

Source-Text-Conditioned Cloze Generation for Discourse Comprehension Assessment in Intelligent Textbooks

Langdon Holmes

Vanderbilt University
langdon.holmes@vanderbilt.edu

Scott Crossley

Vanderbilt University
scott.crossley@vanderbilt.edu

Wesley Morris

Vanderbilt University
wesley.g.morris@vanderbilt.edu

Aiden Min

Vanderbilt University
seungchan.min@vanderbilt.edu

Abstract

Cloze exercises offer scalable comprehension assessment, but their validity depends on which words are selected as gaps. We compared three automated methods for cloze exercises generated from summaries within an intelligent textbook platform. Conditioning masked language model predictions on the source text (contextuality-plus) produced higher quality and more source-dependent gaps.

1 Introduction

Comprehension monitoring during reading is well-established as a goal of formative assessment (McNamara and Magliano, 2009), and intelligent textbook platforms increasingly pursue this goal through embedded activities (Crossley et al., 2024). While constructed-response tasks such as summary writing and short answer questions provide rich signals of understanding, they require sophisticated automated scoring models and impose substantial cognitive and time demands on learners. Cloze exercises—reading tasks in which strategically selected words are deleted for learners to restore—offer a lightweight alternative to constructed-response tasks that can be deployed frequently with minimal learner burden, providing immediate feedback on comprehension without the overhead of free-text evaluation (Oller, 1972).

The validity of cloze exercises as comprehension assessments depends critically on gap selection. Taylor’s (1953) original fixed-ratio deletion (every n th word) was intended to produce gaps that measure readability. For the purposes of reading comprehension assessment, this procedure produces items of highly variable quality, as many deleted words are recoverable from local syntactic or collocational patterns alone. More recent work on “deep cloze” demonstrates that gaps requiring integration of information across sentences or from external sources yield more valid comprehension

assessments than those fillable from sentence-level context (Jensen and Elbro, 2022; Gellert and Elbro, 2013). This distinction aligns with construction-integration theory (Kintsch, 1988), which holds that skilled comprehension involves constructing a situation model that integrates textual propositions with broader discourse context rather than merely processing individual sentences.

We address a specific variant of the cloze task that is concerned with generating gap-fill exercises from summaries of source texts within iTELL (intelligent Texts for Enhanced Lifelong Learning), a web-based platform that transforms educational texts into interactive learning environments with automated formative feedback (Crossley et al., 2024; Morris et al., 2024b). In iTELL, cloze exercises assess whether a student can demonstrate comprehension of a source text by restoring words in a summary, which requires cross-document discourse integration rather than surface-level linguistic knowledge.

We evaluate three automated gap selection methods for increasing sophistication within cloze tests: (1) keyword extraction using YAKE (Campos et al., 2020), which selects topically salient terms without external resources; (2) contextuality scoring adapted from Ondov et al. (2024), which uses a masked language model (MLM) to identify words whose predictability increases with broader passage context; and (3) contextuality-plus, our proposed extension, which conditions MLM predictions on the original source text alongside the summary. This source-text conditioning shifts the construct assessed from intra-summary coherence to cross-document discourse comprehension.

This paper makes three contributions. First, we introduce source-text-conditioned gap selection (contextuality-plus), which credibly outperforms keyword extraction on holistic quality and produces the most favorable distribution of gap types for discourse comprehension assessment. Second, we

present a human annotation study with a four-level gap taxonomy (unpredictable, sentence-, passage-, and source-predictable) that reveals method-level differences in the types of comprehension that gaps elicit. Third, we show that an eligibility filter can be added to the gap selection algorithm to favorably shift the distribution of gap types.

2 Related Work

2.1 Cloze Testing and Gap Selection

The cloze procedure was introduced by Taylor (1953) as a measure of text readability using fixed-ratio deletion. Subsequent research established that gap selection substantially affects what cloze exercises measure. Fixed-ratio deletion produces a mixture of items testing local syntax, vocabulary, and broader comprehension, making interpretation difficult (Shanahan et al., 1982; Watanabe and Koyama, 2008). A meta-analysis of second-language cloze research found that rational deletion—selecting specific words based on explicit criteria—yields more reliable and targeted assessments than fixed-ratio approaches (Watanabe and Koyama, 2008), though it requires manual effort and domain expertise.

