A Survey on Automated Distractor Evaluation in Multiple-Choice Tasks

Luca Benedetto, Shiva Taslimipoor, Paula Buttery
ALTA Institute, Dept. Computer Science and Technology, University of Cambridge
name.surname@cl.cam.ac.uk

Abstract

Multiple-Choice Tasks are one of the most common types of assessment item, due to their feature of being easy to automatically and objectively grade. A key component of Multiple-Choice Tasks are distractors – i.e., the wrong answer options – since *poor* distractors affect the overall quality of the item: e.g., if they are obviously wrong, they are never selected. Thus, previous research has focused extensively on techniques for automatically generating distractors, which can be especially helpful in settings where large pools of questions are desirable or needed. However, there is no agreement within the community about the techniques that are most suited to evaluate generated distractors, and the ones used in the literature are sometimes not aligned with how distractors perform in real exams. In this review paper, we perform a comprehensive study of the approaches which are used in the literature for evaluating generated distractors, propose a taxonomy to categorise them, discuss if and how they are aligned with distractors performance in exam settings, and what are the differences for different question types and educational domains.

1 Introduction

Multiple-Choice Tasks are a very popular form of students' assessment, due to their standardised format: they are easy to (automatically) grade and they remove subjectivity from the scoring process, and can thus be used to quickly and efficiently assess large numbers of students, in both high-stakes and low-stakes settings. A challenging step of curating high-quality Multiple-Choice Tasks – also referred to as Multiple-Choice Questions (MCQs) – is the generation of distractors, i.e., the incorrect options. Indeed, high-quality distractors must satisfy several properties (see §2.3), such as being incorrect but plausible, and consistent with the context but objectively wrong. The generation of high quality distractors has been shown to be challenging even for human experts (Shin et al., 2019), and to target this issue and generate large quantities of distractors (which are needed for large pools of questions) recent research has explored many approaches to automatically generate distractors, as discussed in two recent surveys (Awalurahman and Budi, 2024; Alhazmi et al., 2024). According to the assessment and testing literature (Nunnally and Bernstein, 1994), the most reliable approach to evaluate distractors is pretesting: new MCQs are shown to students in exam settings, and their response patterns are used to assess the distractors. Unfortunately, pretesting is unfeasible when automatically generating large numbers of distractors and undesirable in some settings, e.g., due to exam security concerns (Ha et al., 2019); thus, automatically generated distractors are most commonly evaluated with static approaches or with manual evaluation. However, the best techniques to automatically evaluate generated distractors are not commonly agreed across the community and the ones used in practice are rarely aligned with the performance of distractors in real exam settings. Hence, in this paper, i) we perform a comprehensive review of the approaches used in the literature for automated distractor evaluation, ii) we propose a new taxonomy to categorise them, iii) we discuss which ones are the most aligned with pedagogical theory and with the performance of distractors in real exam settings (also focusing on different educational domain and question types), and iv) provide some guidelines for future research.

2 Related Work

2.1 Distractor Generation

Two very recent surveys provide a good overview of approaches to distractor generation and the trends in the literature (Awalurahman and Budi, 2024; Alhazmi et al., 2024). Similarly to many other domains, distractor generation has seen a

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 55–69

rapid shift in recent years: the majority of approaches are now based on (large) language models, in contrast with research pre-transformers which was primarily based on traditional machine learning. We refer to the two survey papers mentioned above for a detailed description of the different techniques used in distractor generation.

2.2 Distractor Evaluation

The task of distractor evaluation is much less studied than distractor generation, even though it is becoming increasingly relevant: indeed, with modern generative models it is very easy to experiment with different prompts and generate a large set of distractors, and it is thus crucial to have ways to automatically and reliably evaluate them. Unfortunately, neither of the survey papers mentioned above focused sufficiently on the techniques and metrics which are used to automatically evaluate distractors. Considering fully automated metrics, Alhazmi et al. (2024) only mention ranking-based (Precision, Recall, F1-score, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), and Mean Average Precision (MAP)) and n-gram metrics (BLUE, ROUGE, and METEOR), while Awalurahman and Budi (2024) only mentions BLEU, ROUGE and METEOR. While these are all metrics that are indeed used in the literature, this list leaves out many others, which are very relevant and potentially more aligned with the performance of distractors in exam settings.

Few papers have distractor evaluation as main focus, proposing automated approaches for the task. Pho et al. (2015) work on distractors that are Named Entities in a knowledge graph, and propose an approach to evaluate them based on the syntactic and semantic relation between the distractors and the correct answer, and their relatedness in the graph. Ghanem and Fyshe (2023) generate "bad distractors" and train a model to estimate whether a given distractor is good or bad. Finally, Raina et al. (2023) propose an ensemble of three metrics which are meant to measure the incorrectness, plausibility, and diversity of distractors.

