

# Engineering a Semantic Topicality Instrument for Multiple-Choice Question Quality Control

Michelle Banawan, Linh Huynh, Katerina Christhif, & Danielle S. McNamara

Distractor plausibility

Semantic similarity

Topic relevance

*Ensuring that multiple-choice question (MCQ) distractors are conceptually relevant yet incorrect is critical for test validity and learner modeling. This study evaluates semantic topicality scoring as an instrument for automated MCQ quality control. A corpus of 20 reading-comprehension items (80 options) was annotated for topical relevance and analyzed using lexical and semantic similarity measures relative to passage main ideas. SBERT achieved the strongest correspondence with human judgments (Spearman's  $\rho = 0.54$ ; AUC = 0.90). An empirically derived midpoint threshold classified 45% of distractors as on-topic, providing an interpretable operationalization of the rubric. Following a nested learning engineering process, we iteratively designed, implemented, and validated this semantic topicality instrument as an actionable diagnostic for AI-assisted item development. Results demonstrate that embedding-based similarity can serve as a scalable, human-aligned signal for content validity, supporting automated item screening, hybrid human–AI review, and future adaptive question-generation workflows.*

## Introduction

Multiple-choice questions (MCQs) remain among the most prevalent tools in educational assessment due to their efficiency, objectivity, ease of scoring, and reliability in large-scale testing contexts (Haladyna et al., 2002). Within the context of reading comprehension, high-quality MCQs are characterized by content relevance, clarity, and well-designed distractors that differentiate between varying levels of comprehension (Haladyna et al., 2002; Gorgun & Bulut, 2024). Effective items align with intended learning objectives, measure meaningful comprehension rather than surface recall, and employ plausible yet clearly incorrect distractors to maintain discriminative power.

Producing and reviewing such distractors is resource-intensive and difficult to scale. Item writers must ensure conceptual alignment with passage ideas, and quality reviewers must verify this alignment, both of which require time and expertise. As assessment programs increasingly rely on large item pools and AI-assisted item generation, there is a growing need for automated, interpretable methods to evaluate distractor topicality and support sustainable quality-control processes. Prior work shows that AI-generated MCQs often exhibit factual errors, reduced conceptual relevance, and limited higher-order reasoning (Cheung et al., 2023; Law et al., 2025), reinforcing the need for systematic topicality diagnostics regardless of item

source. Effective automation must therefore align with psychometric quality criteria such as appropriateness, clarity, relevance, discriminative power, and cognitive challenge (Feng et al., 2024).

To address this challenge, this study follows a Learning Engineering approach in which distractor topicality is treated as an engineering problem addressed through iterative cycles of design, instrumentation, implementation, and empirical investigation (Goodell, 2023; Craig et al., 2025). Consistent with Learning Engineering efforts to develop measurement frameworks within complex learning environments, we construct and validate a semantic topicality instrument using nested cycles of development and evaluation.

## Methods

### Dataset and Annotation

The dataset comprises 20 reading-comprehension questions evenly distributed between two subject domains: History and Lifespan Development. Each question includes one correct answer and three distractors, yielding 80 options (20 correct and 60 distractors). Each question is associated with a short passage and a main-idea summary written by reading comprehension experts to capture the essential conceptual content. These summaries serve as the reference text for measuring topical overlap between an option and its passage.

Options averaged 5 words and main-idea summaries averaged 82 words. Two expert annotators independently rated each option–main idea pair for topical relevance using a four-point ordinal rubric (Table 1).

Table 1.  
Annotation Rubric for Topical Relevance

Rating	Description	Conceptual coverage
1	Disagree	Off-topic, does not address key concepts
2	Neutral	Peripheral, partially covered in sub-points
3	Agree	Addressed in sub-points of minor ideas
4	Strongly Agree	Fully covered in a main idea

Inter-rater agreement between the two annotators was 75% exact match (weighted Cohen's  $\kappa = 0.44$ ), indicating fair consistency in ordinal judgments of topical relevance. Disagreements occurred on five of the twenty items, typically involving adjacent category ratings (e.g., 3 vs 4). A deterministic majority score was derived for subsequent analyses by taking the mode of the two ratings (lower value in ties). The distribution of majority topical-relevance scores confirms that most answer options, including correct answers by design, were judged as highly relevant to their source text. This is consistent with the construction of pedagogically plausible distractors.