One explicit criterion is the degree to which a word’s predictability depends on comprehension beyond the sentence level (Gellert and Elbro, 2013; Kleijn et al., 2019). Jensen and Elbro (2022) demonstrated in their “deep cloze” test that gaps requiring global inferences predict reading comprehension above and beyond sentence-level items. Kleijn et al. (2019) formalize this in their HyTeC-cloze method, requiring that gaps be unpredictable from local context but predictable from the global context without relying on extra-textual knowledge. Both approaches share an emphasis on discourse-level gap selection but rely on manual annotation. This study builds on these insights by automating the identification of discourse-level gaps using language model probability distributions.

2.2 Automated Cloze Generation with Language Models

Several approaches have applied LLMs to automate cloze generation. Yang et al. (2021) used BERT to identify keyword gaps and found that students who completed the resulting exercises were more engaged with the reading than a control group. Keim and Littman (2022) created “context clozes” for long-form reading using GPT-2 surprisal to target

words whose predictability increases with cross-sentence context. Most directly relevant, Ondov et al. (2024) introduced a contextuality algorithm for pedagogically aligned cloze generation that computes the divergence between a word’s probability given only its local sentence versus the full passage, using a masked language model. Words with high contextuality are substantially more predictable in full context, so selecting them as gaps should better assess reading comprehension.

2.3 Intelligent Textbooks and Formative Assessment

In iTELL, cloze exercises are automatically generated and presented at the bottom of a page of reading material. After reading the page, learners attempt to restore selected words in a summary of the page. Responses are scored by string matching (ignoring whitespace and casing), and learners see their score and correct answers immediately after submitting. Cloze exercises complement other iTELL activities such as summary writing with automated LLM-based feedback (Morris et al., 2023, 2024b) and automatic question generation (Morris et al., 2024a) by providing lower-stakes comprehension monitoring that can be completed in minutes.

3 Methods

3.1 Gap Selection Methods

Three methods of cloze gap generation were assessed: keyword, contextuality, and contextuality-plus (our method). All methods shared common preprocessing implemented with spaCy (en_core_web_sm). The first sentence of each summary was excluded from gap selection to preserve its topic-setting function. Eligibility filters restricted gaps to tokens that are at least three characters long, purely alphabetic, not stop words, and not tagged by spaCy as proper nouns, numbers, punctuation, symbols, or other non-words (the PROP, NUM, PUNCT, SYM, and X part-of-speech tags). A minimum distance of seven word positions between consecutive gaps prevented clustering that could make exercises artificially difficult.

An example exercise illustrating the three methods on a single summary is provided in Appendix B.

3.1.1 YAKE Baseline

The keyword method uses YAKE (Yet Another Keyword Extractor; Campos et al., 2020), an unsupervised algorithm that scores candidates using

term frequency, positional information, and co-occurrence patterns. We extract up to 10 unigram keywords from each summary and gap all valid occurrences (after the first sentence), subject to a minimum 30-character distance between gap positions. This method serves as a baseline representing the assumption that topically important terms make good comprehension gaps. Because YAKE gaps every occurrence, the number of gaps per exercise varies and is typically fewer than 10.

3.1.2 Contextuality

The contextuality method, adapted from Ondov et al. (2024), uses a masked language model to quantify how much each word’s predictability depends on discourse context beyond the immediate sentence. For each eligible word w_i in the summary $\mathcal{S}$, we compute two probability estimates using ModernBERT-large (Warner et al., 2024):

- **Full-context probability** ζ_i : the probability of w_i when masked in the full summary text, $P(w_i \mid \mathcal{S}_{\setminus i})$
- **Sentence-context probability** ψ_i : the probability of w_i when masked in only its containing sentence

The contextuality score is the pointwise KL divergence:

$$\text{score}(w_i) = \zeta_i \cdot \log_2 \frac{\zeta_i}{\psi_i} \quad (1)$$

Words scoring high are substantially more predictable given the full summary than from sentence context alone. For words tokenized into multiple subword units, all subwords are masked simultaneously and the first subword’s probability is used. Gaps are selected greedily, excluding candidates within ± 7 words of an already-selected term.

