Article

LANGUAGE TESTING

Language Testing
2019, Vol. 36(4) 553–572

DOI: 10.1177/0265532219840382
journals.sagepub.com/home/ltj


Cloze testing for comprehension assessment: The HyTeC-cloze

Suzanne Kleijn
Utrecht University, Netherlands

Henk Pander Maat
Utrecht University, Netherlands

Ted Sanders
Utrecht University, Netherlands

Abstract

Although there are many methods available for assessing text comprehension, the cloze test is not widely acknowledged as one of them. Critiques on cloze testing center on its supposedly limited ability to measure comprehension beyond the sentence. However, these critiques do not hold for all types of cloze tests; the particular configuration of a cloze determines its validity. We review various cloze configurations and discuss their strengths and weaknesses. We propose a new cloze procedure specifically designed to gauge text comprehension: the Hybrid Text Comprehension cloze (HyTeC-cloze). It employs a hybrid mechanical-rational deletion strategy and semantic scoring of answers. The procedure was tested in a large-scale study, involving 2926 Dutch secondary school students with 120 unique cloze tests. Our results show that, in terms of reliability and validity, the HyTeC-cloze matches and sometimes outperforms standardized tests of reading ability.

Keywords

Cloze tests, reading assessment, scoring procedures, text comprehension, test validity

Corresponding author:
Suzanne Kleijn, Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, Utrecht, 3512 JK, Netherlands.
Email: s.kleijn1@uu.nl

As a measure of text comprehension, the cloze test is often critiqued (e.g., Klein-Braley & Raatz, 1984; Pearson & Hamm, 2005; Shanahan, Kamil, & Webb Tobin, 1982). However, many of these critiques are only valid for standard cloze tests in which every Xth word has been deleted. In this paper, we present a new cloze procedure that is especially suitable for comparing comprehension levels between large numbers of texts and/or readers. We will start with a short review of the cloze discussion. Then, we introduce our new procedure: "the Hybrid Text Comprehension cloze test" (HyTeC-cloze). We will end with an evaluation of the procedure using data from a large-scale study with 120 cloze tests and 2926 Dutch secondary school students.

The much-disputed cloze test

Cloze tests come in various forms. They have the following in common: "bits of some discourse are omitted and the task set the examinee is to restore the missing pieces" (Oller & Jonz, 1994b, p. 19). Subsequently, the answers of the examinee are scored. The score can be seen as a measure of the readability of the text, but also as a measure of the reading ability of the examinee (e.g., O'Toole & King, 2011; Taylor, 1953). As such, cloze tests have been popular in readability studies as well as in language assessment studies. As a measure of text comprehension, the validity of cloze tests is still under discussion (e.g., Brown, 2013; Chen, 2004; Gellert & Elbro, 2013; Greene, 2001; Kobayashi, 2002a, 2002b, 2004; Oller & Jonz, 1994a; O'Toole & King, 2010, 2011; Trace, Brown, Janssen, & Kozhevnikova, 2017). Critics believe that cloze is not sensitive to intersentential constraints and that it predominantly measures lower-order skills (i.e., grammatical and linguistic knowledge; Alderson, 1979a; Kintsch & Yarbrough, 1982; Klein-Braley & Raatz, 1984; Pearson & Hamm, 2005; Shanahan et al., 1982). A consistent view within the literature is that cloze is not a valid measure of text comprehension because it does not measure discourse level representations.

However, advocates of the cloze challenge this view and claim that a large percentage of cloze gaps does require information processing across sentences (Brown, 1983; Chihara, Oller, Weaver, & Chávez-Oller, 1994; Cziko, 1983; Gellert & Elbro, 2013; Henk, 1982; Jonz, 1994; McKenna & Layton, 1990; among others). Brown ([1983]1994) showed that 56–70% of cloze items in standard cloze tests are cohesive items (following the classification of Halliday & Hasan, 1976). In addition, Jonz (1994) found that on average 32% of the gaps in standard cloze tests require information beyond sentence borders. Finally, analyses of item difficulty show that passage-level variables influence the difficulty of individual cloze gaps (Chávez-Oller, Chihara, Weaver, & Oller, 1994; Kobayashi, 2002a; Oller & Chen, 2007; Trace et al., 2017). All these findings support the stance that standard cloze tests measure beyond sentence boundaries.

Despite the controversy still surrounding them, cloze tests have important advantages over other assessment methods. First, they are relatively easy to construct and can be used on a wide range of texts. For large-scale studies, this is a major advantage. Furthermore, cloze tests are very suitable for systematic investigations of text difficulty. All texts are treated in the same way and the difficulty of the items directly transfers from the text (Klare, 1976; Miller & Coleman, 1967). Thus, texts can be compared by their cloze scores; this is hard to do with standard comprehension questions, as these questions

may differ in difficulty. Another advantage is that although it can be hard to ask even 10 sensible questions about a 300-word text, cloze testing provides many items that are distributed over the entire text. Finally, when used in experimental studies, cloze comprehension tests can be more resilient to experimenter biases. Although comprehension questions may be designed to be specifically sensitive to certain text manipulations, cloze tests are not.

Of course, all these advantages will only apply in a well-configured cloze test. In our new cloze procedure, we have maximized the strengths of cloze testing and have addressed problems previously reported in literature. We will present the Hybrid Text Comprehension cloze and its rationale in the next section.

The Hybrid Text Comprehension cloze

The Hybrid Text Comprehension cloze (HyTeC-cloze) was developed as an alternative to the standard cloze and other standard comprehension assessment measures. We configured the HyTeC-cloze to be as follows:

- a valid and reliable measure of text comprehension;
- without confounds (confounding variables) between question difficulty and text difficulty;
- minimally sensitive to intrasentential (local) constraints;
- applicable to a wide range of texts;
- easy and fast to create;
- sensitive to differences between texts and between text versions
- sensitive to differences between test takers with different reading abilities;
- suitable for high- and low-proficiency test takers.

An overview of the HyTeC-cloze procedure is given in Table 1.[1] The motivations behind the HyTeC-cloze design are described in the next section.

