Prototyping a Concordance-based Cloze Test: Preliminary Results

1. Kunlaphak Kongsuwannakul, Ph.D. student, Validation, School of Education, University of Leicester, kk234@le.ac.uk
2. Glenn Fulcher, Professor, Education and Assessment, School of Education, University of Leicester, gf39@le.ac.uk
3. Nick Smith, Senior Lecturer, Applied Linguistics, School of Education, University of Leicester, ns359@le.ac.uk

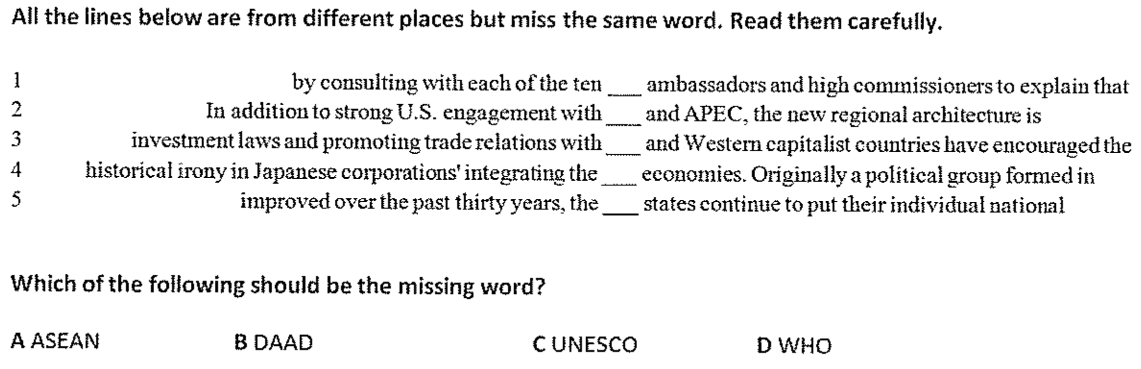
Abstract

Language testing is a dynamic discipline seeking to, for one thing, innovatively measure linguistic competence with accuracy. Intending to satisfy this requirement of an innovative measure and investigate construct validity, this paper presents preliminary results of quantitatively prototyping a concordance-based cloze test (ConCloze). The initial test and item specification is first discussed. Then population and sampling are defined. Test and item responses are analyzed with descriptive statistics. The results are that (a) all of the prototyped ConCloze items have high reliability (α=0.84), (b) their alphas-if-items-deleted are consistently high, and (c) the majority of examinees commented that the test was too difficult. With a collective analysis of multiple pieces of data, an inference is that ConCloze could be a measure of an underlying discrete construct, most likely testing knowledge of word associates in context. However, considering the examinees' comments and high dropout rate, it can be inferred that the difficulty level may be inappropriate for the intended population as it may raise response-validity issues. Some implications of ConCloze towards ASEAN English pedagogy and assessment are also summarized.

Keywords: Concordance-based cloze, validation, prototype, internal consistency

Background

The concordance-based cloze test (henceforth ConCloze) has appeared in the literature for over two decades (cf. Butler, 1990; Butler, 1991). Nonetheless, it is only recently that it has received revived attention for language-testing purposes (Kongsuwannakul, 2014; Kongsuwannakul, in print, a; Kongsuwannakul, in print, b). Figure 1 below illustrates an ASEAN-themed sample item based on the test and item specification (spec) exemplified in Kongsuwannakul (in print, a). Because neither of these publications presented empirical findings which could suggest the construct validity of this item type—a central issue for language testing and measurement (Kane, 2012a; Kane, 2012b; McNamara, 2006; Messick, 1988; Messick, 1989; Messick, 1994; Messick, 1995)—this paper hence deals with preliminary results regarding the quantitative prototyping of this item type, with particular attention to the structural aspect of its construct validity.



All the lines below are from different places but miss the same word. Read them carefully.

1 by consulting with each of the ten ___ ambassadors and high commissioners to explain that
2 In addition to strong U.S. engagement with ___ and APEC, the new regional architecture is
3 investment laws and promoting trade relations with ___ and Western capitalist countries have encouraged the
4 historical irony in Japanese corporations' integrating the ___ economies. Originally a political group formed in
5 improved over the past thirty years, the ___ states continue to put their individual national

Which of the following should be the missing word?