3.1.3 Contextuality-Plus

Contextuality-plus uses an identical algorithm to contextuality, except that the full-context probability conditions on the concatenation of the source page $\mathcal{P}$ and the summary $\mathcal{S}$, rather than the summary alone:

$$\zeta_i^+ = P(w_i \mid \mathcal{P} \oplus \mathcal{S}_{\setminus i}) \quad (2)$$

The sentence-context probability ψ_i remains unchanged. The contextuality-plus score is thus:

$$\text{score}^+(w_i) = \zeta_i^+ \cdot \log_2 \frac{\zeta_i^+}{\psi_i} \quad (3)$$

This modification identifies words whose predictability increases when the source text is available. High-scoring words under contextuality-plus are those predictable from the source but not from the sentence alone—the items that test whether a reader has comprehended the source material. ModernBERT-large’s 8,192-token context window accommodates the concatenated source-plus-summary input for all materials in our dataset.

3.2 Data

Our task of interest is to select appropriate gaps in a summary of a source text such that filling them requires comprehension of the source. Formally, given a source page $\mathcal{P}$ and a summary $\mathcal{S}$ of that page, the system selects a subset of content words in $\mathcal{S}$ to delete, producing a cloze exercise that assesses whether the reader has understood $\mathcal{P}$. This differs from standard cloze generation, where gaps are selected from and conditioned on a single text. Here, the text being gapped (the summary) serves as a vehicle for assessing comprehension of a separate document (the source page).

Materials consisted of 16 pages drawn from three iTELL volumes covering research methods in psychology and business communication. Each page is paired with a reference summary of approximately 150 to 300 words, drafted using generative AI (GPT 3.5 Turbo) and revised by an educational content editor to ensure accuracy and coverage. For each of the 16 summaries, three cloze exercises were generated (one per method), yielding 48 exercises total.

3.3 Annotation Taxonomy

To characterize the types of comprehension that each method’s gaps assess, we developed a four-level taxonomy based on the minimum information source required to fill each gap:

1. **Unpredictable**: The word cannot reasonably be predicted even with access to all available context. These represent poor items that frustrate test-takers.
2. **Sentence-predictable**: The word can be filled using only local sentence context (e.g., collocations, syntactic constraints). These assess surface-level linguistic processing.
3. **Passage-predictable**: The word requires information from other parts of the summary

but not the source text. These assess intra-summary coherence.

4. **Source-predictable:** The word requires information from the original source text. These are the ideal items for discourse comprehension assessment, because they test situation model construction.

This taxonomy maps onto levels of discourse representation in construction-integration theory (Kintsch, 1988; Zwaan and Radvansky, 1998). Sentence-predictable items test textbase processing, passage-predictable items test textbase integration across the summary, and source-predictable items test situation model construction requiring cross-document integration.

4 Human Evaluation Study

4.1 Participants and Procedure

Eleven annotators participated in the evaluation. Three were familiar with the cloze task design and its pedagogical objectives, while eight had experience with the broader iTELL platform but no prior involvement in cloze generation. Annotators were instructed to use the taxonomy to select the *minimum* level required, ensuring that source-predictable labels reflect genuine cross-document dependency. Each annotator completed between 2 and 6 annotation tasks, yielding 56 completed exercises in total (22 contextuality, 18 contextuality-plus, 16 keyword). The imbalance across methods arose from randomized assignment.

For each task, annotators first read the full source page in iTELL, then completed one cloze exercise derived from that page’s summary. The time spent completing the cloze exercise was recorded automatically by the platform as “time on task,” a secondary measure of exercise difficulty. After completing the exercise, annotators provided two types of judgments. At the item level, they categorized each gap according to the four-level taxonomy described in Section 3.3. At the exercise level, they provided a holistic quality rating for the cloze test on a four point scale: (1) did not measure comprehension, (2) reasonably good with some problematic items, (3) good with most items challenging but achievable, (4) excellent with all items challenging but achievable.

Our annotation budget was fixed at approximately one hour per annotator. We prioritized coverage across all 16 source pages over repeated rat-

Method	<i>n</i>	Mean	SD	Median
Keyword	16	2.25	1.13	2.0
Contextuality	22	2.45	0.74	2.0
Ctx-plus	18	2.78	0.88	3.0

Table 1: Holistic quality ratings by method (1–4 scale).

ings of the same exercises. The Bayesian analysis in Section 4.3 accounts for these sample size constraints through crossed random effects and posterior intervals that transparently reflect the resulting uncertainty.

4.2 Inter-Rater Reliability

Because annotators were assigned to different method–passage combinations, only 16 of 56 exercises were rated by two or more annotators. We report model-based intraclass correlation coefficients (ICCs) derived from Bayesian ordinal models’ variance components, which use all observations rather than only the overlapping subset.

