Wojciech Malec, John Paul II Catholic University of Lublin, Poland

Using WebClass for Reduced Redundancy Testing

Abstract

This article examines and reviews two types of reduced redundancy tests, namely cloze tests and C-tests, which involve completing a text from which certain units (whole words or their parts) have been removed. Assessment instruments of this kind are typically used to measure overall language proficiency, for example for the purpose of making placement decisions. The paper also discusses the development of these two measures of reduced redundancy with the help of the WebClass testing system.

Keywords: language testing, reduced redundancy, cloze, C-test, WebClass

1. Introduction

Redundancy can be defined as the degree to which language as a whole or a specific linguistic unit contains "more formal clues than are strictly necessary to convey its message" (Delbridge & Bernard, 1966, p. 107). It is possible to express this degree numerically "as the fraction of the letters in a written text, or the fraction of the phonemes in a spoken utterance that could be eliminated if no random disturbances were present" (Fano, 1961, p. 261). In the case of English printed texts, redundancy is estimated to be approximately 50 percent (Claiborne, 1990; Shannon, 1948). This means that in optimum communication conditions, a message written in English should be understandable even if roughly one-half of the letters were obscured.

By way of illustration, redundancy manifests itself in the sentence *I met Susan yesterday*, where past tense is indicated not only by the verb form but also by the adverbial of time. When the same information is signalled by more than one element, redundancy may be perceived as "information overlap" (Horning, 1987, p. 18). In addition to this, redundancy may be regarded as unnecessary repetition. For example, compared to the English phrase *this little cat*, its Polish equivalent *ta mała kotka* redundantly repeats grammatical agreement as each word is marked for feminine gender. Repetition is also a defining feature of pleonasm and tautology, as in *return back* or *each and every* (Lehmann, 2005, p. 121). Nevertheless, redundancy should not be associated with uselessness. On the contrary, when com-

Wojciech Malec, Katedra Językoznawstwa Teoretycznego, Instytut Językoznawstwa, Katolicki Uniwersytet Lubelski Jana Pawła II, Aleje Racławickie 14, 20-950 Lublin, wojciech.malec@kul.pl, https://orcid.org/0000-0002-6944-8044

munication is hampered by noise, "redundancy enables us to reconstruct missing components of messages" (Oakes, 1998, p. 62). Thus, rather than being useless, redundancy is an essential property of language.

Natural language is a redundant system because it allows the speakers to communicate even when a substantial part of the signal is absent or deformed. In other words, knowing a language means being able "to understand a distorted message, to accept a message with *reduced* redundancy" [emphasis added] (Spolsky, 1968, p. 10). More specifically, it means being able to *restore* those parts of a linguistic code which have been omitted. This particular ability has traditionally been assessed using two language testing procedures, namely the cloze test and the C-test.

2. From cloze tests to C-tests

The word *cloze* is a derivative of *closure*, which in Gestalt psychology refers to the human tendency to perceive an incomplete form as a whole (Renkema, 2004, p. 178). The cloze test is a type of gap-filling procedure which consists in completing a text from which some words have been removed. Although methods of this kind were used by psychologists in the late nineteenth century to test intelligence (Harris, cit. in Oller & Jonz, 1994, p. 10), the first official application of the term *cloze* is attributed to Taylor (1953), who used it to name a device for measuring text difficulty (or 'readability').

Cloze tests may be of several different kinds, depending on the deletion strategy as well as the type of expected response. The standard technique, known as *fixed-ratio cloze*, is to delete the words mechanically at regular intervals (e.g. every eighth word). In *variable-ratio* procedures, by contrast, the words to be replaced with blanks are selected either *randomly* or on the basis of *rational* decisions (e.g. all adjectives or prepositions). Instead of individual words, longer strings of text could be deleted, yet this is not common practice. When the test takers are expected to fill in the blanks by producing their own responses, the task may be referred to as *open* cloze. The other possibility is to provide several alternatives for each deleted word and thus obtain a multiple-choice variant of cloze (sometimes called 'maze', e.g. Cash & Schumm, 2006, p. 266). It should be added that the beginning (and perhaps also the ending) of the passage can be left unchanged so as to allow a better understanding of the text.