Deletion strategy

Standard cloze tests follow a mechanical deletion strategy: every Xth word of the text is deleted. This is usually every fifth word, but there are studies that go as far as every 18th word (Oller & Jonz, 1994a; Watanabe & Koyama, 2008). Except for the title and first sentence (which are often left intact to provide the reader with some contextual support) and very unpredictable words (like numbers and proper names) gaps are chosen mechanically and thereby dispersed at regular intervals throughout the whole text.

Apart from these exclusions, experimenter bias is reduced to the starting point. With a deletion ratio of five, the experimenter can create five different tests. These cloze versions do not tend to differ in the proportion of lexical, syntactic and cohesive items that they sample (Bachman, 1982; Brown, [1983]1994; Jonz, 1994, O'Toole & King, 2010).[3] However, shifting the starting point changes the whole test because a different sample of words is drawn from the text (Brown, 1993). This can lead to sampling error (O'Toole & King, 2010), where cloze versions of the same text differ in levels of difficulty (see also

Table 1. Standard cloze procedure (see Oller & Jonz, 1994a) versus HyTeC-cloze procedure.

Characteristics	Standard cloze	HyTeC-cloze
1. Deletion strategy	Mechanical	Mechanical-Rational
2. Deletion ratio	20%	10%
3. Deletion distance	Fixed: every fifth word	Varied: at least 1 word in between
4. Number of gaps (per 300 words)	60	30
5. Number of starting points	3–5	2
6. Excluded text segments	• Title • First sentence	• Title • First sentence
7. Excluded words	• Proper names • Numbers	• Locally predictable words • Guess words
8. Pre-cloze or post-cloze testing[2]	Pre-cloze	Pre-cloze
9. Open or closed answer format	Open	Open
10. Marking of deletion	Fixed length marking	Fixed length marking
11. Scoring	Exact + spelling errors	Semantic + spelling errors

Alderson, 1979a; Brown, 2002; Porter, 1978). To minimize this "error," experimenters can use more than one cloze version per text (Bormuth, 1969; Porter, 1978; Staphorsius, 1994).

Another option is following a rational deletion strategy, where the experimenter selects the gaps. Selection can be limited to a specific grammatical class of words (e.g., articles, prepositions or connectors; Goldman & Murray, 1992) or based on a specific hypothesis (e.g., which type of information is needed to fill in the gap; Bachman, 1985; Gellert & Elbro, 2013; Levenston, Nir, & Blum-Kulka, 1984). In addition, rational deletion is preferred in experimental studies: it allows for direct comparisons between experimental versions of the same text. Consider a text that is syntactically manipulated as in (1), and investigated via a mechanical cloze, with the starting point at the fifth word and a deletion distance of five. This means that in (1a) the words *growing, burned* and *wounds* disappear (see (2a)), while the gaps in (1b) fall on *its, new*, and *part* (see (2b)). This would create a confound between cloze version and text version.

(1) a. The tree will, by growing new bark over the burned part, heal its own wounds.
b. The tree will heal its own wounds by growing new bark over the burned part.
(2) a. The tree will, by [……] new bark over the [……] part, heal its own [……].
b. The tree will heal [……] own wounds by growing [……] bark over the burned [……].

(Adapted from Davison & Kantor, 1982, p. 192)

In contrast, rational procedures allow the researcher to choose the gaps, thus avoiding the confound in (3). For example:

(3) a. The [……] will, by growing new bark over the burned [……], heal its own [……].
b. The [……] will heal its own [……] by growing new bark over the burned [……].

Another advantage of rational cloze tests is that only those gaps can be chosen that are indicative of text comprehension. Thus, a great amount of noise can be eliminated from the data. For example, one criticism of mechanical cloze tests is that they typically contain a considerable number of function words and that function word gaps tap grammatical knowledge of the examinee rather than text comprehension (e.g., Abraham & Chapelle, 1992; Aitken, 1977; Kobayashi, 2002a). By avoiding such words in the selection process, researchers try to create a "purer" measure of text comprehension. For instance, Gellert and Elbro (2013) and Levenston et al. (1984) selected items that indicated text coherence. Bachman (1985) selected words based on which information was necessary to reconstruct the word. He maximized the number of items that needed information across clauses or sentences.

The drawback of rational procedures is that gap selection may suffer from experimenter bias. If the rationale is not specified clearly and thoroughly, the same procedure will lead to different tests when used by different experimenters (cf. Jonz's 1994 replication of Bachman, 1985). In addition, the experimenter may (unintentionally) prefer certain words over others, resulting in a selection that does not "mirror" the text's overall difficulty (Alderson, 1979a; Klare, 1976).

The HyTeC-cloze procedure combines the strengths of mechanical and rational deletion into a hybrid deletion strategy: *mechanical-rational deletion*. First, a rational strategy is used to create a cleaner measurement of text comprehension based on theoretical and experimental considerations. Words that do not require text level comprehension are left intact. Next, a mechanical strategy is used to draw a random sample from the remaining gap candidates.

Selecting cloze gap candidates

The candidate selection for the HyTeC-cloze is based on the two heuristics presented below.[4]

Heuristics

1. Gaps cannot be too locally predictable; we exclude words that can be reconstructed by knowledge of grammar or usage conventions, as they do not require discourse level comprehension (e.g., Oller & Jonz, 1994b).
2. Gaps cannot be too unpredictable; we exclude words that can only be reconstructed with "extra-textual knowledge" (Bachman, 1985; Levenston et al., 1984).

Heuristic 1: predictable words. Following the first heuristic, many function words are excluded, including articles, prepositions and auxiliary verbs. However, function words that mark referential cohesion and discourse coherence are useful for testing comprehension

at the discourse level (Alderson, 1979a; Gellert & Elbro, 2013; Goldman & Murray, 1992; Levenston et al., 1984). Anaphoric pronouns are allowed, for the reason that they show referential coherence and often require intersentential integration. In addition, they can often be replaced by their antecedent (e.g., "*Peter was very tired. He/Peter slept until twelve o'clock.*"). In contrast, relative and interrogative pronouns are excluded, as they follow local constraints. Furthermore, we allow the deletion of most conjunctions and conjunctive adverbs, as these items measure text comprehension on an intersentential level. An exception is made for conjunctions linking noun phrases (e.g., "*Mary and Bill went to the cinema.*"), which are very predictable.