Holistic scores. The unadjusted ICC was 0.24 (94% HDI [0.05, 0.49]), with an adjusted ICC of 0.35 (94% HDI [0.08, 0.64]) after removing annotator severity variance. This places most posterior mass in the poor-to-fair range, with $P(\text{ICC} > 0.40) = 0.36$ (Cicchetti, 1994). Annotator severity variance dominated the total ($\hat{\sigma}_{\text{annotator}}^2 = 2.82$ vs. $\hat{\sigma}_{\text{passage}}^2 = 1.20$), indicating that low ICC reflects a *calibration* problem rather than an absence of signal. Raters showed minimal absolute agreement but demonstrated non-trivial agreement in ranking the exercises. The Bayesian ordinal model handles this via annotator random intercepts, yielding method effect estimates that are appropriately regularized for rater heterogeneity.

Per-gap annotations. An analogous cumulative logit model for individual gap ratings yielded an adjusted ICC of 0.42 (94% HDI [0.29, 0.54]), with most posterior mass in the fair-reliability range ($P(\text{ICC} > 0.40) = 0.62$). The coincidence matrix (Appendix, Table 6) shows that disagreements concentrate between adjacent categories, suggesting raters broadly agree on whether a gap measures comprehension but disagree on the granularity of context required.

4.3 Holistic Quality Ratings

Holistic score distributions. Table 1 presents holistic quality ratings by method. Contextuality-plus received the highest mean rating ($M = 2.78$,

Contrast	β	94% HDI	$P(\beta > 0)$
Ctx+ vs. Kw	2.00	[0.50, 3.60]	0.994
Ctx+ vs. Ctx	1.25	[0.01, 2.65]	0.971
Ctx vs. Kw	0.75	[-0.59, 2.09]	0.856

Table 2: Pairwise contrasts from Bayesian ordinal model (log-odds scale).

$SD = 0.88$), followed by contextuality ($M = 2.45$, $SD = 0.74$) and keyword extraction ($M = 2.25$, $SD = 1.13$). Contextuality-plus was the only method with a median rating of 3 (good), while both other methods had median ratings of 2 (reasonably good).

Bayesian Ordinal Model. To assess differences across methods, we fit a Bayesian cumulative logit model with method as a fixed effect and crossed random intercepts for annotator and page, implemented in Bambi (Capretto et al., 2022) and PyMC (Abril-Pla et al., 2023):

$$\text{score} \sim \text{method} + (1 \mid \text{rater}) + (1 \mid \text{page}) \quad (4)$$

Four chains of 2,000 tuning and 2,000 sampling iterations yielded good convergence ($\hat{R} = 1.0$ for all parameters; effective sample sizes $> 1,800$). The estimated random effect standard deviations were 1.57 for annotator and 0.95 for page, confirming that both sources of variability are meaningful. Table 2 presents pairwise contrasts on the log-odds scale with 94% highest density intervals (HDIs).

The contextuality-plus versus keyword contrast is strongly supported, with a 99.4% posterior probability that contextuality-plus produces higher holistic ratings ($\beta = 2.00$, 94% HDI [0.50, 3.60]). The contextuality-plus versus contextuality contrast provides suggestive evidence of an advantage ($P(\beta > 0) = 0.971$, $\beta = 1.25$, 94% HDI [0.01, 2.65]), though the wide posterior interval reflects genuine uncertainty about the magnitude. The contextuality versus keyword contrast shows weaker evidence ($P(\beta > 0) = 0.856$), leaving it unclear whether contextuality scoring alone improves upon keyword-based gap selection.

4.4 Gap Type Annotations

Gap type distributions. Contextuality-plus produced the most favorable profile for discourse comprehension assessment (Table 3), with the highest proportion of source-predictable gaps (37%) and a relatively low proportion of unpredictable gaps (17%). The keyword method produced the fewest

Category	Kw ($n=95$)	Ctx ($n=212$)	Ctx+ ($n=163$)
Unpredict.	2 (2%)	42 (20%)	28 (17%)
Sentence	35 (37%)	70 (33%)	45 (28%)
Passage	33 (35%)	36 (17%)	29 (18%)
Source	25 (26%)	64 (30%)	61 (37%)

Table 3: Gap type distribution by method. Percentages within each method.

