

*Special Issue: Ethics and Integrity, Artificial Intelligence, Information and Communication Technologies (ICTs)
and disinformation in Humanities research*

Seventy Years of Cloze Testing: a review

Setenta anos da técnica cloze: uma revisão

Setenta años de la técnica cloze: una revisión

David Špetla*

 <https://orcid.org/0000-0001-7517-9031>

Tomáš Foltýnek**

 <https://orcid.org/0000-0001-8412-5553>

Abstract: The cloze test was originally put forward as a method of measuring text readability. It soon, however, found other uses in various fields of research. The purpose of this literature review is to map the applications the cloze test has found over the 70 years since its inception. The review uncovers six broad areas of use: text readability, students' reading ability and comprehension, language assessment and language learning, cognition and psychiatry, plagiarism detection, and, most recently, training and evaluation of language models. Each application necessitates a specific method for item deletion, varying answering formats, and distinct scoring strategies. The review delves into the academic discussions surrounding the cloze test, thoroughly examining both the criticisms and the advantages it offers in its varied uses. This ongoing dialogue and continuous discovery of new applications have allowed the cloze test to remain a relevant and adaptable tool.

Keywords: Cloze procedure. Literature review. Authorship verification.

Resumo: O teste cloze foi originalmente proposto como um método para medir a legibilidade de textos. No entanto, rapidamente passou a ser utilizado em outras áreas de pesquisa. O objetivo desta 70 anos de literatura é mapear as aplicações que o teste cloze desenvolveu ao longo dos setenta anos desde sua criação. A revisão identifica seis grandes áreas de utilização: legibilidade textual; habilidade e compreensão de leitura dos estudantes; avaliação e aprendizagem de línguas; cognição e psiquiatria; detecção de plágio; e, mais recentemente, treinamento e avaliação de modelos de linguagem. Cada aplicação requer um método específico de remoção de itens, diferentes formatos de resposta e estratégias distintas de pontuação. A revisão aprofunda as discussões acadêmicas em torno do teste cloze, examinando de forma abrangente tanto

* External collaborator at the Department of Machine Learning and Data Processing, Masaryk University (Czech Republic). Master's degree in English-language Translation from Masaryk University. E-mail: <spetla@mail.muni.cz>.

** Assistant professor at the Department of Machine Learning and Data Processing, Masaryk University (Czech Republic). PhD in Economic Informatics from Mendel University in Brno. E-mail: <foltynnek@fi.muni.cz>.

as críticas quanto as vantagens que ele apresenta em seus diferentes usos. Esse diálogo contínuo e a descoberta constante de novas aplicações permitiram que o teste cloze permanecesse uma ferramenta relevante e adaptável.

Palavras-chave: Procedimento cloze. Revisão de literatura. Verificação de autoria.

Resumen: La prueba cloze fue propuesta originalmente como un método para medir la legibilidad de los textos. Sin embargo, pronto encontró otros usos en diversos campos de investigación. El propósito de esta revisión de la literatura es mapear las aplicaciones que la prueba cloze ha tenido a lo largo de los 70 años transcurridos desde su creación. La revisión identifica seis grandes áreas de uso: la legibilidad de los textos; la capacidad y la comprensión lectora de los estudiantes; la evaluación y el aprendizaje de lenguas; la cognición y la psiquiatría; la detección del plagio; y, más recientemente, el entrenamiento y la evaluación de modelos de lenguaje. Cada aplicación requiere un método específico de eliminación de ítems, diferentes formatos de respuesta y estrategias de puntuación distintas. La revisión profundiza en las discusiones académicas en torno a la prueba cloze, examinando exhaustivamente tanto las críticas como las ventajas que ofrece en sus diversos usos. Este diálogo continuo y el descubrimiento constante de nuevas aplicaciones han permitido que la prueba cloze siga siendo una herramienta relevante y adaptable.

Palabras clave: Procedimiento cloze. Revisión de la literatura. Verificación de autoría.

Introduction

The term *cloze* (pronounced /kloʊz/) was coined by Wilson L. Taylor (1953) who was referring to the Gestalt principle of closure. Just like one can mentally complete an unfinished shape, Taylor argued, one can also supply a word that has been deleted from a text. The cloze test consists of a text passage in which some words have been replaced with blanks that are supposed to be filled in. It was initially meant to provide a measure of text readability that could avoid the pitfalls of standard readability formulas. Soon, however, it found numerous additional uses. The purpose of the present paper is to review the most prominent uses which have emerged in the seventy years since the publication of Taylor's seminal paper. The inquiry initially arose from the necessity to confirm the cloze test's compatibility and reliability in authorship verification. Unfortunately, most existing reviews and bibliographies of the cloze test are from the 1970s and early 1980s (e.g., Riley, 1973; Jongsma, 1980; C. G. Robinson, 1981). More recently, there have appeared a meta-analysis by Watanabe and Koyama (2008) and a monograph by Sadeghi (2021), but these only deal with language testing. Despite recent increased interest in the cloze test due to its employment in deep machine learning, there seems to be no current general overview of the cloze test's uses. This is why the present authors decided to conduct a literature review.

Method

The literature review's purpose is to answer two research questions:

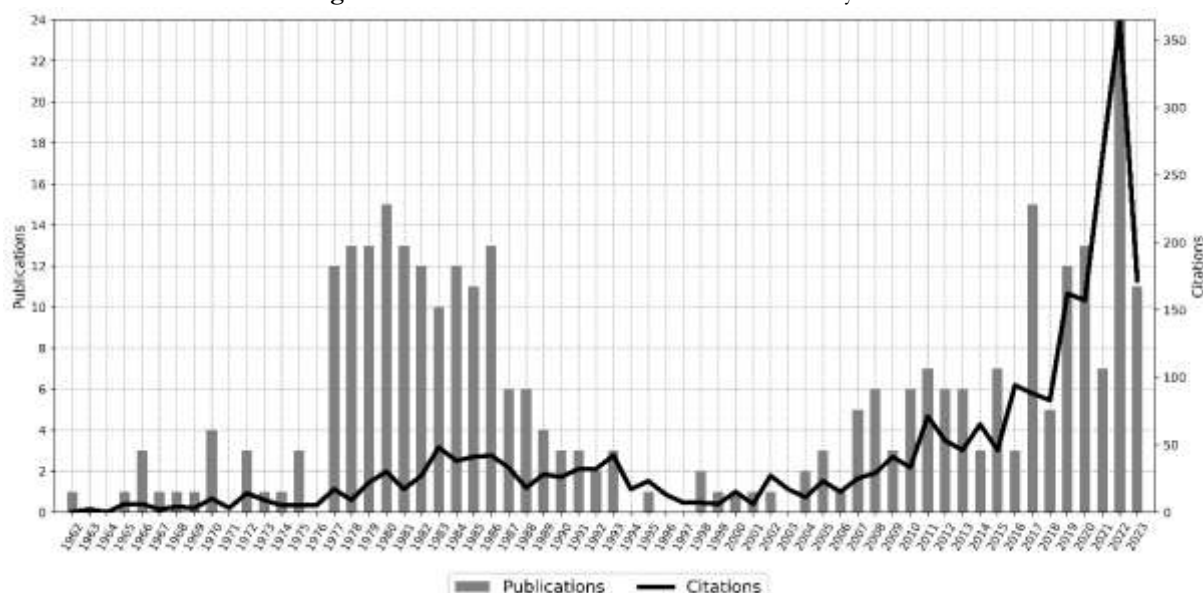
1. What possible applications does the cloze test have?
2. How can the cloze test contribute to plagiarism detection and verification of authorship?