### Similarity Metrics

To estimate topical relevance computationally, three complementary similarity measures were applied to each option–main idea pair. First, TF–IDF cosine similarity quantified the weighted lexical overlap between the two texts, emphasizing shared content words while down-weighting frequent or uninformative terms. Second, Jaccard overlap measured the ratio of shared

tokens to the total unique tokens across both texts, offering a simple indicator of vocabulary intersection. Finally, Sentence-BERT (SBERT) cosine similarity computed the cosine similarity between contextual embeddings generated by the all-MiniLM-L6-v2 model. This embedding-based method captures semantic proximity beyond surface wording, allowing conceptually related expressions to be recognized even when they share few lexical items. Together, these metrics allowed a direct comparison between lexical similarity (TF-IDF, Jaccard) and semantic similarity (SBERT) in predicting human topicality judgments.

Although main-idea summaries were substantially longer than the option texts, this asymmetry reflects authentic item structure. Embedding-based measures such as SBERT are robust to such length differences, whereas lexical metrics (TF-IDF, Jaccard) can exhibit mild sensitivity. These nuances were recognized and considered in the interpretation of results.

## Evaluation Procedures

The alignment between computational measures and human ratings was examined through both correlational and classification-based analyses. Pearson's  $r$  and Spearman's  $\rho$  were computed to assess, respectively, linear and rank-order associations between similarity scores and human topical-relevance ratings. Ratings 3–4 were treated as on-topic and 1–2 as off-topic. Performance was assessed using empirically derived midpoint thresholds, defined as the midpoint between the mean similarity scores of correct and distractor options within each metric. For SBERT, this procedure yielded a midpoint threshold of 0.384, which was subsequently used for semantic classification analyses.

## Results

### Corpus Characteristics

Tokenization and part-of-speech tagging revealed that answer options were noun and adjective-dominant, while main-idea texts exhibited richer syntactic diversity. Readability analysis using the Flesch–Kincaid Grade Level metric showed that options averaged Grade 12.2, typical of tertiary-level assessments, whereas main-idea summaries averaged Grade 14.9, indicating dense, multi-sentence conceptual language. Vocabulary analysis identified 133 unique tokens in options and 251 in main-idea summaries, with minimal direct word overlap (mean Jaccard = 0.034). These findings confirm that topical relevance in this dataset depends primarily on semantic rather than lexical similarity, supporting its suitability for testing embedding-based models.

### Correlation with Human Judgments

All three computational measures showed statistically significant positive associations with the human topical-relevance ratings ( $p < .001$ ). The lexical approaches (TF-IDF and Jaccard) demonstrated moderate correlations with expert judgments, with Pearson's  $r$  values of 0.48 and 0.50, respectively, and comparable rank-order associations (Spearman's  $\rho = 0.55$  for both). These results indicate that word-level overlap captures part of the relationship between option content and the passage's main ideas.

However, the SBERT semantic similarity measure exhibited the strongest correspondence with human ratings, achieving a Spearman's  $\rho$  of 0.54 and an area under the ROC curve (AUC) of 0.90, the highest among all tested models. This pattern suggests that the embedding-based approach more effectively represents the conceptual alignment that human raters perceive, particularly when lexical overlap is limited or paraphrased.

Table 3.

Classification performance at empirical thresholds

Measure	Threshold midpoint	Precision	Recall	$F_1$	Threshold conservative	Precision	Recall	$F_1$
TF-IDF	0.072	1.00	0.41	0.58	0.06	1.00	0.49	0.65
Jaccard	0.050	1.00	0.45	0.62	0.05	1.00	0.51	0.68
SBERT	0.384	1.00	0.50	0.67	0.38	1.00	0.51	0.68

Table 4.  
On-Topic Classification using SBERT

Category	Count	% of Distractors
On-topic	27	45%
Off-topic	33	55%

Applying the SBERT threshold = 0.384 categorized 45% of distractors (27/60) as on topic. These corresponded to human relevance scores  $\geq 3$ , validating the threshold as a practical operationalization of rubric-based topical coverage.

## Example Cases

A qualitative examination of individual items further illustrates the distinction between high- and low-scoring distractors identified by the semantic model.

Distractors receiving high SBERT similarity scores ( $\geq 0.60$ ) typically paraphrased or recontextualized key concepts from the passage while maintaining overall conceptual fidelity. For instance, the option “Language acquisition is entirely dependent on social interactions and caregiver reinforcement” (SBERT = 0.69) closely reflects the thematic content of the passage on language development, emphasizing the role of social interaction. Similarly, the distractor “King James II successfully implemented a Catholic monarchy” (SBERT = 0.67) retains the central historical idea of religious absolutism underlying the text, despite presenting it in an overstated or misleading form.