Whereas the earliest cloze tests were predominantly used to measure readability and comprehension (e.g. Beard, 1967; Bormuth, 1968, 1969; Culhane, 1970; Porter, 1976; Rankin & Culhane, 1969; Schneyer, 1965), subsequent implementations of the method were more frequently employed as measures of global language proficiency (e.g. Oller, 1972; Oller, 1979; Oller & Conrad, 1971; Stubbs & Tucker, 1974). In contrast to discrete-point tests, the cloze procedure was widely acknowledged as an integrative form of assessment (Hale et al., 1989). This view was based on the fact that cloze tests appeared "to be highly correlated with virtu-

ally every other type of language test, and with tests of nearly every language skill and component" (Bachman, 1982, p. 61). Quite understandably, cloze tests were generally recommended as language proficiency tests which were relatively easy to create and administer (Aitken, 1977).

On the other hand, cloze tests did not really prove to be solely measuring language competence. For example, Byrne, Feldhusen, and Kane (1971) found cloze test scores to be related to measures of IQ. Another problem was that many cloze tests lacked consistency from one administration to another in terms of both measurement properties and criterion-related validity. In particular, the reliabilities of cloze tests and their relationships with criterion measures of ESL proficiency often varied significantly (for references, see Brown, 1988, 1993, 2002). Furthermore, a study reported in Alderson (1979) exposed some specific inconsistencies between various modifications of the cloze procedure. First, the deletion frequency had a significant effect on the ability being measured. Second, tests with difficult texts were better correlated with proficiency measures while tests with easy texts were better correlated with dictation. Finally, acceptable-answer scoring resulted in higher validity than did exact-answer scoring. This apparent unpredictability of cloze test scores significantly detracted from the attractiveness of the procedure as a unitary technique previously recommended as a measure of overall language proficiency.

In addition to the above, a fundamental drawback of the cloze method as a measure of L2 proficiency was revealed by Lado (1986), who found that even native speakers may struggle with tests of this kind. This study additionally demonstrated that changing the scoring scheme from exact to acceptable answer failed to improve native speakers' test performance. In view of the fact that it seemed to be absurd to assert that native speakers of English did not know English, Lado (1986) concluded that it was unacceptable for a language test to have "such a heavy non-language factor" (p. 132).

All the problems with cloze tests led many test developers to question the usefulness of this procedure. According to Brown (2002), for example, the standard (fixed-ratio) cloze "is far too inefficient for responsible use in decision-making" (p. 110). Indeed, rational deletion may be a better option (Bachman, 1985) as this variant of the cloze technique gives the test developer more control over the content validity of the test. On the other hand, "the rational deletion cloze is neither a test of reduced redundancy nor a measure of general language competence" (Grotjahn, 2013, p. 121). It seems that those who actually need a language proficiency test can do better by opting for the C-test.

C-tests were introduced in 1981 (Raatz & Klein-Braley, cit. in Klein-Braley, 1985, p. 81) as a solution to the well-documented problems with cloze tests. This more recent method of assessment is not radically different from the cloze test: both involve restoring a mutilated text whose beginning and ending are typically left unchanged. However, there are also some essential differences. First, the

standard form of the C-test is constructed by deleting the second half of every second word (in accordance with the so-called *C-principle*). Another difference is that a complete C-test consists of *several* short passages (but see also Khodadady, 2013, who argues that this should not be viewed as an absolute rule). Moreover, when estimating internal consistency reliability, each passage of text is analysed as a superitem (Raatz, 1985). It is worth adding that in constructing C-tests we normally delete the longer parts of those words that consist of an odd number of letters, and single-letter words are omitted, i.e. left intact.