The review combines a systematic method of literature search with qualitative synthesis. For the literature search, two citation databases were considered: Scopus and Web of Science. Because the results from Scopus included multiple entries for the same publications (Valderrama-Zurián et al., 2015), we settled on the Web of Science Core Collection. The records there go back to 1900, although five of the seven indexes within the collection are much more recent (Clarivate, 2024). Having weighed different alternatives for search terms, including, for instance, *gap-filling tests* and *fill-in tests*, we settled on *cloze*. This word is used only in its technical sense, ensuring focused

and relevant results. Since we wanted to eliminate studies that only used the cloze test as one of several methods instead of focusing specifically on the cloze test, we eventually opted for searching for *cloze* within titles of papers. This singled out studies *about* the cloze test and limited them to a manageable number.

The literature search was carried out on October 19, 2023 and resulted in 323 studies, 64% of which were published in the United States. As seen in Figure 1, their years of publication ranged from 1962 to 2023. Notably, there were two significant surges in publishing activity on cloze. The first one occurred between the late 1970s and late 1980s, and the other surge began around 2017 and has continued, with fluctuations, since. This latter surge can be partly attributed to the cloze task being one of the tasks used for training and evaluating large language models. Overall, almost half of the studies ($N = 150$) have been classified as dealing with Education & Educational Research according to the Research Areas classification. The other three most represented categories are Linguistics ($N = 98$), Psychology ($N = 85$), and Computer Science ($N = 55$). It was not unusual for a paper to be included in two or more categories at once.

Figure 1 - Numbers of Publications and Citations by Year



Source: Authors' visualization made from data retrieved from the Web of Science Core Collection (Clarivate), search conducted October 19, 2023.

Several selection criteria were applied. First, the total number of studies was curbed by setting the minimum of citations to five. This stemmed from the assumption that high-quality and/or impactful papers would be cited more than others. We also eliminated studies that did not deal with written texts. These included, for example, experiments in which researchers omitted frames from a video and had neural networks fill them in (Luo et al., 2020; Yu et al., 2020). In addition, we eliminated studies whose experimental subjects were below secondary-school ages. Finally, three studies were eliminated that turned out to be inaccessible. On the other hand, in order to cover all possible uses of the cloze test, we did not limit the studies by years of publication or document types. This led to a sample of 76 papers published between 1965 and 2021. Additional sources, especially those occurring in the reference lists of the selected papers, were also consulted.

Review

The review has identified six major areas of the cloze test's use, shown in **Tabl**. Subsumed under *other* are some less common uses, including comprehension and readability of computer programs (Cook et al., 1984; Hall & Zweben, 1986), sociolinguistic research (W. P. Robinson, 1965), and music teaching (Mateja, 1982).

Table 1 - Counts of Studies per Each Area of The Cloze Test's Use

Area of Use	Count	Percentage
readability	5	6.58%
reading ability and comprehension	22	28.95%
language assessment and language learning	17	22.37%
cognition and psychiatry	13	17.11%
plagiarism detection	3	3.95%
training and evaluation of language models	4	5.26%
other	12	15.79%
Total	76	100%

Source: Authors' calculations based on data retrieved from the Web of Science Core Collection (Clarivate), search conducted October 19, 2023.

Beginnings of the Cloze Procedure

The cloze test was introduced by Wilson L. Taylor (1953) as an alternative to standard readability formulas such as that by Flesch (1951) or Dale and Chall (1948). Taylor realized that readability may depend not just on intrinsic factors such as sentence lengths, numbers of short words, and proportions of certain parts of speech, but that it may also be impacted by readers' previous knowledge and by the degree to which the linguistic choices in the text align with readers' expectations (see also Guillemette, 1989, pp. 145–146). The cloze procedure commonly consists in deleting every *n*th word from a passage and in having people complete it again. The (pseudo)random method of deletion is supposed to ensure that the blanks “represent proportionately all kinds of words to the extent that they occur” in the text (Taylor, 1953, p. 419). Taylor described this “as throwing all potential readability influences in a pot, letting them interact, then sampling the result” (p. 417). Involving people in the test means that instead of merely *predicting* readability on the basis of textual characteristics, it is actually *measured* while “tak[ing] into account the interaction between and among the reader, the material, and the reading situation” (Hittleman, 1978, p. 119). Compared to standard readability formulas, the cloze procedure would thus seem to have greater validity.

In its canonical version, the cloze procedure starts with a passage of around 300 words. The first and the last sentence of the passage are often left intact to allow readers to orient themselves (e.g., Oller & Jonz, 1994, p. 4; Alderson, 2000, p. 207), and then starting from a random place in the second sentence, words are omitted at regular intervals—often every fifth, seventh, or 12th word is deleted. This is what is called the mechanical or fixed-ratio deletion pattern, and it

leads to something close to a random sample, although it is not truly random (e.g., Meredith & Vaughan, 1978; Alderson, 1979; Henk, 1981; O'Toole & King, 2010). Shorter deletion intervals tend to be avoided because they are thought to make it too difficult for the examinees to restore the items (MacGinitie, 1961; Alderson, 1979, p. 221). The deletions are marked by an underline of a fixed length so as not to give away the original number of characters (but see Rush & Klare, 1978; Entin & Klare, 1978 for alternatives). The subjects are to write each of their guesses in the corresponding blank. Only the original words are scored as correct, though minor deviations in spelling are usually overlooked. This is referred to as exact or verbatim scoring. Most studies talk about overall cloze scores, which are usually expressed as percentages of correct answers. The sum of scores attained by all participants is taken as indicative of the passage's readability.

