

Supporting Distractor Quality Review Through Interpretable Semantic and Lexical Diagnostics

Michelle Banawan, Shubham Chakraborty, Katerina Christhilf,
Linh Huynh, Tracy Arner, and Danielle S. McNamara

Learning Engineering Institute, Arizona State University

{mbanawan, schak103, kchristh, lthuynh1, tarner, dsmcnam}@asu.edu

Abstract

Multiple-choice questions (MCQs) remain a widely used assessment format, yet distractor quality is often evaluated through qualitative review processes that can be difficult to scale and apply consistently. We present MCQ-Diag, a measurement-informed diagnostic tool that supports option-level inspection of distractor quality in both human-authored and AI-generated assessments. The system provides interpretable evidence related to three characteristics of distractor quality: semantic plausibility, semantic uniqueness, and lexical distinctiveness. Semantic plausibility reflects topical alignment between an answer option and the source passage, semantic uniqueness captures relationships among distractors within a question, and lexical distinctiveness identifies surface-level redundancy among options.

Rather than assigning automated quality judgments, MCQ-Diag presents these indicators through an interactive environment that supports item review, revision, and quality assurance. Diagnostic reference bands were derived from a corpus of 200 expert-reviewed MCQs and are intended to support interpretation rather than classification. Expert validation demonstrated modest but statistically significant associations between the semantic plausibility and lexical distinctiveness indicators and expert ratings of distractor plausibility and linguistic quality; semantic uniqueness is presented as a descriptive indicator not yet validated against expert judgment.

By translating established distractor-writing principles into transparent, inspectable indicators, MCQ-Diag provides a practical tool for supporting distractor review in contemporary AI-assisted assessment environments.

1 Introduction

Multiple-choice questions (MCQs) remain one of the most widely used assessment formats in education because they are efficient to administer,

score, and interpret. However, the quality of an MCQ depends not only on the correctness of the keyed response but also on the quality of its distractors. Distractors that are implausible, redundant, or overly attractive can reduce item discrimination, introduce ambiguity, and weaken the validity of score interpretations. For this reason, item-writing guidelines have long emphasized the importance of developing distractors that are plausible, distinct, and aligned with common misconceptions or errors (Haladyna et al., 2002; Afzal and Mitkov, 2014).

Despite their importance, distractor characteristics are often evaluated through qualitative review processes that rely heavily on expert judgment. Experienced item writers can usually recognize when a distractor appears weak or redundant, but identifying the specific source of the problem is not always straightforward. In practice, distractor review is frequently conducted at the item level, making it difficult to inspect how individual options function relative to one another. As a result, important properties such as topical alignment, redundancy among distractors, and excessive similarity to the correct answer may be overlooked during item development.

This challenge has become increasingly relevant with the growth of AI-assisted item generation. Recent work has demonstrated that large language models can generate multiple-choice questions and distractors at scale (Yu et al., 2024; Feng et al., 2024; Tian et al., 2026a). While these approaches improve the efficiency of item production, they also increase the need for tools that support systematic review of generated content. Generating distractors is no longer the primary bottleneck; evaluating and refining distractor quality has become an equally important concern. Existing approaches typically focus on generating distractors or assigning overall quality scores (Raina et al., 2023; Lee et al., 2025). Far fewer tools support inspection of distractor behavior at the option level.

To address this need, we developed MCQ-Diag, a diagnostic tool for reviewing distractor quality through interpretable semantic and lexical indicators. The system focuses on three characteristics of distractor quality: semantic plausibility, semantic uniqueness, and lexical distinctiveness. Semantic plausibility reflects the degree to which an option aligns with the source passage. Semantic uniqueness reflects the extent to which distractors represent distinct alternatives rather than overlapping interpretations. Lexical distinctiveness captures surface-level redundancy that may not be evident from semantic similarity alone. Rather than assigning automated quality judgments, MCQ-Diag presents these indicators through an interactive interface designed to support human review, revision, and quality assurance.

By making option-level relationships visible, MCQ-Diag allows assessment developers to inspect distractor sets, identify potential weaknesses, and compare patterns across individual items and larger item pools.