Compared to the cloze procedure, the C-test offers several important benefits, conveniently summarized by Klein-Braley (1997). First, C-tests require considerably shorter texts than do cloze tests. Second, C-tests can be scored objectively using exact-answer scoring because in the majority of cases only one correct answer is possible. Third, the scoring of C-tests is fast and simple, almost like reading the text. Fourth, native speakers have no problem restoring half-deleted words, whereas people who do not know the language at all typically obtain a score that is close to zero. Fifth, the fact that every second word is mutilated is likely to give rise to a better representation of all the word classes in the text. Sixth, the use of several different texts ensures more adequate sampling of content types. This means that "[e]xaminees who happen to have special knowledge in certain areas no longer have substantial advantages over other examinees" (Klein-Braley, 1997, p. 65).

Admittedly, the C-test is not an absolutely perfect testing technique, and even some of the above benefits have been challenged in the literature. For example, contrary to claims that C-tests are very easy for native speakers, Jafarpur's (1995) study demonstrated that this was true only in the case of one C-test (out of a total of twenty). Huhta (1996), similarly, concluded that certain C-tests "can be very difficult even for native speakers" (p. 215). In another study, the C-principle (i.e. the rule of two) produced tests containing non-functioning items (Jafarpur, 1999b). Moreover, Jafarpur (1999a) used classical item analysis to select the best-performing items and develop an improved version of a C-test, but this did not really have a positive effect on the statistical properties of the test. In addition to these potential problems, some studies have reported rather unfavourable reactions to C-tests on the part of the test takers. For example, in a study by Bradshaw (1990) these reactions were largely negative, and Jafarpur (1995) stated that "C-tests [were] irritating and unacceptable to the subjects" (p. 209). Findings such as these contradict claims by Klein-Braley and Raatz (1984) that C-tests have face validity. In truth, the problem may well be due to the novelty of the format, which is still rarely used in popular textbooks for English language learners. On the plus side, C-tests tend to have satisfactory reliability coefficients (e.g. Coleman, 1996; Klein-Braley & Raatz, 1984; Linnemann & Wilbert, 2010, and many others).

In terms of the construct being measured and the purpose of testing, the C-test bears considerable similarity to the cloze procedure: it was originally intended as

a measure of overall language ability. Indeed, a recent meta-analysis of Pearson's correlation coefficients between C-test and criterion-test scores by McKay et al. (2021) indicates that C-tests are most strongly correlated with general language proficiency. Over the years, C-tests have been found to correlate significantly with various language proficiency measures, such as the Iowa State University English Placement Test (EPT) (Chapelle & Abraham, 1990), the Test of English for International Communication (TOEIC) (Dörnyei & Katona, 1992), TOEFL (Babaii & Ansary, 2001), and Test of German as a Foreign Language (TestDaF) (Eckes & Grotjahn, 2006). As for the purpose of C-tests, they are considered to be particularly useful in situations where selection or placement decisions need to be made (e.g. Mozgalina & Ryshina-Pankova, 2015; Norris, 2006; Odendahl, 2019).

3. Computer-based C-tests

With the growing popularity of digital technologies, language test developers are more and more frequently using computers and other electronic devices to design tests, create test items, assemble them into test forms, administer these to the examinees, score the responses, deliver reports, provide feedback, as well as analyse the scores. Tests of reduced redundancy are, quite naturally, no exception to this general tendency. Recent examples of computerized C-tests include the *Online Language Placement Test*, also known as *onSET* (www.onset.de), which is a test of language proficiency in German or English; *QSAT*, a web-based mC-test constructed by deleting the first half (instead of the second half) of every second word (Boonsathorn & Kaoropthai, 2016); and a computerized C-test in Spanish (Riggs & Maimone, 2018). The crucial question that arises in the context of computerized testing is whether the mode of delivery has a significant impact on the properties of the measurement instrument. This problem has been addressed by Bisping and Raatz (2002), who tested the equivalence of paper-based and computer-based versions of a C-test and found that there were no significant differences between the two modes of administration in terms of the means, standard deviations and reliabilities. The computerized C-test, however, was more appealing to the test takers.