Category	n	Accuracy
Unpredictable	72	11.1%
Sentence	150	78.7%
Passage	98	79.6%
Source	150	80.0%

Table 4: Per-gap accuracy by annotation category.

unpredictable gaps (2%) but also the fewest source-predictable gaps (26%). Standard contextuality showed an intermediate profile but with the highest rate of unpredictable gaps (20%).

Pairwise contrasts from the per-gap cumulative logit model showed contextuality-plus gaps were credibly shifted toward higher-context categories relative to standard contextuality ($\beta = 0.48$, $P(\beta > 0) = 0.944$). Keyword gaps also scored higher than contextuality ($\beta = 0.64$, $P(\beta > 0) = 0.966$), primarily because keyword extraction produced almost no unpredictable items. No credible difference emerged between contextuality-plus and keyword ($\beta = -0.16$), though these methods reach similar ordinal positions through different profiles, with keyword achieving this through minimal unpredictability and contextuality-plus through elevated source-predictability.

Keyword extraction produces fewer gaps per exercise ($M = 5.94$) because it gaps only keyword occurrences, while contextuality methods target 10 gaps each ($M = 9.64$ and $M = 9.06$).

Per-gap accuracy. Table 4 presents gap-fill accuracy by annotation category, aggregated across methods. Accuracy was high and similar for sentence-predictable (78.7%), passage-predictable (79.6%), and source-predictable (80.0%) gaps, but sharply lower for unpredictable gaps (11.1%). Unpredictable gaps are genuinely unfillable, while the three predictable categories are roughly equally challenging once the relevant information source is available.

Time on task. Completion times differed significantly across methods (Kruskal-Wallis $H(2) = 8.67$, $p = .013$). Median times were 208 sec-

Category	Retained	Removed
Unpredictable	8%	31%
Sentence	30%	43%
Passage	18%	17%
Source	44%	9%

Table 5: Gap annotation category distribution for retained vs. removed gaps under the lemma overlap constraint.

onds for keyword, 441 seconds for contextuality-plus, and 655 seconds for contextuality exercises. Post-hoc Mann-Whitney tests showed a significant difference between keyword and contextuality ($U = 81$, $p = .005$), with other pairwise differences not reaching significance at conventional thresholds. The shorter keyword times likely reflect gaps fillable from local context, requiring less source-text consultation. There was no significant difference in response times between contextuality and contextuality-plus methods.

4.5 Eligibility Filter

The human evaluation identified that 17% of contextuality-plus gaps were unpredictable, making them frustrating for annotators. We explored a refinement to address this issue, which should be interpreted as exploratory rather than confirmatory.

Human coding indicated that unpredictable gaps typically arise when the MLM assigns high contextuality to words that do not appear in the source even though the underlying concept may be present. For example, a source text used the word “relieve” where the reference summary used the synonym “alleviate,” so a learner primed by the source might fill the gap with the wrong lexical form. We hypothesized that restricting gaps to words whose lemma appears in the source would preferentially remove such unpredictable gaps while retaining source-predictable ones.

We implemented this as an additional eligibility filter, then re-generated cloze exercises for all 18 contextuality-plus exercises in the annotated dataset and evaluated the impact using the existing human annotations.

The filter achieved its intended effect (Table 5). Unpredictable gaps constituted 31% of removed gaps but only 8% of retained gaps, while source-predictable gaps constituted 44% of retained gaps but only 9% of removed gaps.

The cost was modest. The average number of gaps per exercise decreased from 9.0 to 8.65 be-

cause the greedy algorithm replaces removed high-scoring gaps with lower-scoring lemma-matched alternatives. The filter also removes sentence-predictable gaps (43% of removed gaps), which were not desirable for our assessment goals.

5 Discussion

We present three methods for generating cloze gaps from summaries of source texts, framed as alternative operationalizations of discourse comprehension. The central argument is that our contextuality-plus method identifies words whose predictability depends on having read an external source text. Because our method shares identical algorithmic machinery with standard contextuality, the observed differences isolate source-text conditioning as the differentiating factor: credibly higher holistic ratings than keyword extraction ($\beta = 2.00$, 94% HDI [0.50, 3.60]), suggestive evidence of an advantage over standard contextuality ($\beta = 1.25$, 94% HDI [0.01, 2.65]), and the highest proportion of source-predictable gaps (37% vs. 30% and 26%). The four-level taxonomy, mapped to construction-integration theory (Kintsch, 1988; Zwaan and Radvansky, 1998), supports this construct argument by linking each gap category to a distinct level of discourse representation.