2 Distractor Quality as a Measurement Problem

Distractor quality has long been recognized as an important component of effective multiple-choice assessment. Item-writing guidelines consistently recommend that distractors be plausible, distinct from one another, and aligned with likely misconceptions or errors (Haladyna et al., 2002; Afzal and Mitkov, 2014). Although distractor plausibility and distinctiveness are widely recognized as important characteristics of item quality, they have historically been evaluated primarily through expert review rather than through explicit diagnostic evidence (Haladyna et al., 2002).

One challenge is that distractor quality is not a single characteristic. A distractor may be topically unrelated to the source material, making it easy to eliminate. Two distractors may express essentially the same idea, reducing the effective number of alternatives available to examinees. In other cases, a distractor may be so attractive that it competes directly with the keyed response. Although experienced item writers often identify these issues during review, the evidence supporting such judgments is not always explicit.

This challenge becomes more pronounced in AI-assisted assessment environments. Recent advances in automatic question generation have made

it possible to produce large numbers of candidate items and distractors (Yu et al., 2024; Feng et al., 2024; Tian et al., 2026a). As the volume of generated content increases, manual review remains essential but becomes increasingly difficult to scale. Assessment developers need tools that help make distractor behavior visible and interpretable without replacing expert judgment.

MCQ-Diag was developed to support this review process. The system focuses on three characteristics of distractor quality: semantic plausibility, semantic uniqueness, and lexical distinctiveness. Together, these indicators provide complementary evidence about how distractors function within a question. Rather than assigning quality scores or automated decisions, the system presents these indicators through an interactive diagnostic environment that supports review and quality assurance.

3 MCQ-Diag

MCQ-Diag is an interactive diagnostic environment designed to support the review of distractor quality in multiple-choice assessments. The system provides option-level evidence related to semantic plausibility, semantic uniqueness, and lexical distinctiveness through a web-based interface that supports both single-item inspection and corpus-level analysis. A public version of the tool is available at <https://mcq-diagnostics.lei-tech.org/>.

Figure 2 presents the main diagnostic interface. Users provide a source passage, a question stem, and a set of answer options. The system then computes diagnostic indicators for each option and displays the results through interactive visualizations and summary statistics. In addition to single-item analysis, MCQ-Diag supports batch processing of MCQ collections, allowing users to identify recurring distractor patterns across larger item pools.

The diagnostic framework consists of three complementary indicators. Semantic plausibility evaluates topical alignment between an answer option and the source passage. Semantic uniqueness examines relationships among distractors within a question to identify potential redundancy or imbalance. Lexical distinctiveness evaluates surface-level overlap among options and identifies distractors that may be overly similar in wording. Together, these indicators provide a multidimensional view of distractor quality that can support item development and quality assurance.

3.1 Semantic Plausibility

Semantic plausibility refers to the degree to which a distractor is topically and conceptually related to the source material. Plausible distractors require examinees to engage with the content of the item because they cannot be dismissed immediately as irrelevant or obviously incorrect. In contrast, distractors that are weakly related to the source passage may reduce item difficulty by allowing test-takers to eliminate options without demonstrating understanding (Afzal and Mitkov, 2014; Banawan et al., 2026).

MCQ-Diag estimates semantic plausibility by comparing each answer option to the source passage using sentence-level semantic representations. The passage and each option are encoded independently using Sentence-BERT embeddings (Reimers and Gurevych, 2019). Cosine similarity between the passage representation and the option representation is then calculated to produce an alignment score. Higher values indicate stronger topical alignment between the option and the passage.

The resulting scores are not interpreted as correctness judgments. Instead, they serve as diagnostic indicators that help reviewers inspect how strongly each distractor aligns with the content of the source material. Within the interface, alignment scores are presented alongside qualitative reference bands that support interpretation during item review. Distractors with very low alignment may warrant revision because they are potentially off-topic, whereas distractors with unusually high alignment may require additional inspection because they can become overly competitive with the keyed response.

Semantic plausibility is intended to support review rather than determine distractor quality. The indicator provides evidence about topical alignment that assessment developers can consider alongside other content and instructional considerations.