Computerized C-testing can be largely automated with the aid of dedicated software. One of the earliest tools of this kind, reported by Koller and Zahn (1996), was developed at the Language Centre of the University of Erlangen-Nürnberg to construct, deliver, and score computerized C-tests for placement purposes. This DOS program was also capable of conducting statistical evaluations of the scores. More recently, Odendahl (2019) developed a web-based C-testing platform, which uses a corpus of texts taken from blogs, novels, and newspaper articles as a basis for C-test generation. The system can be used to administer C-tests and then rank the test takers by general language proficiency, thus providing the necessary information for admission and placement decisions. In addition, some mobile apps have been developed for language learners who are interested in practising C-

tests. These include an app that can generate C-tests for individuals preparing for various proficiency or entrance exams (Khokhryakov, 2019) and an app designed to calculate a predicted score for the IELTS (International English Language Testing System) exam (Heidermann & Kebble, 2017; Kebble, 2018). The latter app uses texts from previous IELTS papers to automate C-test generation.

4. WebClass

WebClass (webclass.co) is an online learning management system (LMS) incorporating a fairly extensive testing component that can be utilized for web-based language test development (Malec, 2018). In particular, the system enables the test developer to author, deliver, and conduct statistical analysis of assessments of various types. While the platform is not specifically designed for reduced redundancy testing, it offers a tool that can be helpful in constructing cloze tests and C-tests, either of which, from a technical point of view, is composed of standard gap-filling items arranged into a passage of text.

The tool in question is the text-to-items converter, illustrated in Figure 1 (see also Malec, 2016). The converter is a kind of add-on to the test editor in that it can be used to format a piece of text to be transformed into test items proper. In the case of gap-filling items, the rule is that every word (or a longer string of text) inside square brackets is replaced by a blank and the text inside the brackets is the key. Each gap can contain an unlimited number of correct alternatives, separated by a forward slash.
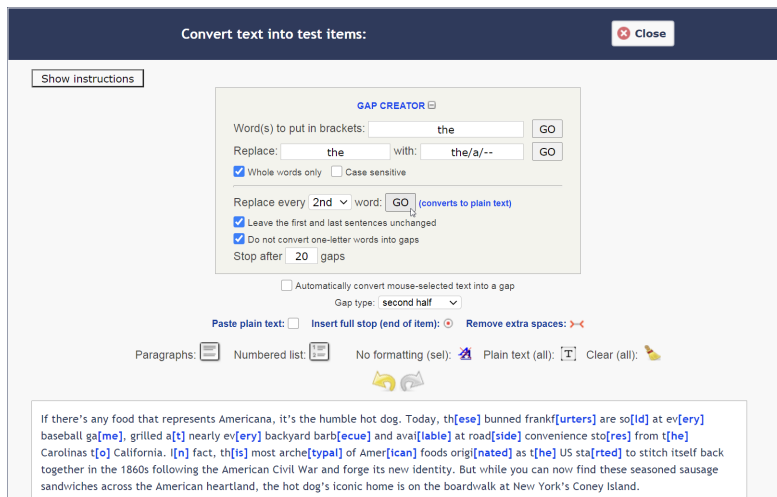

Figure 1. Using the text converter to create a C-test on WebClass
(Text source: www.bbc.com/travel/story/20190702-the-truth-about-the-us-most-iconic-food)

As can be seen in Figure 1, the text converter features a gap creator for automated replacement of every *n*th word with a gap. Three important options relevant to cloze-type formats are integrated into the gap creator, namely leaving the first and last sentences unchanged, skipping one-letter words, and terminating the replacement after a specified number of gaps. The type of gap is selected below the gap creator box[1], and the options available include *selection*, *first half*, *second half*, and *keep first letter*. The first of these options is intended for cloze tests: each whole word (or some other piece of manually selected text) is replaced with a gap. The third option (second half) is meant for the standard variety of C-tests. It is worth adding that numerals (optionally followed by a single letter, as in *1980s*) are always skipped by the gap creator and any other words that need to be left intact (such as proper names) can be marked with an asterisk in front. The text-to-items converter can handle one passage of text at a time. Hence, if more passages are to be included in the entire test, the gap-creating procedure has to be repeated for each of them.