Parts of common expressions, phrasal verbs or antonym pairs are also highly predictable. Even without context, most readers will know by convention that the answer in (4) must be "*time*" and that in (5) the answer is probably "*bad*." Such words are excluded.

(4) Once upon a [……]
(5) Good and [……]

Heuristic 2: unpredictable words. The second heuristic excludes words that are not cued by the context at all. These "guess words" solely depend on extra-textual knowledge. We defined five types of guess words:

1. technical terms
2. proper names
3. units of measurement (e.g., hour, centimeter, year)
4. cardinal directions (e.g., north, west)
5. numbers

As Oller and Jonz (1994b) noted about technical terms: "Such items, if they were deleted, would normally generate little or no variance and could not therefore contribute significantly to the quality of the test" (p. 4). The same holds for the other types of guess words. Even when we accept every answer as long as it is in the same ballpark, the risk of zero or low variation is high. A pre-test confirmed this. When a date was chosen as gap, none of the participants was able to fill in an acceptable answer (e.g., another date). They left these gaps blank.

For technical terms and names, the guessing factor is only present at their first occurrence. For example, in the sentence "*This is called ADHD*," the word "ADHD" can only be known if the reader has prior knowledge on the subject or the text itself. However, if the term is repeated in the text (e.g., "*This is called ADHD. ADHD can be controlled by diet and medicine.*"), then the second mention of ADHD can be inferred from the discourse. Therefore, we exclude terms and names only the first time they are mentioned. This exception is not made for the other types of guess words, since those are, generally, not used co-referentially.

Reliability of candidate selection. The HyTeC-procedure leaves little room for experimenter bias by specifying which types of words do not adhere to the heuristics. However,

deciding whether a word combination is a common expression remains difficult. Jonz's notes show that in half of the cases that he disagreed with Bachman, the reason was that he thought that the item was a collocation or a "multi part lexical item," whereas apparently Bachman did not (Jonz, 1994, p. 321). The same problem could threaten the reliability of the HyTeC-cloze procedure. Hence, we checked the reliability of the candidate selection procedure. A student assistant with no prior knowledge of cloze testing followed the procedure and selected all possible candidates for deletion from three different texts. Agreement between his selection and our own selection was 96% (Cohen's Kappa = .93).[5]

Determining the deletion ratio

Not all candidate gaps can end up in one test. The HyTeC-procedure uses mechanical selection to select gaps out of the possible candidates. The number of words that can be sampled depends on the deletion ratio.

In mechanical cloze tests, it is standard practice to use a deletion ratio of 1 in 5 (see Oller & Jonz, 1994a).[6] A ratio of 1 in 5 is not realistic for the HyTeC-cloze, because the rational selection step provides fewer deletion candidates compared to a mechanical procedure where all words in the text are candidates. In addition, if the texts under investigation are manipulated, even fewer candidates remain, as words that differ between versions clearly cannot become gaps. Another consideration is that for less able reader groups, high ratios may result in a floor effect and no variance can be observed (Robinson, 1981; Staphorsius, 1994).

According to Greene (2001) and Bachman (1985), ratios between "1 in 9" and "1 in 11" are reasonable. We pre-tested ratios of 1 in 10 and 1 in 12 to see what the best ratio would be for our main test population (i.e., Dutch adolescents). Means and standard deviations of the total scores indicated that both were reasonable ratios for our participants. Since there was no difference between the ratios, the HyTeC-cloze uses a ratio of 1 in 10. This deletion ratio results in a higher number of observations per text.

Selecting gaps

With a ratio of 1 in 10 words, a 300-word text would require 30 gaps. If the text has 120 gap candidates, we can create four unique cloze tests, with each test sampling different words. The number of versions is thus determined by the length of the text, combined with the number of candidates.[7] Once the number of versions is determined, candidates can be distributed over the versions by counting them off. In the case of four versions, candidate 1 ends up in version 1, as does candidate 5.

A pre-test showed that all trial texts allowed the construction of at least two different cloze versions. Some texts even allowed five cloze versions. Similar to standard mechanical cloze tests, these cloze versions may vary slightly in difficulty. If we only select one version out of all possible samples, by chance we might end up with a biased sample. Therefore, we advise randomly selecting two versions out of the possible versions and using both of them.

Word repetitions. The heuristics exclude a large number of words. In the resulting smaller sample of gap candidates, some candidates might be overrepresented. In one of our pre-tests, a lemma that occurred seven times in the text ended up as a gap five times. By chance, all instances occurred in the same cloze version. To circumvent such extremes, the HyTeC-procedure uses a limit for lemma repetitions. The number of repetitions of a lemma within a cloze test should not exceed the ratio of repetitions present in the candidate sample. So, if one lemma makes up 10% of the candidates, that lemma can be chosen as a gap three times in a cloze test with 30 gaps. This relative limit prevents overrepresentation of a lemma in the sample while still allowing lemma repetitions to be reflected in the cloze test.

Answer format and scoring procedure

The HyTeC-cloze uses an open answer format with fixed-length blanks (see (3)). In contrast to closed formats such as multiple-choice, answers are not provided or cued in any way. This allows examinees to answer freely without being confused by distractors or cues (e.g., Abraham & Chapelle, 1992; Alderson, 2000) and prohibits guessing the correct answer by chance.

There are two possible scoring procedures for open cloze tests. The most efficient way of scoring is *exact scoring*, in which only originally deleted words – usually including spelling errors – are accepted. *Semantic scoring* (or *acceptable word scoring*) allows originally deleted words, but also semantically correct alternatives. The acceptability of alternative answers is usually scored according to the *global* appropriateness criterion, which means that the answer has to fulfill "all the contextual requirements of the entire discourse context in which it appears" (Oller & Jonz, 1994a, p. 416). In contrast, when the *local* appropriateness criterion is adhered to, the answer only has to fulfill the contextual requirements of the immediate sentence.

Most scholars agree that semantic scoring has higher face validity than exact scoring. When measuring text comprehension, it seems illogical to fault a reader for filling in an acceptable answer (e.g., a synonym), rather than the original word. Nevertheless, many scholars prefer the exact method for its ease, given the high correlations between exact and semantic scores ($r = .9$; Alderson, 1979a; McKenna, 1976; Miller & Coleman, 1967; Staphorsius, 1994).