3.2 Semantic Uniqueness

While semantic plausibility focuses on the relationship between an individual distractor and the source passage, semantic uniqueness focuses on relationships among distractors within the same question. Effective distractors should represent distinct alternatives rather than variations of the same incorrect idea. When distractors overlap substantially in meaning, the effective number of alternatives is

reduced and the item may provide unintended cues to examinees (Foulonneau and Grouès, 2012; Tian et al., 2026b).

MCQ-Diag approaches semantic uniqueness as a relational property rather than an independent score assigned to individual distractors. After computing semantic plausibility scores for all answer options, the system examines the distribution of alignment values within a question. Summary statistics including the mean, range, and standard deviation of option-level alignment scores are used to characterize the internal structure of the distractor set.

Questions with tightly clustered alignment scores may indicate semantic redundancy among distractors because multiple options occupy similar regions of the semantic space. In contrast, large differences among alignment scores may indicate imbalance, where some distractors are substantially less plausible than others. The system also identifies cases in which a distractor exhibits stronger passage alignment than the keyed response, a pattern that may indicate distractor overcompetitiveness or potential ambiguity in the item.

Rather than classifying distractors into predefined categories, MCQ-Diag presents these relational patterns as evidence for review. This approach allows assessment developers to inspect the overall structure of the distractor set and identify potential issues that may not be apparent when distractors are evaluated individually. By emphasizing relationships among options, semantic uniqueness complements semantic plausibility and provides additional insight into how distractors function within a multiple-choice item. Unlike semantic plausibility and lexical distinctiveness, semantic uniqueness has not yet been directly validated against independent expert ratings; it is currently best understood as a descriptive, corpus-derived indicator rather than a validated diagnostic measure, and is treated as such throughout this paper.

3.3 Lexical Distinctiveness

Semantic similarity alone may not fully capture redundancy among distractors. Because modern embedding models are intentionally tolerant of paraphrase, distractors that differ only through minor wording changes can appear semantically distinct while still functioning as near-duplicates from the perspective of an examinee. This issue is particularly relevant in multiple-choice questions, where answer options are often short and small wording differences can mask substantial overlap in mean-

ing.

To address this limitation, MCQ-Diag incorporates a lexical distinctiveness indicator that evaluates surface-level overlap among answer options. For each option, the system computes IDF-weighted Jaccard similarity with all other options in the same question and identifies the highest degree of overlap. Lexical distinctiveness is then defined relative to this worst-case comparison, reflecting the assumption that a single near-duplicate distractor may be sufficient to reduce the diversity of the option set.

The lexical distinctiveness indicator is based on IDF-weighted Jaccard similarity among answer options. IDF weighting emphasizes content-bearing terms while reducing the influence of common function words. Lexical distinctiveness is computed as one minus the maximum pairwise overlap observed for a given option, making the measure sensitive to potentially redundant distractors that share substantial lexical content.

Lexical distinctiveness complements the semantic indicators by identifying surface-level redundancy that may not be apparent from semantic similarity alone. Together, the three indicators provide a more complete view of distractor quality during item review.

3.4 Interactive Diagnostic Environment

MCQ-Diag is implemented as a web-based diagnostic environment that supports both single-item review and corpus-level analysis. The system is publicly available at <https://mcq-diagnostics.lei-tech.org/>. Figure 2 presents the primary review interface used for single-item inspection.

To analyze an item, users provide a source passage, a question stem, a set of answer options, and optionally the keyed response. The dashboard reports semantic plausibility and lexical distinctiveness at the option level, with semantic uniqueness derived from these scores and shown separately as a descriptive summary. This lets users inspect individual distractors rather than relying on item-level summaries alone.

The interface supports common review tasks. Semantic plausibility flags distractors that may be insufficiently related to the source material; semantic uniqueness highlights redundancy and imbalance across the distractor set; lexical distinctiveness catches options with substantial surface-level overlap. Together, these give reviewers complementary evidence for revision.