Further settings are available in the main test editor. These include, for example, the time limit, mastery level, acceptable and partially acceptable responses, feedback (general and answer-specific), as well as gap length. Moreover, the test constructor may decide to automatically award partial credit for responses containing no more than a specific number of spelling errors (modifiable for each item). In addition, the scoring algorithm can be set to ignore spaces, punctuation, and capitalization in the responses to any given item. Generally speaking, either exact-answer or acceptable-answer scoring can be implemented on the platform.
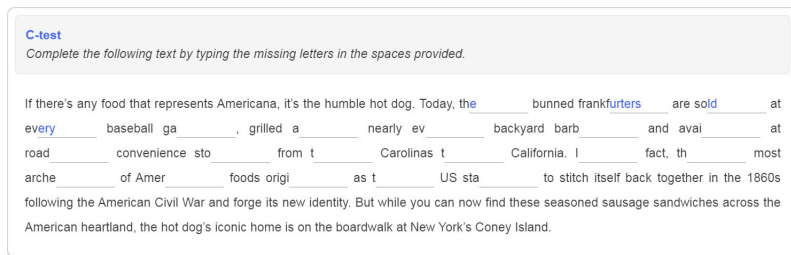

Figure 2: Responding to a C-test on WebClass

Figure 2 shows the passage from Figure 1 in its final form (as displayed to the examinees). The test as a whole can optionally be delivered in full screen mode. In addition, the test takers may be instructed not to navigate away from the test, and if they actually do so, the test is automatically suspended for them. When this happens, the teacher or test administrator is informed and has the option of allow-

---

[1]   The gap-type options are located below the gap creator box (which is collapsed by default) because they are also applicable to two other methods of creating gaps, i.e. by double-clicking a word and (optionally) by selecting a piece of text with the mouse.

ing individual test takers to continue. This security setting may be a vital option in unproctored testing environments, as in distance education. It may be worth adding that the time limit can be modified while the test is being administered.

Test administration is routinely followed on WebClass by human verification of the automated scoring. In the case of gap-filling items, the score verification tool displays the content of each item followed by the key and all the responses not included in the key. The test takers are grouped by the incorrect responses. The score verification procedure consists in inspecting the incorrect answers and, if need be, adding them to the key or to a pool of acceptable answers (deserving either full or partial credit). Naturally, once an answer has been accepted, the change is applicable to all the test takers who have submitted this particular response, and it is not necessary to repeat the procedure for each individual test taker. The score verification tool additionally allows the tester to provide feedback to the examinees on their responses.

The final components of test development on WebClass include statistical analysis of the scores and item banking. The statistics computed by the system include several reliability estimates and item indices. Some of these statistics are specifically applicable to criterion-referenced assessments, while tests of reduced redundancy are intended for norm-referenced interpretations. The statistics which are relevant to norm-referenced tests include Cronbach's alpha, Spearman-Brown split-half reliability coefficient, SEM (standard error of measurement), item facility, and item discrimination. It is currently not possible to compute reliability estimates with C-test passages as superitems. However, all the scores can be downloaded (in .csv format) to be analyzed elsewhere. Finally, items which function properly can be stored in the item bank, and then imported into a new test. Every cloze and C-test passage is stored in the item bank and retrieved in its entirety.

Conclusion

Redundancy as a linguistic concept has existed in language testing since the early 1960s (Barnwell, 1996). For many years the cloze test was considered to be a promising testing technique. However, over the course of several decades many of its shortcomings became increasingly evident until they were finally addressed by the C-test in the 1980s. This paper has discussed cloze tests and C-tests as measures of general language proficiency and briefly presented the WebClass LMS as a tool that can aid the tester in developing tests of this kind.