While few studies in the sample specifically address measuring text readability through the cloze procedure, one notable line of research in this area focuses on its connection to the information-theoretical notion of entropy (Shannon & Weaver, 1949). To understand this connection, it is essential to recognize that English—or any language—is not sequences of words pulled out at random from a bag. Different words have different probabilities at different points. The amount of entropy (H), then, depends on how much choice one has at a given point, and it is understood as a measure of the amount of information. Within this framework, information is akin to uncertainty. Entropy can range from zero (completely certain, that is, only one item can go in the blank) to a maximum value determined by the logarithm of the number of equally likely choices, indicating a state of complete uncertainty. In a similar vein, Finn (1977) argued “that the Cloze Easiness of a word (per cent of subjects supplying the exact word in a cloze task) is a measure of *information* supplied by the word in the passage” (p. 512). The more information a deleted item contains, the more difficult it is to restore the item correctly.

Although the cloze procedure was met with approval and even enthusiasm, it has also been criticized on some points. Among the earliest critics was Wendell W. Weaver (1965) who did not agree with the assertion that the idea of closure from Gestalt theory could be applied to language. Although Ohnmacht, Weaver, and Kohler (1970, p. 214) later found evidence that perceptual closure correlates with cloze factors, it was not strong to the extent “that one would be led to accept the classical gestalt explanation of visual closure phenomena as descriptive of the process of completing cloze blanks”. Another, more serious, point of contention is the test's consistency and construct validity. Taylor (1956) suggested that the cloze procedure may not be useful just for measuring readability of texts, but that it may also be applied to assessing readers' comprehension, “existing knowledge, and success in learning and remembering” (p. 44). But with so many uses, it is not clear what exactly the cloze test measures and how various types of deletions and scoring methods can affect this (e.g., Porter, 1978; Alderson, 1979; Brown, 1980; Abraham & Chapelle, 1992; Kobayashi, 2002; O'Toole & King, 2010). Some of these issues will be explored in more detail when discussing other uses of the test.

Reading Ability and Comprehension

Taylor (1953, p. 415) presented the cloze procedure as a “psychological tool for measuring the effectiveness of communication”. This is because, essentially, readers completing a cloze test are guessing what the writer intended to convey. In Taylor's words, the “cloze procedure takes a measure of the likeness between the patterns a writer has used and the patterns the reader is anticipating while he is reading” (p. 417). This perspective suggests that the procedure can be used not only to evaluate texts, but also to assess readers' comprehension. Since readability can be understood as the interaction between the reader and the text (Hittleman, 1978), the transition of focus from the properties of the text to the ability of the reader is only natural. This shift reframed

the test's purpose: rather than distinguishing texts by difficulty, it became a tool for discriminating between readers—on the basis of their comprehension of a specific text and/or of their general reading ability (C. G. Robinson, 1981).

The cloze test has significant advantages over standard comprehension tests which consist of manually developed and tested questions. It is quicker and easier to produce, it does not suffer from researcher biases, its generation can easily be automated, and the scores are comparable across texts (O'Reilly & Streeter, 1977; C. G. Robinson, 1981; Hall & Zweben, 1986; Gellert & Elbro, 2013; Kleijn et al., 2019). In fact, it is partly thanks to the convenience of the procedure that it has gained more popularity in reading ability and comprehension assessment than in readability (Henk, 1981). The test has been used to assess the reading skills of native speakers, but even more frequently those of language learners, and it has made its name in the field of language testing.

Despite its obvious strengths, there has also been much criticism of the use of the cloze procedure for assessing reading comprehension and language proficiency, and many of the papers in the sample tried to address these shortcomings. A major line of criticism, that has been going on at least from the early 1970s, relates to whether the cloze test is sensitive to context above the sentence level. Carroll (1971, p. 35) claimed that cloze scores depended on redundancy in the immediate context, arguing that "[...] there is no clear evidence that cloze scores can measure the ability to comprehend or learn the major ideas or concepts that run through a discourse. It is even possible to secure cloze scores on the basis of meaningless material so long as grammatical cues are present".

Since reading comprehension is not just about recognizing language patterns but involves the ability to integrate broader context and grasp the overall meaning, this would mean that the cloze procedure was utterly inadequate to the task. In response to this concern, sensitivity to intersentential clues has become probably the most extensively researched aspect of the cloze test. At least 16 studies from our sample, both in the field of reading comprehension and language assessment, were devoted to testing this experimentally. Except for Alderson (1979), Shanahan, Kamil, and Tobin (1982), and Markham (1985), all of the studies found evidence in support of the claim that the cloze test is sensitive to textual relationships beyond the bounds of the sentence.

In the course of addressing the issue, some researchers started studying the nature of individual deletions, often classifying them on the basis of Halliday and Hasan's (1976) typology of cohesive devices and moving in the direction of replacing the mechanical cloze with targeted deletions of items that required wider context for correct completion (Bachman, 1982, 1985; Bridge & Winograd, 1982; Gellert & Elbro, 2013). The type of deletion that is based on specific criteria instead of a mechanical pattern is usually referred to as rational deletion. Since the focus with reading comprehension is not on the text itself but on readers' success at understanding the passage, it seems to be no longer necessary for the deletions to serve as a balanced sample of the text anyway (see Alderson, 1979, p. 226).

Other criticism of the cloze procedure used in reading comprehension comes from Ashby-Davis (1985) who dealt with cloze's validity in qualitative terms. She asserted that in order for a test to be valid, the test situation must copy the real situation, and the test must not test anything else in addition to what it is claimed to test. Because ordinary reading involves the full text without omissions, it proceeds relatively quickly, with long eye movement spans. Cloze testing, on the other hand, is much slower and involves specific eye movements. Furthermore, the general reading strategy is very different when the aim is to fill in words correctly than under normal circumstances, and one is actually producing language, like a writer rather than a reader. Since cloze testing is different from ordinary reading and involves both reading and writing, Ashby-Davis deemed it to

be an invalid test of reading comprehension. On the other hand, the fact that it involves language production in addition to reception is one of the reasons it is so frequently used in language testing.

Language Assessment and Language Learning

When the cloze procedure started to be used to test reading ability in L2 learners, it was a small step to using it for general language proficiency assessment. The idea already appeared in Taylor (1956), not long after his seminal paper, and Carroll, Carton, and Wilds (1959, p. 9) called it “a rather obvious development from the other uses to which cloze technique has been put”. They described the rationale behind this use as follows:

The hearer of an utterance or the reader of a text, to the extent that he is well acquainted with the language, will develop expectancies as to what is going to follow any given part of the text. The more experience he has with the language, the better he will be able to predict how to complete a given context, or to fill in a lacuna. (p. 3).

Indeed, cloze scores were found to correlate with language proficiency test scores (Oller, 1972; Brown, 1980), and to this day various versions of the test appear as integral parts of entrance exams (Trace, 2020, p. 249).