A ASEAN B DAAD C UNESCO D WHO

Figure 1: Example of a ConCloze Item

Given that English is the official lingua franca used in ASEAN, the significance of this study lies in the fact that it provides the first empirical evidence regarding what this item type actually measures. This in turn could inform language practitioners—from teachers to language testers, and from course designers to educators—of (a) how to improve the existing prototype further in order to fit their corresponding contexts, and (b) how to apply the item format to other settings altogether. For example, teachers of English, learning from this study the construct validity of ConCloze, could develop their own ConCloze test for their students from course textbooks. Another example is when ConCloze becomes part of a university-entrance English examination, and the writers of secondary-education curricula may wish to encompass into their courses the areas of

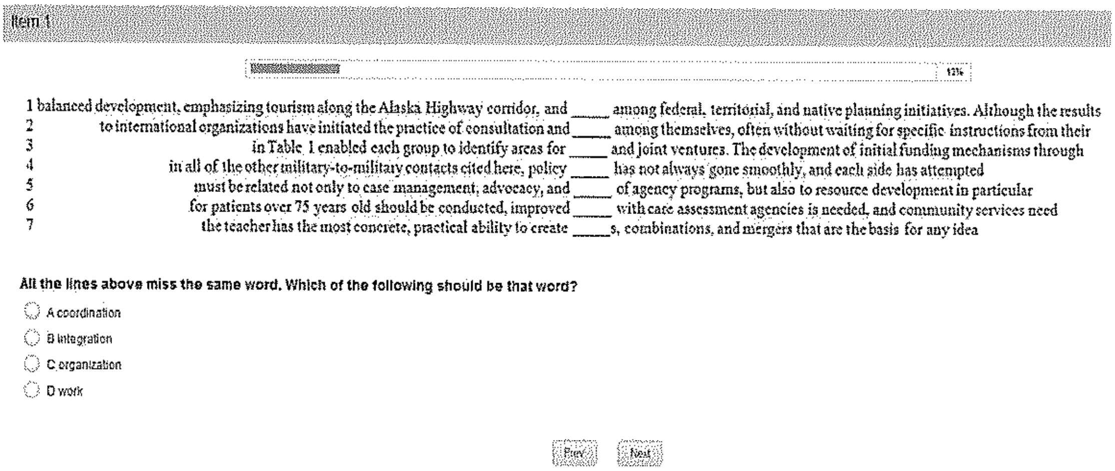
English linguistic competence that can address the domains of knowledge required and to be tested by ConCloze. Therefore, in short, this study is potentially useful for the ASEAN English language practices both on a large scale and for the individual level.

Research Questions

Given the line of inquiry into ConCloze construct validity, there is one primary research question in this prototyping study: Do item responses (IRs) have internal consistency? In order to address this structural-validity question in empirical terms, there are thus two operational research questions formulated: a) Is the ConCloze test produced reliable?; and b) Do the IRs of different items correlate with one another? Other peripheral issues, which would assist in realizing wider implications, will also be addressed along the discussion.

Research Methodology

Test design. Figure 2 below is the first ConCloze item, functioning as a sample test item for the guiding language in Table 1 (cf. Fulcher & Davidson, 2007). From this spec, it is evident that this 39-item test—as a research instrument—is designed with a view to curbing construct-irrelevant variance such as testwiseness to the minimum (cf. Cohen, 2012). This is in order that the test interpretation can represent the ConCloze construct relatively accurately.



1 balanced development, emphasizing tourism along the Alaska Highway corridor, and _____ among federal, territorial, and native planning initiatives. Although the results
2 to international organizations have initiated the practice of consultation and _____ among themselves, often without waiting for specific instructions from their
3 in Table 1 enabled each group to identify areas for _____ and joint ventures. The development of initial funding mechanisms through
4 in all of the other military-to-military contacts cited here, policy _____ has not always gone smoothly, and each side has attempted
5 must be related not only to case management, advocacy, and _____ of agency programs, but also to resource development in particular
6 for patients over 75 years old should be conducted, improved _____ with care assessment agencies is needed, and community services need
7 the teacher has the most concrete, practical ability to create _____s, combinations, and mergers that are the basis for any idea

All the lines above miss the same word. Which of the following should be that word?