References

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

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

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System, 29*(2), 209–219.

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

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

Barnwell, D. P. (1996). *A History of Foreign Language Testing in the United States: From Its Beginnings to the Present*. Tempe, AZ: Bilingual Press.

Beard, J. G. (1967). Comprehensibility of high school textbooks: Association with content area. *Journal of Reading, 11*(3), 229–234.

Bisping, M., & Raatz, U. (2002). Sind computerisierte und Papier&Bleistift-Versionen des C-Tests äquivalent? In R. Grotjahn (Ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Vol. 4, pp. 131–155). Bochum: AKS-Verlag.

Boonsathorn, S., & Kaoropthai, Ch. (2016). QSAT: The web-based mC-test as an alternative English proficiency test. *TESOL International Journal, 11*(2), 91–107.

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement, 5*(3), 189–196.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension. *Reading Research Quarterly, 4*(3), 358–365.

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing, 7*(1), 13–30.

Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*(1), 19–31.

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

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

Byrne, M. A., Feldhusen, J. F., & Kane, R. B. (1971). The relationships among two cloze measurement procedures and divergent thinking abilities. *Reading Research Quarterly, 6*(3), 378–393.

Cash, M. M., & Schumm, J. S. (2006). Making sense of knowledge: Comprehending expository text. In J. S. Schumm (Ed.), *Reading Assessment and Instruction for All Learners* (pp. 262–296). New York: The Guilford Press.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: What difference does it make? *Language Testing, 7*(2), 121–146.

Claiborne, J. D. (1990). *Mathematical Preliminaries for Computer Networking*. New York: John Wiley & Sons, Inc.

Coleman, J. A. (1996). A comparative survey of the proficiency and progress of language learners in British universities. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 367–399). Bochum: Brockmeyer.

Culhane, J. W. (1970). CLOZE procedures and comprehension. *The Reading Teacher, 23*(5), 410–464.

Delbridge, A., & Bernard, J. R. L. (1966). *Patterns in Language*. Sydney: Angus and Robertson.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing, 9*(2), 187–206.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290–325.

Fano, R. M. (1961). Introductory remarks to the third session of the Symposium on the Structure of Language and its Mathematical Aspects. In R. Jakobson (Ed.), *Structure of Language and its Mathematical Aspects: Proceedings of the Twelfth Symposium in Applied Mathematics* (pp. 261–263). Providence, RI: American Mathematical Society.

Grotjahn, R. (2013). Cloze test. In Michael Byram & Adelheid Hu (Eds.), *Routledge Encyclopedia of Language Teaching and Learning* (2nd ed.), pp. 121–122. Abingdon: Routledge.

Hale, G. A., Stansfield, Ch. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6*(1), 47–76. DOI:10.1177/026553228900600106.

Heidermann, D., & Kebble, P. G. (2017). IELTS Score Predictor [Apple app]. Retrieved June 25, 2021, from https://apps.apple.com/us/app/ielts-score-predictor/id1190170836.

Horning, A. S. (1987). *Teaching Writing as a Second Language*. Carbondale: Southern Illinois University Press.

Huhta, A. (1996). Validating an EFL C-test for students of English philology. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 197–234). Bochum: Brockmeyer.

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

Jafarpur, A. (1999a). Can the C-test be improved with classical item analysis? *System, 27*(1), 79–89.

Jafarpur, A. (1999b). What's magical about the rule-of two for constructing C-tests? *RELC Journal, 30*(2), 86–100.

Kebble, P. G. (2018). Designing and assessing a digital, discipline-specific literacy assessment tool. *EuroCALL Review, 26*(1), 79–88.

Khodadady, E. (2013). Authenticity and sampling in C-tests: A schema-based and statistical response to Grotjahn's critique. *The International Journal of Language Learning and Applied Linguistics World, 2*(1), 1–17.