The purpose of proficiency testing is to discriminate between language learners on the basis of their proficiency, and also to distinguish them from native speakers (Alderson, 1980, p. 60). This requirement affects, among other things, the choice of scoring method. For native speakers, exact scoring of the cloze test was deemed almost as good as giving points to any contextually acceptable insertion, while being less laborious (Taylor, 1953; Henk, 1981). However, for non-native speakers, exact scoring could result in too difficult a test (Oller, 1972) and in frustration when valid responses are penalized for not matching the original word (Brown, 1980). This is why many authors suggested using what is often called semantic or acceptable-answer scoring, which allows any answer that fits the context to be accepted as correct. In fact, there have been studies claiming that semantic scoring resulted in better discrimination. Alderson (1980), for instance, concluded that to get the best separation between native and non-native speakers, the semantic scoring method should be applied in combination with difficult texts. On the other hand, O’Toole and King (2011) found that semantic scoring favored the most able students rather than the least able ones, contrary to the intended purpose of the scoring method. The authors’ explanation was that more proficient students have wider linguistic repertoires that they can draw on, making it easier for them to come up with an acceptable answer.

Despite its advantages, semantic scoring can bias judgments of acceptable responses. clozentropy, a portmanteau of cloze and entropy (Darnell, 1968, 1970; Reilly, 1971), mitigates this by pretesting cloze passages on a native-speaker group and using each response’s frequency to compute its information value. Scores are weighted logarithmically to reflect conformity to criterion-group norms (Reilly, 1971), thereby indexing proficiency as alignment with established language conventions (Darnell, 1970, p. 36). Although more informative than semantic scoring, clozentropy requires extensive pretesting and data processing, making it costlier and more labor-intensive (Brown, 1980, p. 315).

Besides exact, semantic, and clozentropy scoring, other scoring methods have also been explored. Alderson (1980), for example, compared exact and semantic scoring to three others: accepting also any grammatically correct replacements, accepting replacements that are of the same parts of speech as the original item, and accepting replacements belonging to any admissible parts of speech. All these scoring methods are expected to make it easier for non-native speakers to earn scores.

But even then the cloze test may be too difficult for them, especially for those at the lower end of the proficiency scale. A solution is to give the examinees items to choose from. Typically, this is done with the multiple-choice cloze, also referred to as maze (e.g., Williams et al., 2011). It is used both in reading comprehension and language proficiency testing. Brown (1980, p. 311) noted that unlike the open cloze, it probably tests just receptive skills, hence its decreased difficulty. On the other hand, it is more laborious to generate because one has to come up with reasonable distractors for each blank. Distractors are typically generated manually, but there have also been attempts to automate the process (O'Reilly & Streeter, 1977; Goto et al., 2010; Mostow et al., 2017; Ren & Zhu, 2021).

Because of the necessity to generate distractors, there have been other proposals that preserve the advantages of the multiple-choice cloze test but are much easier to generate. Ozete (1977), for instance, developed the reading-input test, in which the subjects choose from two alternatives—the correct word and a distractor selected from the surrounding text in the passage with the help of tables of random numbers. Some additional advantages of the reading-input test are that no extraneous vocabulary is introduced through the distractors and that the reading process is much more natural than either with open cloze or multiple-choice cloze. Another possibility is what Propst and Baldauf (1979) named the matching cloze. It was specifically designed so that even elementary-school pupils with English as their second language and with little production and reading skills could be tested with it. In it, the authors divided the text into sections and for each section, “the deleted words were presented in random order in the right margin” (Baldauf & Propst, 1978, p. 453). The task was to select the right word for each gap. In experiments with elementary school students in Mariana Islands, Baldauf and Propst (1978) and Propst and Baldauf (1979) found the matching cloze scores to correlate with scores from the Gates-MacGinitie Reading Tests and the Marianas English Achievement Test.

To general language proficiency assessment apply many of the concerns pertaining to reading comprehension. The cloze test's attractiveness resides in its ease of construction, but it “is perhaps deceptively simple” (Brown, 1980, p. 316). One of the most frequently asked questions posed by researchers was What does the test measure? They realized that some deletions may be much easier to complete than others and that different scoring methods, deletion patterns, and deletion starting points resulted in radically different tests. Studies such as Porter (1978), Alderson (1979), and Brown (1980) have shown that these parameters may significantly affect the resulting cloze scores. Alderson (1979, p. 226), therefore, suggested that since the cloze procedure does not lead to automatically valid tests, “each test produced by the technique needs to be validated in its own right and modified accordingly”.

These findings led to studies that aimed to uncover what variables affect scores the most. Conveniently, Abraham and Chapelle (1992) distinguished between intrinsic and extrinsic determinants of difficulty. While intrinsic factors relate to characteristics of the task itself, extrinsic factors have to do with the abilities and strategies of the persons performing the task. As regards intrinsic factors, researchers have agreed that frequency of the item in the language is associated with higher cloze scores (Finn, 1977; Abraham & Chapelle, 1992; Kobayashi, 2002; Trace et al., 2017). Among other factors associated with easiness were frequency of the item within the passage (Finn, 1977; Kobayashi, 2002; Trace et al., 2017), the items being function rather than content words, the items having simple inflectional paradigms (Abraham & Chapelle, 1992; Kobayashi, 2002), shortness of the items and passage (Trace et al., 2017), and readability of the passage (Finn, 1977).

Apart from language testing, there have also been a few attempts to use the cloze procedure for language learning, specifically for vocabulary acquisition. Lee (2008) successfully leveraged

rational cloze passages in English as a second language (ESL) vocabulary teaching. The author also used the C-test “to determine participants’ ability to extend use of target vocabulary to other contexts” (p. 644). The C-test was developed by Klein-Braley and Raatz (1984) to address the numerous problems of the cloze procedure. Lee (2008) argued that the C-test can be used to find out if students are able to apply the vocabulary they have learned in different settings. Another study using the cloze test for language learning was by Zou (2017), who compared the extents to which cloze exercises, sentence writing, and composition writing promote learning new vocabulary. It was hypothesized that cloze exercises would be the least effective because they require less active involvement from students compared to sentence and composition writing. In an experiment, three groups of students—each assigned one of the tasks—obtained significantly different mean scores on a subsequent vocabulary knowledge test, supporting the hypothesis. Drawing on think-aloud protocols and retrospective interviews, Zou further argued that, unlike the other tasks, cloze exercises do not encourage chunking, hierarchical organization of information, and pre-task planning.