2.3 About *Good* Distractors

The educational literature is rich in recommendations and guidelines on how to create good distractors for MCQs. Ideally, these guidelines should be implemented within the models for automated distractor generation and evaluation, but our literature review suggests that in many cases the approaches used for evaluating automatically generated distractors in the NLP and AI for Education communities are somewhat disconnected from them. It is important to note that there are differences between educational domains – e.g., guidelines for language learning and mathematics cannot be exactly the same – but there are many common aspects. Distractors that are too easy fail to assess students' true understanding, while those that are too difficult or misleading can cause confusion and frustration; thus, distractors should be plausible, but objectively unacceptable (Yeung et al., 2019). Potentially, distractors should try to capture the common errors and misconceptions of students (Lee et al., 2016; Scarlatos et al., 2024), which enables targeted interventions. Also, distractors should be independent from one another, otherwise one or more could be excluded with logical reasoning, thus hindering the quality of the question. Distractors should be semantically and grammatically coherent with the context (Ghanem and Fyshe, 2023; Gao et al., 2019), and similar in length, style, and grammatical form to the correct answer (Pho et al., 2015). In language pedagogy literature, there is the recommendation that the target word and the distractors belong to the same word class (Heaton, 1988), ideally being "false synonyms" (Goodrich, 1977).

3 Taxonomy

Figure 1 presents the taxonomy we propose to categorise approaches from the previous literature. We group the different approaches based on the type of information that they use for evaluation. **Dynamic** approaches are based on learners' answers, and **static** approaches leverage only the textual information from the distractors (and potentially the correct answer, the question, and the reading passage). Dynamic approaches (§4), and specifically *Traditional Distractor Analysis*, can be seen as the *gold standard*, since they are based on students' responses and are an actual measurement of how distractors perform in exam settings; they can be further divided into approaches based on real students and the ones based on *responses from Question Answering (QA) models*. On the other hand, static approaches (§5) can be seen as an alternative to dynamic ones, as they can be used when it is unfeasible to obtain students' responses. Static approaches can be further divided into three groups: i) *comparative* approaches evaluate generated distractors by comparing them to some refer-

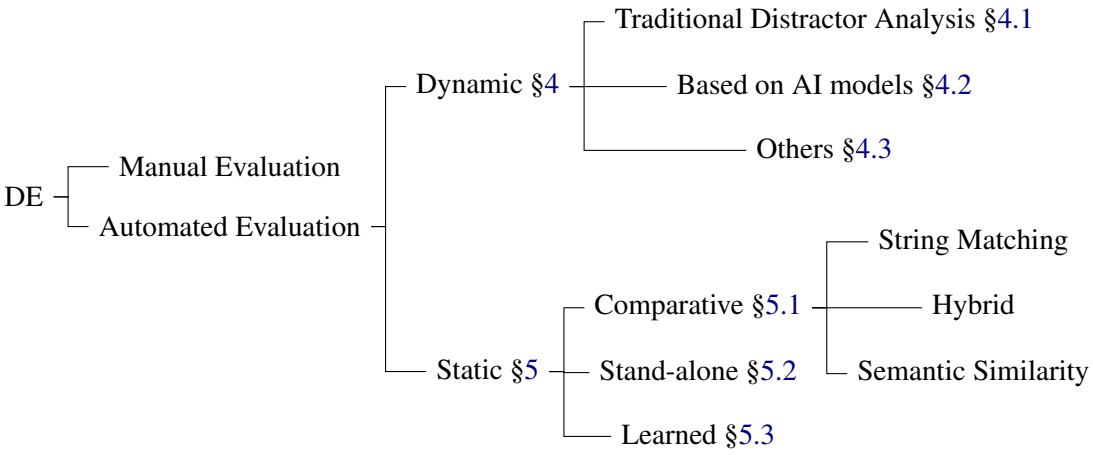

Figure 1: The taxonomy we propose to categorize the different approaches for Distractor Evaluation (DE).

ence ones, which are considered as gold standard, ii) *stand-alone* approaches consist in computing some measures of similarity between distractors and between distractors and the correct answer, and iii) *learned* approaches are machine learning models trained to predict the quality of generated distractors. From a practical point of view, there are notable similarities between distractor generation evaluation and difficulty estimation. In difficulty estimation, the gold standard is difficulty from pretesting – e.g., from Item Response Theory (Hambleton and Swaminathan, 2013) – but approaches have been proposed for difficulty estimation from text for when students' responses are not available (Benedetto et al., 2022).

Previous approaches are described in Sections 4 and 5, and Table 1 provides an overview of all the papers we discuss in this survey, grouped according to the proposed taxonomy. The table also shows the educational domain which each paper worked on, whether manual evaluation is used in addition to automated evaluation metrics, and whether distractors are evaluated individually or as a set.

4 Dynamic Approaches

Dynamic approaches to distractor analysis use students' responses to measure how well distractors perform in exams. They can be further divided into **traditional distractor analysis** §4.1 and **AI-based dynamic approaches** §4.2, depending on whether human or virtual students' responses are used.

Traditional distractor analysis is most commonly used in the Education and Assessment literature: it studies how distractors perform in real exams, observing the response patterns of human students, and can thus be considered the *optimal* approach to distractor evaluation. When it is unfeasible to use traditional distractor analysis due to cost, time constraints, or concerns about safety, AI-based dynamic approaches can be used. These are based on same techniques, but use the responses of QA models as a proxy for the responses from real students. Similar to difficulty estimation tasks, which are ideally performed via pretesting with real learners, research explored the possibility of using machine learning and AI to simulate it (Benedetto et al., 2022; AlKhuzaey et al., 2021). This includes the setting of virtual pretesting, which became more popular in recent years (Park et al., 2024; Uto et al., 2024; Benedetto et al., 2024).