○ A coordination

○ B integration

○ C organization

○ D work

Figure 2: ConCloze Item 1 as sample item for the guiding language

Entry	Guiding Language
1	There is one example item (Item 0) explained at the beginning of the test.
2	Each item has four options, only one of which is the correct answer (key). Test takers are assumed to be familiar with this task type (selecting one correct answer).
3	Selecting the correct answer of each item is scored 1. Selecting any of the other options is scored 0. Not selecting any option is scored 0. Selecting more than one option in each item is scored 0.
4	The prompt is made up of seven concordance lines, each marked with its line number at the front. This emphasizes the fact that they are from different origins.
5	The concordance lines are centered and truncated.
6	Each concordance line contains ten words on both sides of the Key-Word-In-Context (KWIC).
7	There is no modification to the words in the concordance lines. The only exception is when a giveaway of the correct answer would pose a construct-irrelevant threat.
8	Concordance lines are sampled from the Corpus of Contemporary American English (COCA)'s (Davies, 2008–) Academic Genre.
9	All the concordance lines presented to the test takers are sorted right to the KWIC blanks in ascending order.
10	All the KWIC blanks are fixed at an equal length. In Figure 2, this is five underscores long.
11	The stem wording of each item is constant. It states the problem, "All the lines above miss the same word," and urges action, "Which of the following should be that word?"
12	Since "vocabulary with a middle frequency" is the most problematic one to the learners of any disciplines and is in fact academic vocabulary (reviewed in Thurstun & Candlin, 1998, p. 268), the key is sampled purposefully from Gardner & Davies's (2014) Academic Vocabulary List (AVL) 1K–1.3K.
13	The distractors are drawn from a close semantic field. They can have either a collocational, analytical, or paradigmatic relationship to the key (the correct KWIC). In Figure 2, integration has a paradigmatic relationship with coordination, and organization and work an analytic relationship.
14	The forms of the distractors must be changed so as to be identical to that of the key. In Figure 2, all the distractors (B–D) are nouns or equivalent.
15	All the options are checked against the given concordance lines and, when deemed necessary, added alternative suffixes to. This must be done such that no construct-irrelevant testwiseness of suffixes (either derivational or inflectional morphemes) can

	give away the correct answer.
16	All the options are arranged in ascending alphabetical order.
17	The three purposefully selected word classes are noun, verb, and adjective. Each has 13 items, arranged in order: noun, verb, adjective, and so forth.

Table 1: Guiding language

Population defined. Because ConCloze being prototyped is based on Gardner & Davies's (2014) Academic Vocabulary List (AVL) words (see Table 1 above), the intended population can be defined as non-native speakers of English engaging or having engaged with academic English. Therefore, an operational definition for this population encompasses those non-native speakers of English studying in or graduating from the university level, who are very likely to have had exposure to academic English.

Sampling. In a prototyping process, a small sample—as little as ten for item respondents—is usually adequate for trying out a particular prototype (Nissan & Schedl, 2012). Accordingly, with convenience and snowball samplings, this study has sampled 38 online examinees, who were all studying in a university program or graduating from the higher-education level (undergraduate and postgraduate).

Data collection. SurveyMonkey was used as the test platform (open for one month, March 2014, at https://www.surveymonkey.com/s/AcadEnglishVocabTest). Since this research follows BERA's (2011) guidelines of research ethics and so allowed the participants to opt out anytime if they wished, only 13 of the 38 examinees above completed the entire test and filled out the feedback form at the end of the test.

Sample description. Table 2 below shows the gender distribution of the ConCloze participants, which represented a moderately good distribution of genders in this sample. Then Table 3 displays their educational levels, which seemed to skew moderately towards the postgraduate classification. Their ages ranged from 19 to 53 (range=34), with the median of 33. Those speaking Thai as their first language were 22 in number (64.71%), those with Arabic seven (20.59%), and those with Kurdish three (7.89%). The majority of the participants (n=20, 58.82%) had had 0–3 months of experience in English-speaking countries. Seventeen of the participants had taken an IELTS test during the past three years. Among them, 13 also reported their overall scores, the mean of which was 5.54 on Bands 0–9. Twelve of the other participants had not taken any standardized test of English during the past seven years.

Gender	Count	Percent
Male	14	38.9
Female	21	58.3
Prefer not to answer	1	2.8
Total response	36	100

Table 2: Gender of ConCloze 1 participants

Highest Education Level*	Count	Percent
Presessional course to an undergraduate level	4	11.43
Presessional course to a postgraduate level	12	34.29
Year 1, undergraduate	1	2.86
Year 3, undergraduate	2	5.71
Year 1, taught postgraduate	2	5.71
Year 1, research postgraduate	3	8.57
Year 4, research postgraduate	2	5.71
Year 5, research postgraduate	1	2.86
Holds a bachelor's degree	1	2.86
Holds a master's or a Ph.D., or studies at an year-unspecified postgraduate level	7	20.00
Total response	35	100

* Zero-response and N/A categories (the response='none') excluded from presentation

Table 3: Educational level of ConCloze 1 participants

Data analysis. The statistics used are descriptive statistics: average, percentage, Cronbach's alpha reliability index, and alpha-if-item-deleted for the test and IRs. A simple textual analysis is used for written feedback from the respondents.