Cognition and Psychiatry

Much research involving the cloze procedure has been conducted in the fields of cognitive science, neuroscience, psychology, and psychiatry. Already in 1897, Hermann Ebbinghaus used a similar procedure, called *Combinationsgabe test* or completion test, as a measure of intellectual ability. Like the cloze test it consists of a passage in which words (or syllables) have been deleted and the task is to fill them in. In the completion test, however, the lengths of the missing elements are indicated, and the passage may be read to the subjects in whole before they are given the test. The latter feature widens the range of cognitive processes involved to listening and recall. As for the cloze test, Taylor (1956) also suggested using it to test intelligence because intelligence is what comprehension depends on. Indeed, Taylor found Air Force trainees’ cloze scores to correlate strongly both with comprehension tests and an intelligence test. However, measuring intelligence is not the only application in this area that the cloze procedure has been put to. The present section discusses three others: cognitive styles, predictive processing, and psychiatric disorders.

Researchers have questioned whether examinees’ cognitive styles—specifically field independence, or the ability to detach from surrounding context—bias cloze-test performance. Readence et al. (1980) used the 25-item Group Embedded Figures Test (GEFT; Oltman et al., 1971), in which participants locate simple shapes within complex designs (Witkin, 1971), and found significant GEFT–cloze correlations—but observed similar correlations with a standard comprehension measure, challenging the idea that field independence uniquely benefits cloze tasks. Stansfield and Hansen (1983) reported a GEFT–cloze correlation of .43—stronger than GEFT’s links to other Spanish-proficiency tests—and argued this reflects a cognitive-style bias. Abraham and Chapelle (1992), who identified field independence as an extrinsic factor in cloze difficulty, found GEFT correlations to be lower than expected. Collectively, these studies offer little support for the field-independence hypothesis.

Regarding predictive processing, there has been a line of research that started with Kutas and Hillyard’s (1980) study of event-related brain potentials (ERPs). Studies from our sample dealing with this research include Arcuri et al. (2001), Block and Baldwin (2010), Delong et al. (2011), and Staub et al. (2015). Such studies dealt with what is called sentence completion norms. These norms are derived by presenting participants with sentences missing the final word and recording their responses. From these, cloze probabilities are calculated, indicating “the proportion of subjects using [a given] word to complete a particular sentence” (Kutas & Hillyard, 1984, p. 161). In ERP experiments, subjects are shown the normed sentences that end either with a

predictable word (high cloze probability) or an unpredictable one (low cloze probability). At the same time, researchers measure the brain's electrical activity via electroencephalogram (EEG) to observe how the brain reacts to expected versus unexpected sentence completions. Kutas and Hillyard (1980) found that the N400 component of the ERP is highly sensitive to the cloze probability of a word at the end of a sentence. This "suggested that the N400 indexed something fundamental about the processing of meaning" (Kutas & Federmeier, 2011, p. 624).

Other studies used the cloze test to gauge how predictable the speech of persons affected by psychiatric disorders is. To this end, they usually generated a cloze test from recorded speech of such persons and had a panel of neurotypical raters complete it. The resulting cloze scores were taken to indicate the level of predictability. The focus was on persons living with schizophrenia, usually in comparison to others, including people with an alcohol use disorder (Cheek & Amarel, 1968), people not affected by any disorder (Manschreck et al., 1980), and people with a bipolar disorder (Newby, 1998). Cloze passages of schizophrenic patients tended to have lower scores than those of the other groups.

Plagiarism Detection

Cloze tests are usually generated from passages that the examinees have not seen before because familiarity with the passage might skew the scores upwards. However, Glatt and Haertel (1982) exploited the opposite assumption—"that students should be able to reproduce their own writing far more accurately than that of anyone else" (p. 127)—as a way to distinguish between authors and non-authors. Glatt and Haertel cited Rankin's (1959, p. 132) description of the test as "an objective measure of language correspondence between a reader and writer" arguing that if both are, in fact, the same person, they should obtain a higher score on the cloze test.

Glatt and Haertel's data showed that both plagiarists and non-plagiarists could obtain similar scores, but only plagiarists got very low scores. Having determined a posterior probability of correctly guessed words, Glatt and Haertel applied Bayes' theorem, which resulted in a formula that estimates the likelihood of plagiarism based on the number of errors made. There are two limitations to this method—first, it does not provide any evidence of plagiarism (just probability), and second, it requires an accurate estimate of the overall rate of plagiarism in the population, which can be challenging to determine. Thus, while this method can cast suspicion on some writers, it does not guarantee detection of all instances of plagiarism. On the other hand, it does not depend on finding the original source of the text, making it possible to use it for detecting contract cheating.

Standing and Gorassini (1986) were not happy with this and suggested a different way of inferring plagiarism from cloze scores. Their method involved testing individuals on short passages extracted from their own work, as well as on passages from comparable but unfamiliar texts. They hypothesized that plagiarists would show consistently low scores across both types of passages, reflecting a lack of familiarity with the content. In other words, the scores from the non-authored passages would provide a convenient baseline for comparison. Both of Standing and Gorassini's experiments confirmed that students scored higher on cloze passages constructed from their own texts than someone else's texts. In the end, the authors stressed, the results are statistical in nature and cannot be used to prove one's guilt. Instead, such results can at best serve to exonerate the innocent.

Given that the cloze procedure was originally designed to measure readability, Torres and Roig (2005) investigated how the readability of source texts affected cloze scores when students paraphrased them. They found that scores were higher for paraphrased texts than for plagiarized ones, using both exact and semantic scoring. Scores were also higher when the source text was easy

to read. While the interaction between readability and production mode was not significant, the lowest mean score occurred with plagiarized texts from difficult sources. Like the previous two studies, Torres and Roig used mechanical cloze, but for future work they suggested that deleting rationally key terms could result in better discrimination. In fact, recent graduate work has shown that deleting items rationally using predictions from a large language model leads to better results in authorship verification and offers new possibilities (Dobeš, 2022; Kancko, 2023).

Training and Evaluation of Language Models

The cloze test has also inspired natural language processing engineers, be it in pretraining and finetuning language models (Hoang et al., 2018; Baevski et al., 2019), or in creating datasets for model training and evaluation (Xie et al., 2018; Lewis et al., 2019; Schick & Schütze, 2021).

Baevski et al. (2019) highlighted the use of a “cloze-style objective” for pretraining large language models. This method involves masking tokens (usually subwords, i.e., elements smaller than words) within the input text, rendering them invisible to the model during training. The model is then tasked with predicting these masked tokens on the basis of the left and right context. When the model’s predictions differ from the original tokens, the error gradient is computed and backpropagated to adjust the model’s parameters. This iterative process of prediction, error evaluation, and parameter adjustment enables the model to learn and improve its ability to infer and generate contextually appropriate responses.

Xie et al. (2018) developed a large dataset of multiple-choice cloze tests that can be used to evaluate natural language understanding models. The deletions were selected in such a way that answering them correctly requires grasping distant contextual clues. On the assumption that English examinations contain a large proportion of questions that require reasoning ability, they collected such tests from the web. They cleaned the data up and divided it into middle-school and high-school questions. They also had the questions manually classified as questions that require knowledge of grammar, intrasentential reasoning, extraction of the answer from surrounding text, and intersentential reasoning. Overall, it is applications like these that are currently giving the cloze test new opportunities to shine.