We opt for semantic scoring following the global appropriateness criterion for several reasons. First of all, our pre-tests showed moderate correlations between exact and semantic scoring (range .759 to .873) and were not nearly as high as correlations reported in previous studies.[8] In addition, it is unknown whether the .9 correlations hold for all types (or combinations) of items, as well as for texts and readers of all levels. For instance, McKenna (1976) found that semantic scoring benefitted high-proficiency students more than low-proficiency students (see also Brown (2002) for similar results for L2 students). McKenna also found that semantic scores correlated significantly higher with reading comprehension scores from the Stanford Achievement Test than exact scores, indicating higher concurrent validity of semantically scored cloze tests. However, O'Toole and King (2011) warned against semantic scoring, since

it may lead to an underestimation of text difficulty and an overestimation of reading competence. Although O'Toole and King (2011) made a valid point with regard to anchoring and floor/ceiling effects, in our view the total reverse can be said for exact scoring. That is: it may lead to an overestimation of text difficulty and an underestimation of reading competence. Given the results of these studies, it seems ill-advised to generalize over scoring methods. Although high correlations have been found, exact scoring is not equivalent to semantic scoring.

Furthermore, semantic scoring has been shown to be more reliable than exact scoring. In a meta-analysis of 24 ESL/EFL-studies, Watanabe and Koyama (2008) found a mean reliability estimate of .74 for semantic scoring ($k = 97$) compared to .64 for exact scoring ($k = 122$). In addition, reliability estimates for semantic scoring were more stable, ranging from .60 to .97, whereas exact scoring ranged from .14 to .99.

Thus, both from a conceptual as from a statistical point of view, semantic scoring is more suited for measuring text comprehension than exact scoring. That is why, although it is much more time consuming, answers will be scored semantically.[9]

Evaluation of the HyTeC-cloze procedure

In this final section, we evaluate the HyTeC-cloze procedure on the basis of the results of a large-scale study on 60 Dutch texts. Two cloze versions were created for each text (i.e., 120 unique tests). The cloze tests contained 30–42 cloze gaps depending on the text length (range 300–420 words) and were presented digitally on computers (see Appendix). They were administered to 2926 Dutch secondary school students in grades 8–10. Most students filled in a total of four cloze tests divided over two sessions. Students were enrolled in different levels of the Dutch education system, ranging from the lowest pre-vocational level ("vmbo-bb") to pre-university level ("vwo").[10] Answers were scored by two independent judges and a possible third judge to cast the deciding vote.

The data was used to address issues related to scoring procedure, internal reliability, response rates, sensitivity to local constraints, concurrent validity, and known-group validity.

Semantic versus exact scoring

The data was scored using the semantic scoring procedure outlined in the "Answer format and scoring procedure" section above and using an exact scoring procedure in order to compare the results. The exact and semantic scores correlated highly ($r_s = .862$; $p < .001$), but not as highly as previously reported (cf. the "Answer format and scoring procedure" section). Furthermore, the correlation was not completely stable; it decreased in strength going from the lowest level of education to the highest (from .859 to .789) and varied between individual cloze tests (from .637 to .951; see Table 2). For completeness purposes we will report findings for both the semantic scoring method and the exact scoring method wherever possible, but the semantic score outperformed the exact score in all tests.

Table 2. Summary of Spearman's rho correlations calculated over cloze test versions ($k = 120$).

Correlation	*M* r_s	*SD* r_s	*Mdn* r_s	Min. r_s	Max. r_s
Semantic scoring/Exact scoring	.848	.061	.855	.637	.951

Table 3. Summary of internal reliability scores calculated over cloze test versions.

Scoring method	*M* α	*SD* α	*Mdn* α	*Min.* α	*Max.* α
Semantic	.828	.038	.831	.707	.899
Exact	.738	.075	.742	.519	.894

Internal reliability

For each cloze test version ($k = 120$), internal reliability was measured using Cronbach's alpha. A summary of the results is given in Table 3. Semantic scoring and exact scoring were both relatively reliable, but semantic scoring was systematically more reliable. In addition, semantic scoring was more stable across cloze tests and never dropped below .70. These scores are high, especially given the fact that many studies have reported dramatically low alphas for cloze tests (Brown, 2013; Watanabe & Koyama, 2008).

Response rates and data loss

The validity of any test is threatened when test takers do not answer seriously or when they do not answer at all. We explored our data to see whether these threats were present.

First, we performed a qualitative check on the cloze gaps that were left blank (9.3%). Most gaps seemed to be left blank because the student did not know the answer to that particular gap. These blanks were dispersed throughout the test and surrounded by serious answers. One third of the gaps was probably left blank because the student was not motivated enough to continue or ran out of time. The surrounding gaps were left blank as well or throughout the test the student filled in mostly nonsense answers. Out of all the blanks, only these cases are considered to be real cases of data loss. We checked cases where students did fill in something but the answer did not seem to be serious (e.g., "adfgd" or "…") in the same way. Tests containing comments such as "I hate this test" were removed altogether. In total, 9.66% of the filled out cloze tests was removed. Finally, we checked whether data loss was equally distributed over students differing in reading ability. We compared the standardized readability scores of the students that were removed to those of the students that remained in the dataset and found no significant difference ($F(1,4756) = 0.028$; $p = .866$). The data loss did not result in an underrepresentation of low-ability students.

Table 4. Means, standard deviations and medians for forward and backward log-probability.

Words	Forward log-probability			Backward log-probability		
	M	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Overall ($N = 46274$)	−2.468	1.652	−2.091	−2.461	1.790	−2.190
Not a cloze gap ($N = 38456$)	−2.239	1.573	−1.826	−2.202	1.722	−1.864
Cloze gaps ($N = 7818$)	−3.594	1.566	−3.528	−3.734	1.556	−3.804

Level of comprehension measurement

The HyTeC-cloze test was designed to measure text comprehension. Other cloze tests have been criticized because their gaps seem to rely more on local linguistic predictability (on the basis of grammatical knowledge and knowledge of collocations) than intersentential or context dependent comprehension. Since predictable words are not included as gaps in the HyTeC-cloze (see the "Heuristic 1: predictable words" section), the relation between local predictability and cloze scores should be weak. Two analyses were done to check this claim.