Findings

Internal consistency. From a Cronbach's alpha coefficient test of the 13 test-completers' scores, the overall reliability is 0.8405, which indicates high test reliability. It can therefore be interpreted that ConCloze 1 items received quite consistent IRs throughout the test.

With regard to alphas-if-items-deleted, Table 4 below presents ConCloze alpha coefficients if each particular item is deleted. For example, deleting Item 27 can increase the scale reliability to 0.8534 at best. The table shows that the alpha variation can range from 0.82–0.85. Given that the actual test alpha is 0.8405, it can therefore be interpreted that deleting any particular item does not make much change to the overall test reliability. In other words, each ConCloze item seems to contribute to the scale variance fairly equally.

Item	Alpha if item deleted	Item	Alpha if item deleted	Item	Alpha if item deleted
1	0.8337	14	0.8265	27	0.8534*
2	0.8433	15	0.8382	28	0.8295
3	0.8476	16	0.8482	29	0.8235
4	0.8337	17	0.8405	30	0.8348
5	0.8227	18	0.8497	31	0.8449
6	0.8418	19	0.8302	32	0.8302
7	0.8235†	20	0.8405	33	0.8348
8	0.8395	21	0.8358	34	0.8276
9	0.8414	22	0.8479	35	0.8295
10	0.8342	23	0.8385	36	0.8348
11	0.8316	24	0.8418	37	0.8250
12	0.8295	25	0.8295	38	0.8331
13	0.8540	26	0.8446	39	0.8388

* Highest

† Lowest

Table 4: ConCloze alphas-if-items-deleted

Overall difficulty. In the section of research methodology, a low test-completion rate—i.e., at 34%, with the dropout rate at 66%—has been presented. While analyzing item facility is not the focus of this study, an inappropriate level of test difficulty may pose response-validity threat (cf. Henning 1987). Accordingly, participation rates are analyzed, showing in Figure 3 below that (a) a sharp decline in participants occurred when the sample ConCloze question was presented to the examinees, and (b) by the end of Item 8, the number of respondents became constant. Therefore, it can be interpreted that once the respondents explored the sample question, many of them did not wish to continue the testing, and a few more participants exited the test webpage—closing or leaving idle the browser tab—after trying answering the initial test items.

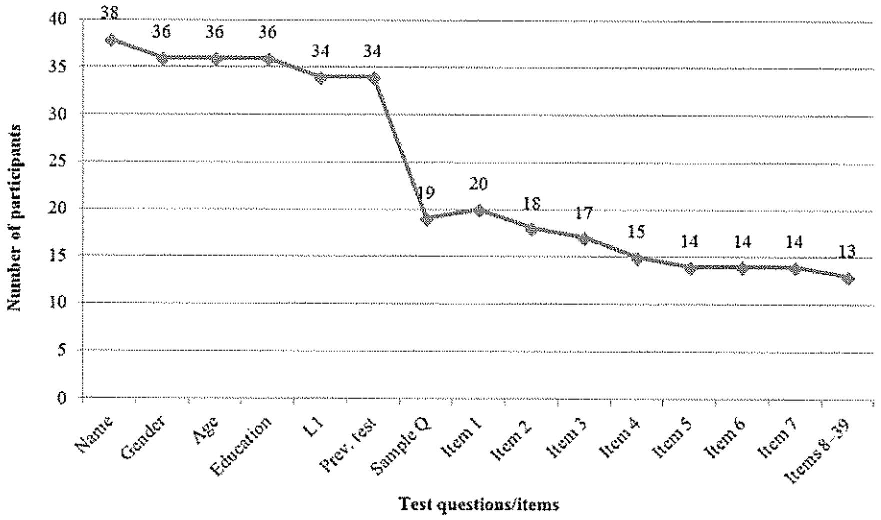
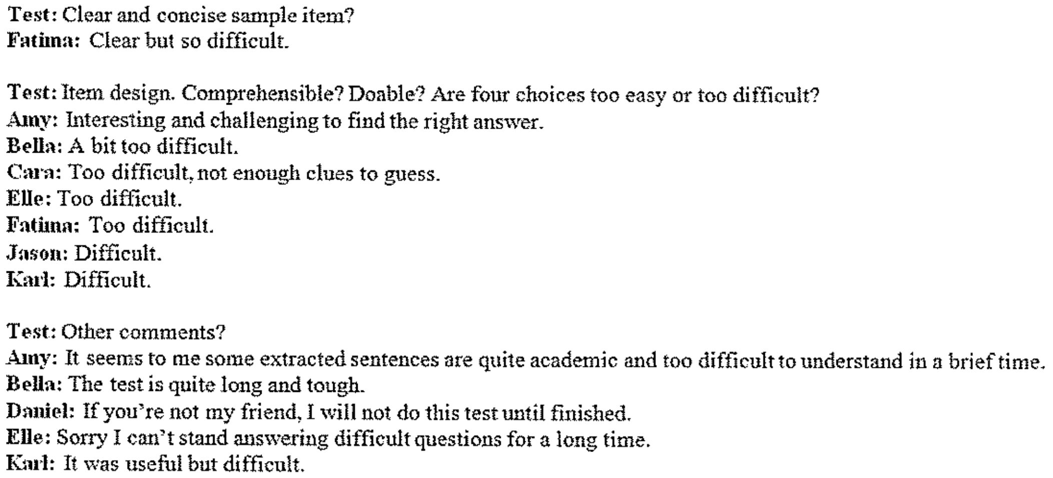