Conclusion

The past seventy years have seen major developments in the use of the cloze test. Although initially introduced as a measure of text readability, it has subsequently been applied to reading assessment, language testing, cognitive and psychiatric research, plagiarism detection, and, most recently, deep learning. The procedure itself has also diversified considerably, with variation in deletion principles, response formats, and scoring methods. The literature reviewed here therefore suggests that cloze is better understood not as a single test but as a family of procedures whose properties depend substantially on how the task is constructed, administered, and scored. This diversity has contributed to the procedure’s adaptability, but it also makes clear that future research must attend more closely to the relationships among task design, the construct being measured, and the population or context in which the procedure is used.

These conclusions should nevertheless be interpreted in light of several limitations of the present review. First, although the search was supposed to include translated titles from the Web of Science Core Collection (Matthews, 2021), no non-English studies were found, possibly introducing language bias (see Vera-Baceta et al., 2019). Second, the sample is not comprehensive; given the volume of cloze research, we limited inclusion to studies with at least five citations to

focus on the most influential work. Third, while the studies span multiple research areas, our expertise covers only some. These limitations constrain the scope of the conclusions that can be drawn, but the literature nevertheless points to a coherent research agenda centred on three broad questions: what different cloze procedures measure, how their properties vary across populations and contexts, and what they can reveal when applied to emerging problems.

The most fundamental priority concerns the construct validity of different cloze procedures. A question that has accompanied the method throughout its history is what successful restoration actually measures. Reading comprehension, for instance, draws on linguistic knowledge, cognitive resources, and prior knowledge, and different deletions may place very different demands on the reader. Some gaps can be restored largely from local lexical or syntactic information, whereas others require integration across sentences, inference, or wider knowledge. The validity of cloze as a measure of reading comprehension may therefore be less a matter of determining whether cloze measures comprehension *per se* than of establishing which aspects of comprehension a particular implementation targets, under what conditions, and how adequately. Future research should examine systematically how deletion principles, response formats, scoring methods, and characteristics of the text influence the construct being measured.

A related priority is differential validation across populations and contexts. Because cloze procedures have been used with different age groups, proficiency levels, languages, educational settings, and research populations, evidence obtained with one implementation cannot necessarily be assumed to generalize to another. Comparative studies are needed to determine how task characteristics interact with linguistic proficiency, prior knowledge, educational background, and other population variables. Such work would help establish more clearly which forms of the procedure are appropriate for particular purposes and would make comparisons across studies more meaningful.

These questions of task design and validity are equally important in more specialized applications. One example is plagiarism detection and authorship verification, where the available evidence remains comparatively limited. Although plagiarism detection was addressed by only three studies, they identified both important strengths and limitations of the cloze approach. Unlike many other methods, it does not require prior identification of the source text, which may be advantageous when the original material is unavailable. On the other hand, its results are inherently probabilistic rather than definitive. Notably, all three studies used an every-fifth-word deletion pattern, while the available evidence suggests that more rational item selection may yield substantially better results. Future research should therefore examine how different deletion principles affect the effectiveness of cloze-based plagiarism detection and authorship verification, as well as the factors underlying the apparent advantage of authors over non-authors.

The growing use of cloze-like procedures in language-model research extends this same set of questions into a new domain and creates an opportunity to connect educational, psycholinguistic, and computational traditions that have largely developed separately. Cloze-like prediction has become an important objective in machine language learning, and the ability of models trained to recover or predict linguistic material to acquire syntactic, lexical, and semantic regularities raises the question of how much linguistic knowledge can be acquired from distributional information alone. At the same time, limitations in grounding, inference, and pragmatic interpretation raise the question of where contextual prediction ceases to provide an adequate account of language knowledge. Comparisons among language models, ordinary readers, and authors on carefully controlled cloze tasks could therefore do more than improve test construction: they could help distinguish what is recoverable from linguistic context in general

from what depends on broader comprehension, prior knowledge, or an individual's particular relationship to a text.

Taken together, these directions suggest that the future value of the cloze procedure lies not simply in extending it to new applications, but in specifying more precisely what different implementations measure, for whom, and under what conditions. Its continued relevance will depend on linking methodological choices to clearly defined constructs and purposes while also using new computational tools to revisit questions that have accompanied the procedure since its introduction.

References

- Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, 76(4), 468–479. <https://doi.org/10.1111/j.1540-4781.1992.tb05394.x>
- Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13(2), 219–227. <https://doi.org/10.2307/3586211>
- Alderson, J. C. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning*, 30(1), 59–76. <https://doi.org/10.1111/j.1467-1770.1980.tb00151.x>
- Alderson, J. C. (2000). *Assessing Reading*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732935>
- Arcuri, S., Rabe-Hesketh, S., Morris, R., & McGuire, P. (2001). Regional variation of cloze probabilities for sentence contexts. *Behavior Research Methods Instruments & Computers*, 33(1), 80–90. <https://doi.org/10.3758/BF03195350>
- Ashby-Davis, C. (1985). Cloze and comprehension: A qualitative analysis and critique. *Journal of Reading*, 28(7), 585–589.
- Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, 16(1), 61–70. <https://doi.org/10.2307/3586563>
- Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, 19(3), 535–556. <https://doi.org/10.2307/3586277>
- Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., & Auli, M. (2019). Cloze-driven pretraining of self-attention networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5359–5368. <https://doi.org/10.18653/v1/D19-1539>
- Baldauf, R. B., & Propst, I. K. (1978). Preliminary evidence regarding the validity of a modified cloze procedure for lower elementary ESL students. *Educational and Psychological Measurement*, 38(2), 451–455. <https://doi.org/10.1177/001316447803800229>
- Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, 42(3), 665–670. <https://doi.org/10.3758/BRM.42.3.665>