Khokhryakov, I. (2019). C-Test Generator [Android app]. Retrieved June 25, 2021, from https://play.google.com/store/apps/details?id=com.appdroid.develop.c_testgenerator.

Klein-Braley, Ch. (1985). A cloze-up on the C-test: A study in the construct validation of authentic tests. *Language Testing, 2*(1), 76–104.

Klein-Braley, Ch. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*(1), 47–84.

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

Koller, G., & Zahn, R. (1996). Computer based construction and evaluation of C-tests. In R. Grotjahn (Ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 401–418). Bochum: Brockmeyer.

Lado, R. (1986). Analysis of native speaker performance on a cloze test. *Language Testing, 3*(2), 130–146.

Lehmann, Ch. (2005). Pleonasm and hypercharacterisation. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 2005* (pp. 119–154). Dordrecht: Springer.

Linnemann, M., & Wilbert, J. (2010). The C-test: A valid instrument for screening language skills and reading comprehension of children with learning problems? In R. Grotjahn (Ed.), *The C-Test: Contributions from Current Research* (pp. 113–124). Frankfurt am Main: Peter Lang Verlag.

Malec, W. (2016). Automating the construction of selected-response items with a text-to-items converter. *CBU International Conference Proceedings, 4*, 864–872.

Malec, W. (2018). *Developing Web-Based Language Tests*. Lublin: Wydawnictwo KUL.

McKay, T. H., Teimouri, Y., Sağdıç, A., Salen, B., Reagan, D., & Malone, M. E. (2021). The cagey C-test construct: Some evidence from a meta-analysis of correlation coefficients. *System, 99*, 102526.

Mozgalina, A., & Ryshina-Pankova, M. (2015). Meeting the challenges of curriculum construction and change: Revision and validity evaluation of a placement test. *The Modern Language Journal, 99*(2), 346–370.

Norris, J. M. (2006). Development and evaluation of a curriculum-based German C-test for placement purposes. In R. Grotjahn (Ed.), *The C-Test: Theory, Empirical Research, Applications* (pp. 45–83). Frankfurt am Main: Peter Lang Verlag.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Odendahl, W. (2019). Gateways, placements, and grouping: Automating the C-Test for language proficiency ranking. *Interface – Journal of European Languages and Literatures, 8*, 29–67.

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), 151–158.

Oller, J. W. (1979). *Language Tests at School*. London: Longman.

Oller, J. W., & Conrad, Ch. A. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*(2), 183–195.

Oller, J. W., & Jonz, J. (1994). Why cloze procedure? In J. W. Oller, & J. Jonz (Eds.), *Cloze and Coherence* (pp. 1–20). London, Toronto: Associated University Presses.

Porter, D. (1976). Modified cloze procedure: A more valid reading comprehension test. *ELT Journal, 30*(2), 151–155.

Raatz, U. (1985). Better theory for better tests? *Language Testing, 2*(1), 60–75.

Rankin, E. F., & Culhane, J. W. (1969). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading, 13*(3), 193–198.

Renkema, J. (2004). *Introduction to Discourse Studies*. Amsterdam: John Benjamins.

Riggs, D., & Maimone, L. L. (2018). A computer-administered C-test in Spanish. In J. M. Norris (Ed.), *Developing C-Tests for Estimating Proficiency in Foreign Language Research* (pp. 265–294). Frankfurt am Main: Peter Lang Verlag.

Schneyer, J. W. (1965). Use of the cloze procedure for improving reading comprehension. *The Reading Teacher, 19*(3), 174–179.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656.

Spolsky, B. (1968). *What does it mean to know a language, or how do you get someone to perform his competence?* Paper presented at the Second Conference on Problems in Foreign Language Testing, University of Southern California, 7–9 November. Washington, DC: ERIC.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *The Modern Language Journal, 58*(5/6), 239–241.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*(4), 414–438.