Completion-time patterns offer converging evidence. Keyword exercises were completed fastest (median 208s), consistent with gaps that are largely predictable from local context. Contextuality-plus exercises took roughly twice as long (441s), consistent with the need to consult the source text. Standard contextuality exercises took longest (655s), likely reflecting extended deliberation on its higher rate of unpredictable gaps (20%).

An exploratory lemma overlap filter reduced unpredictable gaps from 17% to 8% while preserving source-predictable items, though it requires validation on independent data. The resulting pipeline is lightweight (a masked language model and spaCy), domain-agnostic, and suitable for integration with intelligent textbook platforms. Future work should validate with student populations, evaluate across domains, and explore distractor generation for multiple-choice variants.

Limitations

This study has several limitations. The sample size is modest, with 56 completed exercises from 11 annotators (16–22 per method), which limits

statistical power. The post-hoc refinement in Section 4.5 was developed on the same dataset as the primary evaluation and should be tested experimentally on new data. Our annotator pool consisted of lab members rather than students encountering the material in a course, and their interaction with the exercises may differ from the target population. All analyses used a single masked language model (ModernBERT-large), and we have not evaluated sensitivity to model choice. Finally, we did not measure learning outcomes, which are the ultimate validity criterion for formative assessment and require longitudinal studies comparing cloze-augmented instruction to control conditions.

References

- Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. 2023. [PyMC: a modern, and comprehensive probabilistic programming framework in Python](#). *PeerJ Computer Science*, 9(e1516).
- Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. [YAKE! keyword extraction from single documents using multiple local features](#). *Information Sciences*, 509:257–289.
- Tomás Capretto, Camen Piho, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A Martin. 2022. [Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python](#). *Journal of Statistical Software*, 103(15):1–29.
- Domenic V. Cicchetti. 1994. [Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology](#). *Psychological Assessment*, 6(4):284–290.
- Scott Crossley, Joon Suh Choi, Wesley Morris, Langdon Holmes, David Joyner, and Vaibhav Gupta. 2024. [Using intelligent texts in a computer science classroom: Findings from an iTELL deployment](#). In *Proceedings of the 8th Educational Data Mining in Computer Science Education Workshop*.
- Anna S. Gellert and Carsten Elbro. 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.
- Katrine Lyskov Jensen and Carsten Elbro. 2022. [Clozing in on reading comprehension: A deep cloze test of global inference making](#). *Reading and Writing*, 35(5):1221–1237.
- Greg Keim and Michael Littman. 2022. [Selecting context clozes for lightweight reading compliance](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 167–172. Association for Computational Linguistics.
- Walter Kintsch. 1988. [The role of knowledge in discourse comprehension: A construction-integration model](#). *Psychological Review*, 95(2):163–182.
- Suzanne Kleijn, Henk Pander Maat, and Ted Sanders. 2019. [Cloze testing for comprehension assessment: The HyTeC-cloze](#). *Language Testing*, 36(4):553–572.
- Danielle S. McNamara and Joseph Magliano. 2009. [Toward a comprehensive model of comprehension](#). In *Psychology of Learning and Motivation*, volume 51, pages 297–384. Elsevier.
- Wesley Morris, Joon Suh Choi, Langdon Holmes, Vaibhav Gupta, and Scott Crossley. 2024a. [Automatic question generation and constructed response scoring in intelligent texts](#). In *Proceedings of the Workshop on Large Language Models in and for Natural Language Generation, Evaluation, and Translation (L3MNGE)*.
- Wesley Morris, Scott Crossley, Langdon Holmes, Chaohua Ou, Mihai Dascalu, and Danielle McNamara. 2024b. [Formative feedback on student-authored summaries in intelligent textbooks using large language models](#). *International Journal of Artificial Intelligence in Education*, 35(3):1022–1043.
- Wesley Morris, Scott A. Crossley, Langdon Holmes, Chaohua Ou, Danielle McNamara, and Mihai Dascalu. 2023. [Using large language models to provide formative feedback in intelligent textbooks](#). In *Proceedings of the International Conference on Artificial Intelligence in Education*, pages 484–489. Springer.
- John W. Oller. 1972. [Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language](#). *The Modern Language Journal*, 56(3):151.
- Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2024. [Pedagogically aligned objectives create reliable automatic cloze tests](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3961–3972. Association for Computational Linguistics.
- Timothy Shanahan, Michael L. Kamil, and Aileen Webb Tobin. 1982. [Cloze as a measure of intersentential comprehension](#). *Reading Research Quarterly*, 17(2):229–255.
- Wilson L. Taylor. 1953. [“cloze procedure”: A new tool for measuring readability](#). *Journalism Quarterly*, 30(4):415–433.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.