- Bridge, C. A., & Winograd, P. N. (1982). Readers' awareness of cohesive relationships during cloze comprehension. *Journal of Reading Behavior*, 14(3), 299–312. <https://doi.org/10.1080/10862968209547457>
- Brown, J. D. (1980). Relative merits of 4 methods for scoring cloze tests. *Modern Language Journal*, 64(3), 311–317. <https://doi.org/10.2307/324497>
- Carroll, J. B. (1971). *Defining language comprehension: Some speculations* (ED057698). ERIC.
- Carroll, J. B., Carton, A. S., & Wilds, C. P. (1959). *An investigation of "cloze" items in the measurement of achievement in foreign languages* (ED021513). ERIC.
- Cheek, F. E., & Amarel, M. (1968). Studies in the sources of variation in cloze scores: II. The verbal passages. *Journal of Abnormal Psychology*, 73(5), 424–430. <https://doi.org/10.1037/h0026187>
- Clarivate. (2024, April 8). *Web of Science Core Collection Training & Support*. Web of Science Group. <https://clarivate.com/webofsciencegroup/support/wos/wos-core-collection/>.
- Cook, C., Bregar, W., & Foote, D. (1984). A preliminary investigation of the use of the cloze procedure as a measure of program understanding. *Information Processing and Management*, 20(1–2), 199–208. [https://doi.org/10.1016/0306-4573\(84\)90050-5](https://doi.org/10.1016/0306-4573(84)90050-5)
- Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27(1), 11–28.
- Darnell, D. K. (1968). *The development of an English language proficiency test of foreign students, using a clozentropy procedure: Final report* (ED024039). ERIC.
- Darnell, D. K. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. *Speech Monographs*, 37(1), 36–46. <https://doi.org/10.1080/03637757009375647>
- Delong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, 48(9), 1203–1207. <https://doi.org/10.1111/j.1469-8986.2011.01199.x>
- Dobeš, E. (2022). *Ghostwriting detector* [Master's thesis, Mendel University in Brno, Faculty of Business and Economics]. <https://theses.cz/id/108kwc/>.
- Ebbinghaus, H. (1897). Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern [On a new method for testing mental abilities and its use with school children]. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, 13, 401–459.
- Entin, E. B., & Klare, G. R. (1978). Some inter-relationships of readability, cloze and multiple choice scores on a reading comprehension test. *Journal of Reading Behavior*, 10(4), 417–436. <https://doi.org/10.1080/10862967809547294>
- Finn, P. J. (1977). Word-frequency, information-theory, and cloze performance: A transfer feature theory of processing in reading. *Reading Research Quarterly*, 13(4), 508–537. <https://doi.org/10.2307/747510>
- Flesch, R. (1951). *How to test readability*. Harper.

- Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, 31(1), 16–28. <https://doi.org/10.1177/0734282912451971>
- Glatt, B. S., & Haertel, E. H. (1982). The use of the cloze testing procedure for detecting plagiarism. *The Journal of Experimental Education*, 50(3), 127–136. <https://doi.org/10.1080/00220973.1982.11011813>
- Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning-an International Journal*, 2(3), 210–224. <https://doi.org/10.34105/j.kmel.2010.02.016>
- Guillemette, R. A. (1989). The Cloze procedure: An assessment of the understandability of data processing texts. *Information and Management*, 17(3), 143–155. [https://doi.org/10.1016/0378-7206\(89\)90016-5](https://doi.org/10.1016/0378-7206(89)90016-5)
- Hall, W. E., & Zweben, S. H. (1986). The cloze procedure and software comprehensibility measurement. *IEEE Transactions on Software Engineering*, SE-12(5), 608–623. <https://doi.org/10.1109/TSE.1986.6312957>
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
- Henk, W. A. (1981). Effects of modified deletion strategies and scoring procedures on cloze test-performance. *Journal of Reading Behavior*, 13(4), 347–357. <https://doi.org/10.1080/10862968109547423>
- Hittleman, D. R. (1978). Readability, readability formulas, and cloze: Selecting instructional materials. *Journal of Reading*, 22(2), 117–122.
- Hoang, L., Wiseman, S., & Rush, A. (2018). Entity tracking improves cloze-style reading comprehension. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1049–1055. <https://doi.org/10.18653/v1/D18-1130>
- Jongsma, E. R. (1980). The cloze procedure: A survey of the research. *Occasional Papers in Reading*, 3–46.
- Kancko, T. (2023). *Authorship verification via cloze-test* [Master's thesis, Masaryk University, Faculty of Informatics]. <https://is.muni.cz/th/bz1lo/>.
- Kleijn, S., Maat, H. P., & Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing*, 36(4), 553–572. <https://doi.org/10.1177/0265532219840382>
- Klein-Braley, C., & Ratz, U. (1984). A survey of research on the C-Test. *Language Testing*, 1(2), 134–146. <https://doi.org/10.1177/026553228400100202>
- Kobayashi, M. (2002). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal*, 86(4), 571–586. <https://doi.org/10.1111/1540-4781.00162>
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62(1), 621–647. <https://doi.org/10.1146/annurev.psych.093008.131123>

- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205. <https://doi.org/10.1126/science.7350657>
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161–163. <https://doi.org/10.1038/307161a0>
- Lee, S. H. (2008). Beyond reading and proficiency assessment: The rational cloze procedure as stimulus for integrated reading, writing, and vocabulary instruction and teacher-student interaction in ESL. *System*, 36(4), 642–660. <https://doi.org/10.1016/j.system.2008.04.002>
- Lewis, P., Denoyer, L., & Riedel, S. (2019). Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4896–4910). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1484>
- Luo, D., Liu, C., Zhou, Y., Yang, D., Ma, C., Ye, Q., & Wang, W. (2020). Video cloze procedure for self-supervised spatio-temporal learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 11701–11708. <https://doi.org/10.1609/aaai.v34i07.6840>
- MacGinitie, W. H. (1961). Contextual constraint in English prose paragraphs. *The Journal of Psychology*, 51(1), 121–130.
- Manschreck, T. C., Maher, B. A., & Rucklos, M. E. (1980). Cloze procedure and written language in Schizophrenia. *Language and Speech*, 23(4), 323–328. <https://doi.org/10.1177/002383098002300402>
- Markham, P. L. (1985). The rational deletion cloze and global comprehension in German. *Language Learning*, 35(3), 423–430. <https://doi.org/10.1111/j.1467-1770.1985.tb01085.x>
- Mateja, J. (1982). Musical cloze: Background, purpose, and sample. *Reading Teacher*, 35(4), 444–448.
- Matthews, T. (2021). *Web of Science Core Collection Reference Guide*. Clarivate. https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/2021/04/2021_04-New-WoS-CC-QRG.pdf
- Meredith, K., & Vaughan, J. (1978). Stability of cloze scores across varying deletion patterns. In P. D. Pearson & J. Hansen (Eds.), *Reading: Disciplined inquiry in process and practice* (ED227441; pp. 181–184). ERIC.
- Mostow, J., Huang, Y.-T., Jang, H., Weinstein, A., Valeri, J., & Gates, D. (2017). Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children’s comprehension while reading. *Natural Language Engineering*, 23(2), 245–294. <https://doi.org/10.1017/S1351324916000024>
- Newby, D. (1998). “Cloze” procedure refined and modified. “Modified Cloze”, “reverse Cloze” and the use of predictability as a measure of communication problems in psychosis. *British Journal of Psychiatry*, 172(FEB.), 136–141. <https://doi.org/10.1192/bjp.172.2.136>
- Ohnmacht, F. W., Weaver, W. W., & Kohler, E. T. (1970). Cloze and closure: A factorial study. *Journal of Psychology: Interdisciplinary and Applied*, 74(2), 205–217. <https://doi.org/10.1080/00223980.1970.9923731>
- Oller, J. W. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *The Modern Language Journal*, 56(3), 151. <https://doi.org/10.2307/324037>