Figure 3: Diminishing participant number of ConCloze

Textual feedback. Because of the anonymous and online nature of the test, it is unknown what exactly their reasons were for the high dropout above. Still, a textual analysis of the comments from the test-completers presented in Figure 4 below seems to help pinpoint that one of the main reasons could be great test difficulty perceived by the respondents. That is, the figure has one theme clearly recurring throughout for interpretation: ConCloze was difficult for the sample.



Test: Clear and concise sample item?
Fatima: Clear but so difficult.

Test: Item design. Comprehensible? Doable? Are four choices too easy or too difficult?
Amy: Interesting and challenging to find the right answer.
Bella: A bit too difficult.
Cara: Too difficult, not enough clues to guess.
Elle: Too difficult.
Fatima: Too difficult.
Jason: Difficult.
Karl: Difficult.

Test: Other comments?
Amy: It seems to me some extracted sentences are quite academic and too difficult to understand in a brief time.
Bella: The test is quite long and tough.
Daniel: If you're not my friend, I will not do this test until finished.
Elle: Sorry I can't stand answering difficult questions for a long time.
Karl: It was useful but difficult.

Figure 4: ConCloze excerpted comments (pseudonymized)

Discussions

Internal consistency. Based on the findings about the ConCloze internal consistency, it can be inferred that the items likely measure the same domain of linguistic competence. Given the theoretical considerations outlined in Kongsuwannakul (in print, b), a tentative inference is therefore that the construct involves the knowledge of word associates in context.

Test difficulty and textual feedback. While the opinions and the decline in the participant number as discussed in the findings may be largely subjective in nature and therefore may vary from sample to sample, it can be inferred that the test may be perceived greatly difficult for, at the very least, part of the intended population. In other words, an inference is that the test and item spec as-is may not be suitable for the intended population because it presumably produces overly difficult items, a source of construct-irrelevant variance.

Implications. As introduced earlier, given that the English language is (a) set to be the official language shared by ASEAN countries, and (b) certainly, an unofficial global language (Crystal, 1997, 2003; Graddol, 1997), the potential implications of the ConCloze item type are huge. As Kongsuwannakul (Kongsuwannakul, in print, b) pointed out, two significant impacts that could be directly applicable to ASEAN English-language classroom practices are word-knowledge profiling and evaluation, and reconceptualized English-language pedagogy.

Concerning word-knowledge profiling and evaluation, it implies that, rather than producing single scores for vocabulary breadth, teachers and assessors may need to consider measuring the depth of vocabulary knowledge of their students. This means that the students could have a developmental profile of their vocabulary-depth progress along their formal education, for example. Not only is this more informative for the test-score users (e.g., at entrance examinations and job applications), but it can also provide diagnostic feedback to the students themselves. This can enhance lifelong learning of English for the ASEAN community.