Yukiko Watanabe and Dennis Koyama. 2008. [A meta-analysis of second language cloze testing research](#). *Second Language Studies*, 26(2):103–133.

Albert C. M. Yang, Irene Y. L. Chen, Brendan Flanagan, and Hiroaki Ogata. 2021. [Automatic generation of cloze items for repeated testing to improve reading comprehension](#). *Educational Technology & Society*, 24(3):147–158.

Rolf A. Zwaan and Gabriel A. Radvansky. 1998. [Situation models in language comprehension and memory](#). *Psychological Bulletin*, 123(2):162–185.

A Inter-Annotator Agreement Details

	Unpr.	Sent.	Pass.	Source
Unpredictable	42	41	7	20
Sentence	41	80	29	34
Passage	7	29	26	22
Source	20	34	22	80

Table 6: Coincidence matrix for per-gap annotations from 138 overlapping gaps. Diagonal entries indicate agreement; off-diagonal concentration between adjacent “unpredictable” and “sentence” categories indicates disagreement.

B Example Cloze Exercise

Figure 1 presents an example summary gapped by all three methods, illustrating their differences. The keyword method gaps only repetitions of “cultural”/“cultures.” Standard contextuality selects diverse words but includes low-information items (e.g., “like,” “evolves”) fillable from sentence context alone. Contextuality-plus replaces those with domain-specific terms (“masculinity,” “uncertainty,” “significance”) that require source-text comprehension.

Original summary (page: *Divergent Cultural Characteristics*)

It contrasts individualistic cultures, such as the United States, where personal freedom and independence are emphasized, with collectivist cultures, which prioritize community and group needs. The work of Geert Hofstede is highlighted, focusing on how different cultures perceive individualism, collectivism, and other dimensions like time orientation and masculinity-femininity. Uncertainty reduction theory by Charles Berger and Richard Calabrese is discussed, demonstrating how communication evolves as uncertainty decreases. The text also examines the cultural significance of materialism versus relational values, showing how these differences influence personal and business interactions across the globe. Understanding these cultural nuances is essential for effective intercultural communication and navigating global business landscapes.

Keyword (6 gaps, all from keyword family *cultural/cultures*):

It contrasts individualistic _____, such as the United States, where personal freedom and independence are emphasized, with collectivist _____, which prioritize community and group needs. ... focusing on how different _____ perceive individualism, collectivism, and other dimensions like time orientation and masculinity-femininity. ... The text also examines the _____ significance of materialism versus relational values, showing how these differences influence personal and business interactions across the globe. Understanding these _____ nuances is essential for effective inter_____ communication and navigating global business landscapes.

Contextuality (10 gaps):

It contrasts individualistic _____, such as the United States, where personal _____ and independence are emphasized, with collectivist _____, which prioritize community and group needs. ... focusing on how different _____ perceive individualism, collectivism, and other dimensions _____ time orientation and masculinity-femininity. ... demonstrating how communication _____ as uncertainty decreases. The text also _____ the cultural significance of materialism versus relational values, showing how these _____ influence personal and business interactions across the globe. _____ these cultural nuances is essential for effective intercultural communication and _____ global business landscapes.

Contextuality-plus (10 gaps):

It contrasts individualistic _____, such as the United States, where personal _____ and independence are emphasized, with collectivist _____, which prioritize community and group needs. ... focusing on how different _____ perceive individualism, collectivism, and other dimensions like time orientation and _____-femininity. ... demonstrating how communication evolves as _____ decreases. The text also examines the cultural _____ of materialism versus relational values, showing how these _____ influence personal and business interactions across the globe. _____ these cultural nuances is essential for effective intercultural communication and _____ global business landscapes.

Figure 1: Example cloze exercise for one summary gapped by all three methods. Keyword selects only repetitions of *cultural/cultures*. Contextuality and contextuality-plus share 7 of 10 gap positions; the 3 positions that differ (sentences 3–5) show contextuality-plus replacing sentence-predictable items with source-dependent domain terms.