- Oller, J. W., & Jonz, J. (Eds.). (1994). *Cloze and coherence*. Bucknell University Press. <https://doi.org/10.13140/RG.2.1.2666.2889>
- Oltman, P. K., Raskin, E., & Witkin, H. A. (1971). *Group embedded figures test*. Consulting Psychologists Press.
- O'Reilly, R. P., & Streeter, R. E. (1977). Report on the development and validation of a system for measuring literal comprehension in a multiple-choice cloze format: Preliminary factor analytic results. *Journal of Reading Behavior*, 9(1), 45–69. <https://doi.org/10.1080/10862967709547206>
- O'Toole, J. M., & King, R. A. R. (2010). A matter of significance: Can sampling error invalidate cloze estimates of text readability? *Language Assessment Quarterly*, 7(4), 303–316. <https://doi.org/10.1080/15434303.2010.515699>
- O'Toole, J. M., & King, R. A. R. (2011). The deceptive mean: Conceptual scoring of cloze entries differentially advantages more able readers. *Language Testing*, 28(1), 127–144. <https://doi.org/10.1177/0265532210375687>
- Ozete, O. (1977). The cloze procedure: A modification. *Foreign Language Annals*, 10(5), 565–568. <https://doi.org/10.1111/j.1944-9720.1977.tb03033.x>
- Porter, D. (1978). Cloze procedure and equivalence. *Language Learning*, 28(2), 333–341. <https://doi.org/10.1111/j.1467-1770.1978.tb00138.x>
- Propst, I. K., & Baldauf, R. B. (1979). Use matching cloze tests for elementary ESL students. *Reading Teacher*, 32(6), 683–690.
- Rankin, E. F. (1959). The cloze procedure: Its validity and utility. In O. S. Causey & W. Eller (Eds.), *Eighth yearbook of the National Reading Conference* (pp. 131–144). National Reading Conference.
- Readence, J. E., Baldwin, R. S., Bean, T. W., & Dishner, E. K. (1980). Field dependence-independence as a variable in cloze test performance. *Journal of Literacy Research*, 12(1), 65–67. <https://doi.org/10.1080/10862968009547353>
- Reilly, R. R. (1971). A note on “Clozentropy: A procedure for testing English language proficiency of foreign students.” *Speech Monographs*, 38(4), 350–353. <https://doi.org/10.1080/03637757109375730>
- Ren, S., & Zhu, K. Q. (2021). Knowledge-driven distractor generation for cloze-style multiple choice questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5), 4339–4347. <https://doi.org/10.1609/aaai.v35i5.16559>
- Riley, P. M. (1973). *The cloze procedure: A selected annotated bibliography* (ED106749). ERIC.
- Robinson, C. G. (1981). Cloze procedure: A review. *Educational Research*, 23(2), 128–133. <https://doi.org/10.1080/0013188810230206>
- Robinson, W. P. (1965). Cloze procedure as a technique for the investigation of social class differences in language usage. *Language and Speech*, 8(1), 42–55. <https://doi.org/10.1177/002383096500800104>
- Rush, R. J., & Klare, G. R. (1978). Re-opening the cloze blank issue. *Journal of Reading Behavior*, 10(2), 208–210. <https://doi.org/10.1080/10862967809547271>

- Sadeghi, K. (2021). *Assessing second language reading: Insights from cloze tests*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-84470-7>
- Schick, T., & Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269. <https://doi.org/10.18653/v1/2021.eacl-main.20>
- Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, 17(2), 229–255. <https://doi.org/10.2307/747485>
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.
- Standing, L., & Gorassini, D. (1986). An evaluation of the cloze procedure as a test for plagiarism. *Teaching of Psychology*, 13(3), 130–132. https://doi.org/10.1207/s15328023top1303_7
- Stansfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly*, 17(1), 29–38. <https://doi.org/10.2307/3586422>
- Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82, 1–17. <https://doi.org/10.1016/j.jml.2015.02.004>
- Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Quarterly*, 30(4), 415–433. <https://doi.org/10.1177/107769905303000401>
- Taylor, W. L. (1956). Recent developments in the use of “cloze procedure.” *Journalism Quarterly*, 33(1), 42–48, 99. <https://doi.org/10.1177/107769905603300106>
- Torres, M., & Roig, M. (2005). The cloze procedure as a test of plagiarism: The influence of text readability. *The Journal of Psychology*, 139(3), 221–232. <https://doi.org/10.3200/JRLP.139.3.221-232>
- Trace, J. (2020). Clozing the gap: How far do cloze items measure? *Language Testing*, 37(2), 235–253. <https://doi.org/10.1177/0265532219888617>
- Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing*, 34(2), 151–174. <https://doi.org/10.1177/0265532215623581>
- Valderrama-Zurián, J.-C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics*, 9(3), 570–576. <https://doi.org/10.1016/j.joi.2015.05.002>
- Vera-Baceta, M.-A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics*, 121(3), 1803–1813. <https://doi.org/10.1007/s11192-019-03264-z>
- Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies*, 26(2), 103–133.
- Weaver, W. W. (1965). Theoretical aspects of the cloze procedure. In E. L. Thurston & L. E. Hafner (Eds.), *The philosophical and sociological basis of reading*. National Reading Conference.

Williams, R. S., Ari, O., & Santamaria, C. N. (2011). Measuring college students' reading comprehension ability using cloze tests. *Journal of Research in Reading*, 34(2), 215–231. <https://doi.org/10.1111/j.1467-9817.2009.01422.x>

Witkin, H. A. (1971). *Group Embedded Figures Test* [Dataset]. <https://doi.org/10.1037/t06471-000>

Xie, Q., Lai, G., Dai, Z., & Hovy, E. (2018). Large-scale cloze test dataset created by teachers. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2344–2356. <https://doi.org/10.18653/v1/D18-1257>

Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., & Kloft, M. (2020). Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events. In *MM 2020—Proceedings of the 28th ACM International Conference on Multimedia* (pp. 583–591). Association for Computing Machinery, Inc. <https://doi.org/10.1145/3394171.3413973>

Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research*, 21(1), 54–75. <https://doi.org/10.1177/1362168816652418>

Received: 08/06/2026

Final version received: 30/08/2026

Accepted: 01/09/2026

Published online: 08/09/2026