First, we examined whether the HyTeC-cloze procedure was successful in selecting cloze gaps that were not highly locally predictive. If the procedure was successful, the words that were used as cloze gaps should have a lower local probability compared to words that were not turned into cloze gaps. T-Scan – a tool for automatic Dutch text complexity analysis (Pander Maat et al., 2014) – was used to determine the forward log-probability (probability of $Word_N$ given $Word_{N-2}$ and $Word_{N-1}$) and backward log-probability (probability of $Word_N$ given $Word_{N+1}$ and $Word_{N+2}$) of all words. The probabilities of words that were used as cloze gaps were significantly lower than the probabilities of words that were not used as cloze gaps (see Table 4; forward probability: $U = 77859698.500$; $z = -67.300$; $p < .001$; $r = -.31$; backward probability: $U = 75133694.500$; $z = -69.842$; $p < .001$; $r = -.32$). Reversing the $\log_{10}$-transformation shows us that the words used as cloze gaps are on average almost 23 times less probable (based on the two words preceding them) compared to non-clozed words. Based on the two words following them, they are 34 times less probable compared to the non-clozed words. The selection of non-locally predictive words was thus successful.

Second, the probability measures were entered in a logistic regression to see how much variance they can explain as predictors of the cloze scores. Combining the measures gives us a window of four words surrounding the cloze gap. Together forward and backward log-probability only explained 2.4% of the variance observed in the semantic scores and 7.3% for the exact scores (see Table 5). The difference between semantic and exact scores was to be expected. The probability measures indicate the probability of the exact word that was deleted in the text, not the probability of all semantically correct words that could occur there. Therefore, we expected the probability measures to explain more variance for exact scores than for semantic scores. Yet, even for exact scores the explained variance is low and it is therefore very unlikely that the HyTeC-cloze test only measures local level predictability.

Table 5. Explained variance by log-probability measures at item level.

Scoring method	Model	Explained variance (r^2_N a)
Semantic	Forward and backward probability	.024
	Forward probability	.022
	Backward probability	.018
Exact	Forward and backward probability	.073
	Forward probability	.068
	Backward probability	.046

Table 6. Summary of Spearman's rho correlations calculated over cloze test versions ($k = 120$).

Correlation	*M* r_s	*SD* r_s	*Mdn* r_s	Min. r_s	Max. r_s
Semantic score – Reading ability	.606	.137	.621	.161	.856
Semantic score – Vocabulary	.604	.125	.609	.220	.839
Exact score – Reading ability	.564	.154	.587	.047	.832
Exact score – Vocabulary	.558	.143	.569	.105	.870

Concurrent validity

Correlations of the cloze scores with other text comprehension measures or with standardized ability tests, can give us an idea of the convergent validity of our cloze test: do the tests measure the same construct? For the majority of the students, standardized reading ability and vocabulary scores were available. These scores were collected using the reading ability and vocabulary tests from the test batteries RSM14 and VVO ("Student monitoring system secondary education") developed by Cito ("The Dutch institute for Educational Measurement"), which were administered at the same time or within 90 days of the cloze tests.[11]

The summed semantic cloze scores[12] correlated on average .606 with the reading ability scores and .604 with the vocabulary scores (see Table 6).[13] The exact cloze score correlated slightly lower (reading ability: *mean* r_s= .564; vocabulary: *mean* r_s= .558; cf. McKenna, 1976). Given that well-established, standardized tests of reading ability have been found to correlate moderately with each other at between .31 and .79 (Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008); a mean correlation of .60 (with an even higher median) suggests that the HyTeC-cloze test does not underperform compared to other established reading ability tests.

In addition to the standardized test scores, we have some data that can indicate how our participants would have performed had we used the same texts but a different assessment method. A selection of eight texts was used in an eye-tracking study (see Kleijn, 2018). Within this study, text comprehension was assessed with multiple-choice questions. One hundred and eighty-one ninth-grade students participated in this study. After reading each text, the students answered eight questions that appeared one-by-one on the

computer screen. The questions were designed to assess comprehension of the main points of the text. The students were not able to look back in the text when answering these questions. Mean scores were calculated per text and education level and then compared to the corresponding mean cloze scores. The multiple-choice score correlated .525 ($p = .008$) with the semantic cloze scores and .389 ($p = .060$) with the exact cloze scores.

Known-group validity

A good assessment method has to be sensitive to known differences in comprehension levels (i.e., known-group validity). The readers in the sample were enrolled in different levels of education and differ in age. Furthermore, 30 texts that were clozed were taken from school books written for different education levels and grades, the other 30 were public information texts. Thus, we have strong reasons to believe there should be a lot of variance in the sample: between students and between texts. The HyTeC-cloze must be sensitive enough to show this variance. It must be able to discriminate between students with different reading abilities and it must also be able to discriminate between different texts and even between text versions in experimental setups.

First, we investigated the overall amount of variance in the cloze scores by plotting the frequency distributions. If the data is normally distributed and there is no evidence of either floor or ceiling effects, we can proceed to investigate the expected variance between groups.

As shown in Figure 1, the frequency distribution of the semantic scores was close to normal with a mean score of 16 out of 30 items correctly answered. The mean exact score was of course lower (semantic score = exact answers + semantically correct answers) and the distribution had a heavier left-tail. These are distributions of all students. When we compare distributions of the different education levels in the sample, we find that exact scoring was particularly problematic for the lowest levels of the Dutch education system with many observations of zero or close to zero. Again, semantic scores were more normally distributed and show variation especially in the lower education levels.

A lot of variance exists in the sample, but can we attribute this variance to known differences between students and/or texts? Linear mixed effects modeling was used to answer this question. The data was hierarchically structured: students were nested in schools and cloze tests were nested in texts and semi-crossed with students. This structure was tested in a stepwise procedure. *Level of education* and *Grade* were introduced as fixed factors. The final model is shown in Table 7.

All factors improved the fit of the model. As expected, scores varied between students, texts and slightly between cloze versions. The fixed factors of *level of education* and *grade* explained a large part of the variance between students. The analysis showed significant differences in cloze scores between all education levels and grades. The effects were in the expected direction: cloze scores were higher for students enrolled in higher education levels and grades. The analysis shows that the HyTeC-cloze is sensitive to these known differences.