In contrast, low-scoring distractors (SBERT < 0.15) tended to introduce information that was either tangential or factually irrelevant to the passage’s conceptual focus. Examples include “Experience is not necessary at all” (SBERT = 0.07), which directly contradicts the discussion of experiential learning in the source text, and “It improved relations with Native Americans” (SBERT = 0.09), which diverges entirely from the historical themes addressed. These comparisons demonstrate that high semantic similarity corresponds to distractors that remain topically grounded yet incorrect, whereas low similarity identifies options that deviate substantially from the main conceptual domain of the passage.

Correlations between similarity scores and answer correctness were weak ( $r = 0.06$ – $0.27$ ), confirming that topical relevance is distinct from factual correctness. This distinction reinforces the rubric’s focus on conceptual relevance rather than accuracy per se.

## Discussion

The findings confirm that semantic embeddings capture conceptual relationships underlying human topicality judgments more effectively than word-overlap metrics. SBERT's performance (Spearman's  $\rho = 0.54$ , AUC = 0.90) demonstrates that embedding-based similarity can operationalize topic relevance in a scalable and interpretable way. The empirically derived midpoint threshold (0.384) provides a practical rule for identifying distractors that are conceptually aligned with a passage's key ideas. Correlations between similarity scores and answer correctness were weak ( $r = 0.06$ – $0.27$ ), confirming that topical relevance is distinct from factual accuracy—a distinction central to the rubric's focus on conceptual alignment rather than truth value.

The qualitative cases further illustrate the distinction that high-similarity distractors paraphrased or reframed key ideas while remaining conceptually grounded, whereas low-similarity options introduced tangential or irrelevant content. This pattern reinforces that high semantic similarity captures meaningful topical fidelity rather than mere wording overlap. Viewing this work through a Learning Engineering lens clarifies its next steps and practical value. The SBERT threshold provides an initial instrument for automated item screening, but real-world deployment requires iterative cycles of refinement: (a) scale testing across diverse domains, (b) psychometric linking of topicality to item difficulty/discrimination, (c) teacher-in-the-loop evaluation to calibrate tolerances for “plausible” distractors, and (d) integration into AI-item generation pipelines as a hybrid human–AI filter. Following these cycles will move the instrument from a promising diagnostic to a production-ready quality-control tool that aligns NLP diagnostics with educational validity requirements.

## Conclusion and Future Work

This study demonstrated that computational similarity measures can approximate expert judgments of topical relevance in multiple-choice questions. Among the evaluated approaches, SBERT semantic embeddings showed the highest correspondence with human annotations (Spearman's  $\rho = 0.54$ , AUC = 0.90), outperforming lexical baselines such as TF–IDF and Jaccard overlap. These findings indicate that embedding-based topicality scoring can serve as a quantitative, interpretable indicator of content validity, supporting the automated review and generation of assessment items.

Future work will extend this analysis across larger and more diverse domains to test generalizability. Further, we will explore hybrid models that integrate lexical and semantic features, and evaluate how topicality scores relate to psychometric properties such as item difficulty and discrimination. Embedding-based topicality scoring could also be integrated into AI-assisted item development systems to flag weak and off-topic distractors, guide generative models, and provide conceptual diagnostics of learner misunderstanding. Collectively, these directions aim to refine topicality scoring as a scalable, human-aligned tool for maintaining the validity and quality of educational assessments.

## Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305N210041 and R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## References

- Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., ... & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). *PLoS one*, 18(8), e0290691.
- Craig, S. D., Avancha, K., Malhotra, P., C., J., Verma, V., Likamwa, R., Gary, K., Spain, R., & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In *International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale* (pp. 115-132). <https://doi.org/10.59668/2109.21735>

- Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., ... & Lan, A. (2024, June). Exploring automated distractor generation for math multiple-choice questions via large language models. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 3067-3082).
- Goodell, J., Kessler, A., & Schatz, S. (2023). Learning Engineering at a Glance. *Journal of Military Learning*.
- Gorgun, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*, 29, 24111–24142. <https://doi.org/10.1007/s10639-024-12771-3>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334. [https://doi.org/10.1207/S15324818AME1503\\_5](https://doi.org/10.1207/S15324818AME1503_5)
- Law, A. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Kei-ching Hung, K., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education*, 25(1), 208.