MCQ-Diag also includes a batch analytics mode for larger item collections. Users upload item sets and get aggregate summaries of distractor behavior across a corpus, making it easier to spot recurring issues, such as weak distractors, redundant options, and overly competitive options, that aren't obvious item by item.

Revision is iterative, with users able to reanalyze an item immediately after editing a distractor and compare the updated profile against the original, giving assessment developers real-time feedback on their edits and supporting a human-in-the-loop workflow for item development and quality assurance.

4 Evaluation

4.1 Expert Validation

To examine whether the diagnostic indicators captured meaningful characteristics of distractor quality, we conducted an expert validation study using a subset of the calibration corpus. After excluding questions with purely numeric answer options, the validation set comprised 180 MCQs containing 720 answer options.

Domain experts evaluated all answer options using a four-point rubric assessing semantic plausibility and linguistic quality. Final ratings were established through expert consensus and served as external criteria for evaluating the diagnostic indicators. Ratings were established through expert discussion and adjudication rather than independent double-rating; disagreements were resolved through consensus before a final rating was recorded.

Semantic plausibility scores were positively associated with expert ratings of distractor plausibility ($\rho = .44$, 95% CI [.38, .50], $p < .001$), while lexical distinctiveness scores were positively associated with expert ratings of linguistic quality ($\rho = .37$, 95% CI [.31, .43], $p < .001$). These associations are modest in magnitude but statistically significant, and suggest that the indicators capture characteristics associated with distractor quality identified by human reviewers.

The observed associations are consistent with the intended role of MCQ-Diag as a decision-support tool rather than an automated evaluation system. In addition, informal review of diagnostic outputs suggested that flagged patterns often corresponded to recognizable distractor quality concerns, including redundancy, imbalance among options, and surface-level overlap.

4.2 Distractor Revision Analysis

Beyond supporting inspection, MCQ-Diag was designed to facilitate iterative distractor revision. To evaluate this capability, we examined a set of 142 distractors that were flagged by the system as potentially problematic because of low semantic plausibility, low lexical distinctiveness, or both. Revised distractors were generated using the revision tools integrated within the platform and were subsequently reanalyzed using the same diagnostic indicators.

Across the revised distractors, lexical distinctiveness increased on average by 0.05 points ($SD = 0.03$). To contextualize this change, a 0.05-point increase represents a meaningful shift relative to the reference threshold of 0.82 used elsewhere in this work to flag potentially redundant distractors: an increase of this magnitude is often sufficient to move a distractor from below to above the threshold, though the practical effect for any individual distractor depends on its starting score. Revisions generated through back-translation produced slightly larger improvements in lexical distinctiveness than revisions generated through abstractive paraphrasing (0.06 versus 0.04). At the same time, semantic similarity between the original and revised distractors remained high (mean cosine similarity = 0.88, $SD = 0.05$), indicating that revisions generally preserved the intended meaning of the distractor while increasing differentiation from competing options.

Overall, 67% of flagged distractors moved into acceptable diagnostic ranges following revision. These results suggest that the indicators can support practical item-development workflows by helping assessment developers identify distractors that may benefit from revision and by providing immediate feedback regarding the effects of those revisions.

Not all revisions resulted in improvement. Approximately one-third of flagged distractors remained outside the reference ranges after revision. These cases often reflected issues that extended beyond wording and required substantive changes to the distractor content itself. This finding further supports the role of MCQ-Diag as a review tool rather than an automated correction system. The diagnostic indicators can identify potential concerns and support revision, but expert judgment remains necessary when determining how distractors should ultimately be modified.

4.3 Corpus-Level Analysis

In addition to supporting review of individual items, MCQ-Diag can be used to examine distractor patterns across larger collections of questions. To illustrate this capability, we compared two MCQ corpora representing different item-generation approaches: a baseline set of MCQs generated using a general-purpose large language model and a set of MCQs generated using the ReQUESTA framework (Tian et al., 2026a,b).

Semantic and lexical diagnostics revealed systematic differences between the two corpora. ReQUESTA distractors exhibited higher average passage alignment than baseline distractors (0.36 versus 0.32) and higher lexical distinctiveness (0.91 versus 0.87). Using the diagnostic reference threshold of 0.82, 30% of baseline distractors were flagged as potentially redundant compared with 10% of ReQUESTA distractors.