In addition to different texts, the data also included two different versions of each text. Texts were manipulated in one of three ways: (1) words were substituted for less or more

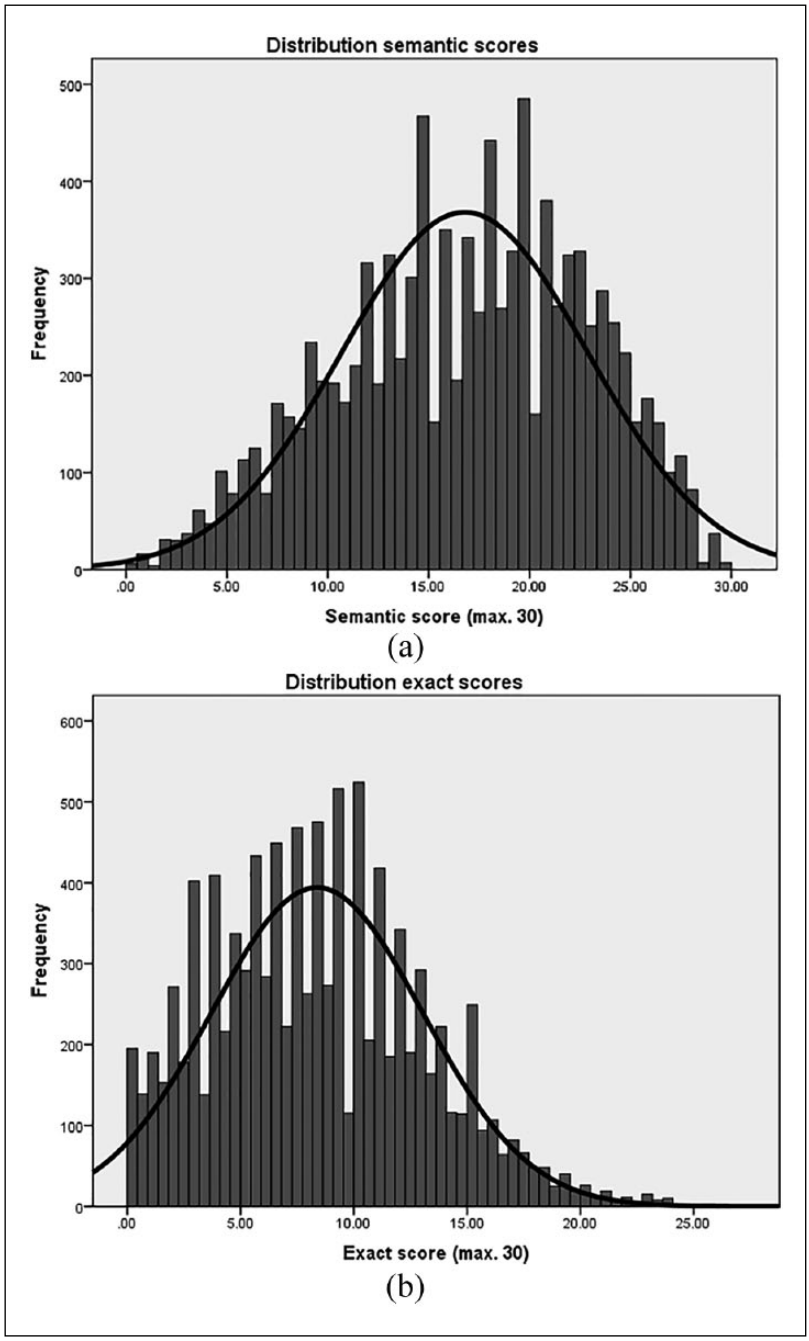

Figure 1. Frequency distribution of semantic and exact summed scores.

Table 7. Final model semantic score.

Random effects	Estimates	*SD*		
School	1.562	1.250		
School: Student	5.838	2.416		
Text	10.586	3.254		
Text: Cloze version	1.483	1.218		
Residual	8.357	2.891		
Fixed effects	**Estimates**	***SE***	***t*-value**	***p***
Intercept	9.836	0.544	18.097	<.001
Education level: pre-voc. low	0[a]			
Education level: pre-vocational medium	2.620	0.257	10.207	<.001
Education level: pre-vocational high	4.934	0.271	18.200	<.001
Education level: general	7.555	0.341	22.139	<.001
Education level: pre-university	9.866	0.323	30.547	<.001
Grade 8	0[a]			
Grade 9	1.159	0.149	7.771	<.001
Grade 10[b]	1.982	0.307	6.462	<.001

[a]Set as reference level. [b] Unbalanced, no 10th-grade pre-vocational students were present in the sample.

familiar alternatives; (2) syntactic dependency lengths were increased or decreased; or (3) connectives were removed or added. Each manipulation was expected to influence text comprehension. In separate analyses, we tested whether the HyTeC-cloze scores reflected these subtle manipulations of text difficulty. For reasons of space we will not go into detail here and will only point out that the HyTeC-cloze was sensitive to text version differences (see Kleijn, 2018). However, not all effects could be observed in the summed cloze score. For some manipulations, it was necessary to zoom in on the gaps directly surrounding the manipulation to find a significant effect. This shows that the HyTeC-cloze is also capable to detect very small, localized effects.

Discussion

Cloze is a popular assessment method in readability studies and language proficiency testing, but it has never been widely accepted as a valid measure of text comprehension. According to its critics, cloze tests are "beset with problems" (Klein-Braley & Raatz, 1984, p. 134). The critics' biggest concern seems to be that gaps can be answered correctly without understanding the text. They claim that localized, low-level processing is enough to fill in the gaps successfully and that it is not necessary to integrate sentences into a discourse level representation. However, most cloze "problems" do not hold for all types of cloze tests and can be successfully addressed in cloze design. In this paper we presented such an improved cloze procedure: the HyTeC-cloze. This hybrid cloze procedure combines the strengths of mechanical and rational cloze tests. The rational strategy is used to exclude words that do not rely on text-level comprehension from becoming

cloze gaps (e.g., articles, copula, multi-word expressions, and guess words). The remaining words in the text are candidates for deletion and a sample of them is mechanically selected. This procedure results in a cloze test that has a very low sensitivity to local predictability and is still fast and easy to produce since it does not require an in-depth analysis of the texts. Furthermore, the HyTeC-procedure is widely applicable. It can be used to assess a wide range of texts without confounding text difficulty with question difficulty and is suitable for test takers of high and low ability provided that semantic scoring is used. Most importantly, our results show that the HyTeC-cloze matches and sometimes even outperforms standardized tests of reading ability when it comes to validity and reliability. These qualities, together with its sensitivity to discriminate between texts, text versions and readers, make the Hybrid Text Comprehension cloze an appealing method for experimental and correlational studies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research presented here was part of the LIN-project (“Readability index for Dutch”) funded by NWO (“The Netherlands Organisation for Scientific Research”), Cito (“The Dutch Institute for Educational Measurement”) and Nederlandse Taalunie (“The Dutch Language Union”) [NWO grant number 321 89 002].

