbury House.

Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. A. Orem & B. P. Taylor (Eds.), *On TESOL '83* (pp. 109-119). Washington, DC: TESOL.

*Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*, 19-31.

Brown, J. D. (1989). Cloze item difficulty. *JALT Journal, 11*, 46-67.

*Brown, J. D. (1993). What are the characteristics of *natural* cloze tests? *Language Testing, 10*, 93-116.

Brown, J. D. (1994). A closer look at cloze: Validity and reliability. In J. Oller., & J. Jon (Eds.), *Cloze and coherence*. Lewisburg, PA: Associated University Presses.

Brown, J. D. (1998). An EFL readability index. *JALT Journal, 20*(2), 7-36.

Brown, J. D. (2002). Do cloze tests work, or is it just an illusion? *Second language studies: Working papers of the Department of Second Language Studies, University of Hawaii, 21*(1), 79-125.

Brown, J. D., Yamashiro, A., & Ogane, E. (1999). Tailored cloze: three ways to improve cloze tests. *University of Hawaii Working Papers in ESL, 17*(2), 107-129.

*Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The Emperor's new cloze: Strategies for revising cloze tests. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development* (pp. 143-161). Honolulu, HI: University of Hawai'i Press.

*Chapelle, C. A. (1988) Field independence: a source of language test variance? *Language Testing, 5*, 62-82.

*Chapelle, C. A., & Abraham. R. G. (1990). Cloze method: What differences does it make? *Language Testing, 7*, 121-146.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning, 35*, 181-206.

Chihara, T., Oller, J. W., Jr., Weaver, K. A., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning, 27*, 63-73.

Cohen, A. (1980). *Testing language ability in the classroom*. Rowley, MA: Newbury House.

Conrad, C. (1970). *The cloze procedure as a measure of English proficiency. Unpublished m*aster's thesis, University of California Los Angeles, Los Angeles, CA.

Crawford, A. (1970). *The cloze procedure as a measure of reading comprehension of elementary level Mexican-American and Anglo-American children.* Unpublished doctoral dissertation, University of California Los Angeles, Los Angeles, CA.

Darnell, D. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. *Speech Monographs, 37,* 36-46.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. *TESOL Quarterly, 16,* 43-59.

*Farhady, H., & Keramati, M. N. (1996). A text-driven method for the deletion procedure in cloze passages. *Language Testing, 13*(2), 191-207.

*Flahive, D. E. (1980). Separating the g factor from reading comprehension. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 34-46). Rowley, MA: Newbury House.

Gallant, R. (1965). Use of cloze tests as a measure of readability in the primary grades. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 286-287). Newark, DE: International Reading Associates.

*Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 121-128). Rowley, MA: Newbury House.

*Hinofotis, F. B., Snow, B. G. (1980). An alternative cloze testing procedure: Multiple-choice format. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 129-133). Rowley, MA: Newbury House.

*Ilyin, D., Spurling, S., & Seymour, S. (1987). Do learner variables affect cloze correlations? *System, 15,* 149-160.

*Irvine, P., Atai, P., & Oller, J. W. Jr. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning, 24,* 245-252.

*Jafarpur, A. (1995). Is C-testing superior to cloze? *Language Testing, 12*(2), 194-216.

*Jafarpur, A. (1996). Native speaker performance validity: In vain or for gain? *System, 24*(1), 89-95.

*Jonz, J. (1976). Improving on the basic egg: The M-C cloze. *Language Learning, 26,* 255-256.

*Jonz, J. (1987). Textual cohesion and second language comprehension. *Language Learning, 37,* 409-438.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24*(1), 61-83.

*Jonz, J. (1991). Cloze item types and second language comprehension. *Language Testing, 8*(1), 1-22.

*Klein-Braley, C. (1983). A cloze is a cloze is a question. In J. Oller (Ed.), *Issues in language testing research* (pp. 218-228). Rowley, MA: Newbury House.

*Kobayashi, M. (2002). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *Modern Language Journal, 86*(4), 571-586.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

*Mackay, R., & Williamson, B. (1979). A trial use of cloze test procedure in the assessment of reading comprehension. *Studia Anglica Posnaniensia, 10,* 211-220.

Markham, P. L. (1985). The rational deletion cloze and global comprehension in German. *Language Learning, 35,* 423-430.

*Mullen, K. (1979). More on cloze tests as tests of proficiency in English as a second language. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 21-32). Washington, DC: TESOL.

Norris, J. M., & Ortega, L. (2006). *Synthesizing research on language learning and teaching.* Philadelphia, PA: John Benjamins.

Oller, J. W. (1972a). Dictation as a test of ESL proficiency. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 346-354). New York: McGraw-Hill.

*Oller, J. W. (1972b). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal, 56,* 151-158.

Oller, J. (1979). *Language tests at school: A pragmatic approach.* London: Longman.

*Oller, J. W. Jr., & Inal, N. (1971). A cloze test of English prepositions. *TESOL Quarterly, 5,* 315-326.

Perkins, K., & Linnville, S. E. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing, 4*(2), 125-141.

*Pike, L. W. (1979). An evaluation of alternative formats for testing English as a second language. *TOEFL Research Report 2*. Princeton, NJ: ETS.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63-74). London: Academic Press.

Revard, D. (1990). Tailoring the cloze to fit: Improvement of cloze tests through classical item analysis. Unpublished scholarly paper. University of Hawaii at Manoa, Honolulu, HI.

Ruddell, R. (1964). A study of the cloze comprehension technique in relation to structurally controlled reading material. *Improvement of Reading Through Classroom Practice, 9*, 298-303.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of ESL proficiency for Arab students. *Modern Language Journal, 58*, 239-241.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 414-438.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44*, 307-336.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris, & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279-298). Philadelphia, PA: John Benjamins.

*Wilson, C. B. (1980). Can ESL cloze tests be contrastively biased?—Vietnamese as a test case. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 208-215). Rowley, MA: Newbury House.