With regard to the applications to language pedagogy, it is suggested that learners and teachers alike should take into consideration the complex nature of formulaic language—the reality of language use that is not fully recognized in teaching, learning and testing (cf. Read, 1993; Read, 1998; Read, 2000; Read, 2012; Read & Chapelle, 2001; Read & Nation, 2004; Schmitt, 2010). This means that their focus should be shifted from individual, discrete words to interconnected and context-embedded nature of language use. If realized, this will not only benefit the results of their English-assessment performance, but it will also enable them to use language naturally and meaningfully in non-test situations.

Recommendations

Based on the sample description in this study, it can be inferred that ConCloze has reached an adequate number of prospective respondents for prototyping purposes, but failed to find respondents of satisfactorily varied education levels and L1 backgrounds. The demographic variation in terms of educational background, and first language was insufficient. While this does not have a serious impact on the test-score interpretation in this prototyping, it implies that subsequent research phases should reach participants of much more various backgrounds to the extent that a better distribution of education-level and L1 diversity are achieved. That is, even though an equal distribution in education levels is not theoretically driven for answering the research questions, L1 diversity is and will potentially help to test a hypothesis of cross-L1 differences (or a lack thereof) in ConCloze performance, thereby obtaining a potential rebuttal against language-group bias. Therefore, it is recommended that other ConCloze prototyping studies may seek respondents of several L1 backgrounds and, if research resources allow, should carry out investigations comparing and contrasting their test-task performance.

With regard to my ongoing thesis, the focus is on the structural and substantive aspects of the ConCloze construct validity, using mixed methods. It is therefore recommended that confirmatory structural equation modelling of the IRs be carried out on a very large scale on top of this study.

References

BERA. (2011). *Ethical guidelines for educational research.* London: British Educational Research Association.

Butler, J. (1990). Concordancing, teaching and error analysis: Some applications and a case study. *System, 18*(3), 343–9.

Butler, J. (1991). Cloze procedures and concordances: The advantages of discourse level authenticity in testing expectancy grammar. *System, 19*(1–2), 29–38.

Cohen, A. D. (2012). Test-taking strategies and task design. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 262–77). Oxon: Routledge.

Crystal, D. (1997, 2003). *English as a global language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Davies, M. (2008–). The corpus of contemporary American English: 450 million words, 1990–present. Retrieved from http://corpus.byu.edu/coca/

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment.* Oxon, UK: Routledge.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–27. doi:10.1093/applin/amt015

Graddol, D. (1997). *The future of English? A guide to forecasting the popularity of the english language in the 21st century* (2000th ed.). London: The British Council.

Henning, G. H. (1987). *A guide to language testing: Development, evaluation, research.* USA: Newbury House.

Kane, M. T. (2012a). All validity is construct validity. or is it? *Measurement: Interdisciplinary Research and Perspectives, 10*(1–2), 66–70.

Kane, M. T. (2012b). Validating score interpretations and uses: Messick lecture, language testing research colloquium, Cambridge, April 2010. *Language Testing, 29*(1), 3–17. doi:10.1177/0265532211417210

Kongsuwannakul, K. (2014). Language processes of the concordance-based cloze item type: Bridging a theoretical gap between language testing and second language acquisition. *The International Journal of Communication and Linguistic Studies,* Advance access.

Kongsuwannakul, K. (in print, a). Six techniques for creating variety in the concordance-based cloze item type. *International Journal of Assessment and Evaluation,*

Kongsuwannakul, K. (in print, b). Theoretical considerations of applications and implications of concordance-based cloze tests. *Literary and Linguistic Computing, advanced access* doi:10.1093/llc/fqu033

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly: An International Journal, 3*(1), 31–51.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 33–45). New Jersey: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education; Collier Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances a Scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–9.

Nissan, S., & Schedl, M. (2012). Prototyping new item types. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 281–94). Oxon: Routledge.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355–71. doi:10.1177/026553229301000308

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. *Validation in language assessment: Selected papers from the 17th language testing research colloquium, long beach* (pp. 41-60). New Jersey: Lawrence Erlbaum Associates.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Read, J. (2012). Piloting vocabulary tests. In G. Fulcher, & F. Davidson (Eds.), *The routledge handbook of language testing* (pp. 307–20). Oxon: Routledge.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1–32.

Read, J., & Nation, I. S. P. (2004). Measurement of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 23–35). The Netherlands: John Benjamins.

Schmitt, N. (2010). *Researching vocabulary a vocabulary research manual.* Basingstoke: Palgrave Macmillan.

Thurstun, J., & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic english. *English for Specific Purposes, 17*(3), 267–80.