ORCID iD

Suzanne Kleijn https://orcid.org/0000-0003-3613-1507

Notes

1. A stepwise construction manual is available upon request; please contact the first author.
2. In a post-cloze test the reader reads the original non-clozed text before making the cloze test. In pre-cloze tests the first time the reader sees the text, the gaps are already in place.
3. As long as the test is long enough the sample will naturally approach the distribution of word types in the text.
4. Although there are similarities with Bachman’s levels of constraint (within clause; across clause but within sentence; across sentence; extra-textual; Bachman, 1985), our classification is not centered on clause or sentence boundaries.
5. Calculated over three different texts, with 1052 words in total.
6. Increasing the context surrounding a specific gap does not seem to influence the score (Alderson, 1979b, MacGinitie, 1961; Rankin & Thomas, 1980; Taylor, [1956]1994), but only for analyses where just the items that are present in all deletion ratio versions are compared. Some scholars have found an effect of deletion ratio on relative total scores, but the direction of this effect seems unpredictable (see Alderson, 1979b).
7. The number of remaining candidates is divided by the number of gaps to be chosen to see how many versions can be made.
8. This difference was to be expected since our cloze procedure does not allow predictable (closed class) grammatical items, while the standard cloze tests in these studies do.
9. Including spelling and typing errors.

10. The Dutch system distinguishes multiple levels of education. Going from practice-oriented education to academically-oriented education, the levels included in the study are as follows: vmbo-bb, vmbo-kb, vmbo-gt, havo, and vwo.
11. The tests were written multiple-choice tests. The vocabulary tests measure the size and depth of the receptive vocabulary knowledge. The reading ability tests measure the ability to understand written texts and to reflect on text function and goals. Questions target local and global processes, including comprehension of the main ideas of the texts, comprehension of words, sentences, paragraphs in context and the relations that hold between them, and understanding differences between facts, opinions, claims, arguments, and conclusions (Van Til & Van Boxtel, 2015).
12. The summed scores were calculated by adding up the scores of the cloze gaps for each participant for each cloze test. Because cloze tests had a different number of gaps, the summed scores were normalized to a 30-gap test.
13. The standardized reading ability and vocabulary scores correlated .569 with each other.

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. doi:10.1111/j.1540–4781.1992.tb05394.x

Aitken, K. G. (1977). Using cloze procedure as an overall language proficiency test. *TESOL Quarterly*, *11*(1), 59–67.

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

Alderson, J. C. (1979b). The effect on the cloze test of changes in deletion frequency. *Journal of Research in Reading*, *2*(2), 108–119. doi:10.1111/j.1467–9817.1979.tb00198.x

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

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly*, *16*(1), 61–70. doi:10.2307/3586563

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535–556. doi:10.2307/3586277

Bormuth, J. R. (1969). *Development of readability analyses*. (Final Report, Project No. 7–0052, Contract No. 1, OEC-3–7–070052–0326). Office of Education, U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research.

Brown, J. D. ([1983]1994). A closer look at cloze validity. In J. W. Oller & J. Jonz (Eds.), *Cloze and Coherence* (pp. 189–196). Lewisburg, PA: Bucknell University Press. (Reprinted from J. W. Oller (Ed.), *Issues in Language Testing* (pp. 237–250). Rowley, MA: Newbury House, 1983).

Brown, J. D. (1993). What are the characteristics of natural cloze tests? *Language Testing*, *10*(2), 93–116. doi:10.1177/026553229301000201

Brown, J. D. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies*, *21*(1), 79–125.

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what. *International Journal of Language Studies*, *7*(1), 1–32.

Chávez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W. (1994). When are cloze items sensitive to constraints across sentences? In J. W. Oller, & J. Jonz (Eds.), *Cloze and coherence* (pp. 229–245). Lewisburg, PA: Bucknell University Press. (Revised from *Language Learning*, *35*(2), 181–206, 1985.)

Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format interactions: A reply to Kobayashi, 2002. *Language Testing*, *21*(2), 228–234. doi:10.1191/0265532202lt227oa

Chihara, T., Oller, J. W., Weaver, K. A., & Chávez-Oller, M. A. (1994). Are cloze items sensitive to constraints across sentences? In J. W. Oller, & J. Jonz (Eds.), *Cloze and coherence* (pp. 135–147). Lewisburg, PA: Bucknell University Press. (Revised from *Language Learning*, *27*, 63–73, 1977.)

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277–299. doi:10.1207/s1532799xssr1003_5

Cziko, G. A. (1983). Another response to Shanahan, Kamil, and Tobin: Further reasons to keep the cloze case open. *Reading Research Quarterly*, *18*, 361–365.

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, *17*(2), 187–209. doi:10.2307/747483

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.

Goldman, S. R., & Murray, J. D. (1992). Knowledge of connectors as cohesion devices in text: A comparative study of native-English and English-as-a-second-language speakers. *Journal of Educational Psychology*, *84*(4), 504–519.

Greene, B. B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, *24*(1), 82–98.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Henk, W. A. (1982). A response to Shanahan, Kamil, and Tobin: The case is not yet clozed. *Reading Research Quarterly*, *17*(4), 591–595.

Jonz, J. (1994). Cloze item types and constraint on response. In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 317–344). Lewisburg, PA: Bucknell University Press.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281–300.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, *74*(6), 828–834.