These findings are not intended to establish the superiority of one generation approach over another. Rather, they illustrate how the diagnostic indicators can be used to identify and compare distractor characteristics across item collections. By exposing patterns that are not readily visible through item-level review alone, MCQ-Diag provides a mechanism for examining distractor quality at scale and for supporting quality assurance workflows in assessment environments.

5 Discussion

MCQ-Diag's goal is not to automate distractor evaluation but to make characteristics of distractor quality more visible and interpretable. Across expert validation, revision analyses, and corpus-level comparisons, semantic plausibility and lexical distinctiveness capture variation in distractor behavior that aligns with expert judgment, while semantic uniqueness, derived from the distribution of plausibility scores, offers a complementary descriptive layer that can inform item development and quality assurance.

From a measurement perspective, the tool addresses a practical gap in item review. Distractor-writing guidelines often call for plausibility, distinctiveness, and alignment with misconceptions, but these qualities are typically judged qualitatively (Haladyna et al., 2002; Rodriguez, 2005). MCQ-Diag translates these principles into inspectable, option-level indicators, helping reviewers move past holistic judgments toward more targeted ex-

amination of specific distractor characteristics.

The need for such tools may become increasingly important as AI-assisted item generation becomes more common (Yu et al., 2024; Feng et al., 2024; Tian et al., 2026a). As the volume of generated content increases, efficient review processes become increasingly important. MCQ-Diag was designed to support this stage of the assessment workflow by helping reviewers identify weak, redundant, or overly competitive distractors more efficiently.

An important design decision was to treat the diagnostic indicators as evidence rather than decisions. The observed correlations with expert ratings provide support for the relevance of the indicators, but they also demonstrate that distractor quality cannot be reduced to a single numerical value. Content relevance, instructional goals, domain-specific considerations, and intended misconceptions remain important components of item review that require human expertise. Accordingly, the system is intended to complement expert judgment rather than replace it.

Taken together, the findings suggest that interpretable semantic and lexical diagnostics can provide a useful foundation for supporting distractor review in MCQ assessments. By combining option-level inspection with corpus-level analytics, MCQ-Diag offers a practical approach for integrating measurement-informed review processes into contemporary assessment development workflows.

6 Conclusion

MCQ-Diag is a measurement-informed diagnostic tool designed to support the review of distractor quality in multiple-choice assessments. The system delivers interpretable, option-level evidence across the three diagnostic dimensions through an interactive environment that supports inspection, revision, and corpus-level analysis. Rather than assigning automated quality judgments, MCQ-Diag makes characteristics of distractor behavior visible and interpretable for assessment developers.

As AI-assisted assessment continues to expand the scale of item generation, efficient review processes become increasingly important. MCQ-Diag contributes to this need by translating established distractor-writing principles into inspectable indicators that complement expert judgment and support quality assurance in both human-authored and AI-generated assessments.

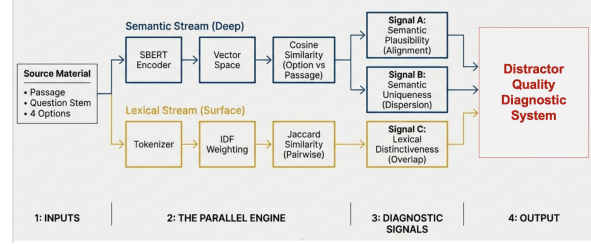


Figure 1: Underlying NLP Signals Pipeline.

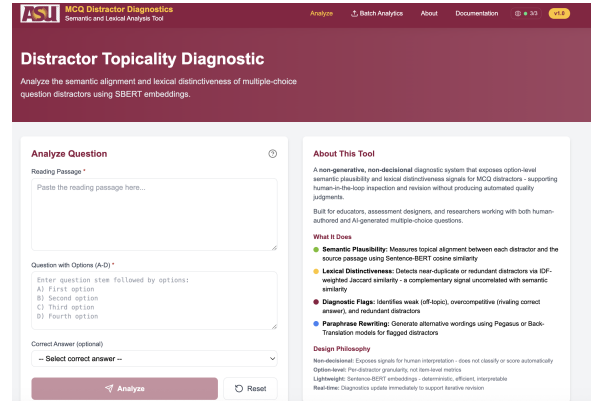


Figure 2: Interactive distractor diagnostics interface.

Limitations

Several limitations should be considered when interpreting the results. First, semantic plausibility is estimated through embedding-based alignment and therefore reflects topical relatedness rather than the full range of considerations experts may use when evaluating distractor quality. Second, semantic uniqueness is inferred from relationships among distractors rather than measured or validated directly; it should currently be treated as descriptive rather than validated, and future work should establish its reliability and validity against independent expert judgments. Third, the diagnostic reference bands were calibrated and evaluated on the same corpus of 200 expert-reviewed MCQs and should be interpreted as corpus-anchored guidelines rather than universal cut scores; assessing generalizability on an independent holdout set is a priority for future work. Fourth, the corpus-level comparison between baseline and ReQUESTA-generated items relied on the same diagnostic tool used throughout this work and did not include a neutral third-party comparison condition, so this comparison should be interpreted as illustrative rather than evaluative. Finally, MCQ-Diag is intended as a decision-support tool rather than an automated evaluation system: the diagnostic indicators provide evidence to support distractor review, but final judgments of item

quality remain the responsibility of assessment developers and subject-matter experts.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant 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

- Naveed Afzal and Ruslan Mitkov. 2014. Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing*, 18(7):1269–1281.
- Michelle Banawan, Linh Huynh, Katerina Christhlf, and Danielle S. McNamara. 2026. [Engineering a semantic topicality instrument for multiple-choice question quality control](#). In *Proceedings of the Learning Engineering Research Network Convening (LERN 2026): From Insights to Implementation, Learning Engineering in Action*. Learning Engineering Research Network.
- Wanyong Feng, Jaewook Lee, Hunter McNichols, Alexander Scarlatos, Digory Smith, Simon Woodhead, Nancy Ornelas, and Andrew Lan. 2024. [Exploring automated distractor generation for math multiple-choice questions via large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3067–3082, Mexico City, Mexico. Association for Computational Linguistics.
- Muriel Foulonneau and Valentin Grouès. 2012. Common vs. expert knowledge: making the semantic web an educational model. In *LiLe@ WWW*.
- Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.
- Yooseop Lee, Suin Kim, and Yohan Jo. 2025. [Generating plausible distractors for multiple-choice questions via student choice prediction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23669–23692, Vienna, Austria. Association for Computational Linguistics.
- Vatsal Raina, Adian Liusie, and Mark Gales. 2023. [Assessing distractors in multiple-choice tests](#). In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 12–22, Bali, Indonesia. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Michael C Rodriguez. 2005. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational measurement: issues and practice*, 24(2):3–13.
- Yu Tian, Shubham Chakraborty, Linh Huynh, Katerina Christhlf, Micah Watanabe, and Danielle S. McNamara. 2026a. [ReQUESTA: A hybrid agentic framework for generating cognitively diverse multiple-choice questions](#). In *Proceedings of the Learning Engineering Research Network Convening (LERN 2026): From Insights to Implementation, Learning Engineering in Action*. Learning Engineering Research Network.
- Yu Tian, Linh Huynh, Katerina Christhlf, Shubham Chakraborty, Micah Watanabe, Tracy Arner, and Danielle S. McNamara. 2026b. Cognitively diverse multiple-choice question generation: A hybrid multi-agent framework with large language models. *Electronics*, 15(6):1209.
- Han Cheng Yu, Yu An Shih, Kin Man Law, Kai Yu Hsieh, Yu Chen Cheng, Hsin Chih Ho, Zih An Lin, Wen-Chuan Hsu, and Yao-Chung Fan. 2024. [Enhancing distractor generation for multiple-choice questions with retrieval augmented pretraining and knowledge graph integration](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11019–11029, Bangkok, Thailand. Association for Computational Linguistics.