Klare, G. R. (1976). A second look at the validity of readability formulas. *Journal of Literacy Research*, *8*(2), 129–152. doi:10.1080/10862967609547171

Kleijn, S. (2018). *Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing*. Utrecht: LOT.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing*, *1*(2), 134–146.

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

Kobayashi, M. (2002b). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*(2), 193–220.

Kobayashi, M. (2004). Investigation of test method effects: Text organization and response format. A response to Chen, 2004. *Language Testing*, *21*(2), 235–244.

Levenston, E. A., Nir, R., & Blum-Kulka, S. (1984). Discourse analysis and the testing of reading comprehension by cloze techniques. In A. J. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes: Studies and practices in native and foreign languages* (pp. 202–212). London: Heinemann Educational Books.

MacGinitie, W. H. (1961). Contextual constraint in English prose paragraphs. *Journal of Psychology*, *51*, 121–130.

McKenna, M. (1976). Synonymic versus verbatim scoring of the cloze procedure. *Journal of Reading*, *20*(2), 141–143.

McKenna, M. C., & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension. *Journal of Educational Psychology*, *82*(2), 372.

Miller, G. R., & Coleman, E. B. (1967). A set of thirty-six prose passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 851–854.

Oller, J. W., & Chen, L. (2007). Episodic organization in discourse and valid measurement in the sciences. *Journal of Quantitative Linguistics*, *14*(2–3), 127–144.

Oller, J. W., & Jonz, J. (Eds.). (1994a). *Cloze and coherence*. Lewisburg, PA: Bucknell University Press.

Oller, J. W., & Jonz, J. (1994b). Why cloze procedure? In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 1–20). Lewisburg, PA: Bucknell University Press.

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.

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.

Pander Maat, H., Kraf, R., Van den Bosch, A., Van Gompel, M., Kleijn, S., Sanders, T., & Van der Sloot, K. (2014). T-scan: A new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal*, *4*, 53–74.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices – Past, present, and future. In S. G. Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Mahwah, NJ: Lawrence Erlbaum.

Porter, D. (1978). Cloze procedure and equivalence. *Language Learning*, *28*(2), 333–341. doi:10.1111/j.1467–1770.1978.tb00138.x

Rankin, E. F., & Thomas, S. ([1980]1994). Contextual constraints and the construct validity of the cloze procedure. In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 165–175). Lewisburg, PA: Bucknell University Press. (Reprinted from M. L. Kamil & A. J. Moe (Eds.), *Perspectives on reading: Research and instruction* (pp. 47–55). Washington, DC: National Reading Conference, 1980).

Robinson, C. G. (1981). Cloze procedure: A review. *Educational Research*, *23*(2), 128–133. doi:10.1080/0013188810230206

Shanahan, T., Kamil, M. L., & Webb Tobin, A. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*(2), 229–255.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument*. Arnhem: Cito.

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

Taylor, W. L. ([1956]1994). Recent developments in the use of cloze procedure. In J. W. Oller & J. Jonz (Eds.), *Cloze and coherence* (pp. 81–90). Lewisburg, PA: Bucknell University Press. (Reprinted from *Journalism Quarterly*, *33*, 42–48, 1956).

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across learner backgrounds. *Language Testing*, *34*(2), 151–174. doi:10.1177/0265532215623581

Van Til, A., & Van Boxtel, H. (2015). *Wetenschappelijke verantwoording Toets 0 t/m 3, tweede generatie*. Arnhem: Cito. Retrieved from https://www.cito.nl/-/media/Files/kennisbank/cito-bv/96_wetenschappelijke-verantwoording-volgsyteemvo-gen2.pdf?la=nl-NL

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies*, *26*(2), 103–133.

Het landelijke fietsdiefstalregister

In het landelijke fietsdiefstalregister kunt u controleren of een fiets als gestolen staat geregistreerd. Heel handig als u een tweedehandsfiets wilt kopen, of ter controle van de aangifte bij de politie van uw gestolen fiets. In het register vindt u vooral aangiftes van na 1 januari 2008. Aangiftes van vóór 1 januari 2008 zijn slechts beperkt verwerkt in het register .

Om te controleren of een fiets als gestolen staat geregistreerd heeft u een _____ van het framenummer en het _____ van de fiets nodig. Een _____ mogelijkheid is met behulp van het _____ , ook wel diefstalpreventiechip (dpc) of protagtor. Als de _____ als gestolen staat geregistreerd is het antwoord ja. Staat de fiets _____ als gestolen geregistreerd dan is het _____ nee. Bij 'nee' kan de fiets _____ nog steeds uit diefstal afkomstig zijn. Want op dit moment wordt slechts in 15% van de gevallen van _____ aangifte gedaan bij de politie. En als er _____ aangifte wordt gedaan, is een fiets niet herkenbaar als gestolen. Het kan ook zijn dat de _____ nog niet is verwerkt in het fietsdiefstalregister. De politie heeft namelijk 10 dagen de tijd om te _____ of een aangifte voorzien is van de juiste _____ zoals framenummer en/of chipnummer. _____ daarna wordt de fiets als gestolen _____ .

Om het probleem van fietsdiefstal uit de wereld te helpen is het van belang om aangifte te doen. Aangifte doen is _____ zinvol! Alleen dan wordt een fiets herkenbaar als "_____" en kunnen de dief, de heler en de fiets worden _____ . De politie kan pas in actie komen als er aangifte is gedaan. Dus als u _____ slachtoffer wordt van fietsdiefstal, doe _____ altijd aangifte. Want door aangifte te doen maakt u de fiets _____ als gestolen. Bovendien vergroot u de _____ dat de fiets weer terugkomt. 7 procent van de _____ leidt namelijk tot terugkeer van de fiets bij de eigenaar. Dus _____ uw fiets en zorg dat _____ de unieke kenmerken van de fiets _____ framenummer, merk of chipnummer kent of op een _____ plek bewaart. Want _____ met deze unieke kenmerken kunt u een _____ aangifte doen en wordt uw fiets als gestolen geregistreerd in het _____ . Vervolgens kan de _____ aan de slag met de opsporing van de fiets en de _____ .

Appendix. Example of a HyTeC-cloze test as presented on screen. Fields were given a lavender background color and were underlined. All fields were blank when the participant started. (Source original text: RDW; (www.rdw.nl/Particulier/Paginas/Fiets.aspx).