Previous research also experimented with some approaches based on the responses of human learners but different from the ones used in traditional distractor analysis; they will be discussed in §4.3.

4.1 Traditional Distractor Analysis

Traditional distractor analysis is based on studying how often distractors are selected, and which is the (average) skill level of the learners selecting different distractors. Again, these metrics are based on how distractors perform in real exam settings, thus can be considered as the *optimal* ones.

Distractors that are never (or rarely) selected by students are *poor* distractors (Nunnally and Bernstein, 1994); the rule of thumb mentioned in several papers is that each distractor should be selected by at least 5% of the students (Haladyna and Downing, 1993), with the exception of very easy MCQs, which are correctly answered by more than 90% of the students (Gierl et al., 2017). Only three articles evaluate automatically generated distractors using the frequency with which participants select each

	Paper	Manual Eval.	Single	Set	Dynamic §4		Static Comparative §5.1			Static Stand-Alone §5.2	Static Learned §5.3
					Traditional	QA Models	Lexical	Hybrid	Semantic		
Language learning	(Pho et al., 2015)		X							X	
	(Lee et al., 2016)	X	X		X						
	(Gao et al., 2019)	X	X				X				
	(Zhou et al., 2019)	X	X				X				
	(Chung et al., 2020)		X	X		X	X				
	(Qiu et al., 2020)	X	X				X				
	(Maurya and Desarkar, 2020)	X	X				X	X	X		
	(Offerijns et al., 2020)	X	X	X		X	X				
	(Kalpakchi and Boye, 2021)	X	X	X			X			X	
	(Rodriguez-Torrealba et al., 2022)		X				X			X	
	(Xie et al., 2022)	X	X				X	X			
	(Chiang et al., 2022)		X				X				
	(Panda et al., 2022)	X	X				X				
	(Qu et al., 2023)	X	X	X			X		X	X	
	(Wang et al., 2023)	X	X				X				
	(Raina et al., 2023)		X	X						X	X
	(Yoshimi et al., 2023)		X	X	X		X				
	(Login, 2024)		X				X	X	X		
	(Zhou and Li, 2024)	X	X				X	X			
	(Qu et al., 2024)		X	X			X		X	X	X
	(Taslimipoor et al., 2024)	X	X	X			X			X	
	(Lin et al., 2024)		X				X	X			
	(Wang et al., 2025)	X	X	X			X			X	X
Maths	(McNichols et al., 2023)		X				X				X
	(McNichols et al., 2024)		X				X				
	(Feng et al., 2024)	X	X				X				X
	(Scarlatos et al., 2024)	X	X				X				
	(Fernandez et al., 2024)		X				X				
Science	(Aldabe and Maritxalar, 2010)		X	X	X						
	(Liang et al., 2018)		X				X				
	(Zhang and VanLehn, 2021)		X	X	X						
	(Dutulescu et al., 2024)	X	X				X				
Misc	(Mitkov and Ha, 2003)		X	X	X						
	(Lee et al., 2025)	X	X	X	X						X
Multi-domain	(Ren and Zhu, 2021)	X	X				X		X		
	(Bitew et al., 2022)	X	X				X				
	(Ghanem and Fyshe, 2023)	X	X								X
	(De-Fitero-Dominguez et al., 2024)		X				X			X	
	(Yu et al., 2024)	X	X				X				
	(Luo et al., 2024)	X	X	X	X		X				

Table 1: List of the papers discussed in this survey, grouped according to the proposed taxonomy.

of them: (Aldabe and Maritxalar, 2010; Zhang and VanLehn, 2021; Lee et al., 2016).

Another indication of distractor quality from the Education literature is the difference between the number of students selecting each distractor and the number of students selecting the correct answer: if a distractor is chosen more often than the correct answer, this is probably an indication of poor instructions or a misleading question (Nunnally and Bernstein, 1994). We did not find any paper evaluating generated distractors with this metric.

Lastly, since a good distractor is one that is selected by students who perform poorly and ignored by those who perform well (Gronlund, 1968), distractors that are selected by students that are (on average) of higher skill level than the students selecting the correct choice are poor distractors. We found only two papers using this factor to evaluate automatically generated distractors: Mitkov and Ha (2003) and Lee et al. (2025) divide students into a group of highly skilled students and a group of beginners, and label distractors that are selected by more students in the upper group than by students in the lower group as poor distractors.

4.2 AI-based Dynamic Approaches

Fundamentally, these use measurements similar to the ones from traditional distractor analysis, but based on the responses from QA models rather than human learners. Using machine learning models as a proxy of students, they should be validated accordingly. This is rarely done in the literature.

Chung et al. (2020) make the assumption that poor distractors will reduce the difficulty of the MCQ task for a QA model, thus use accuracy as an indicator of distractor quality, by comparing distractors generated with different models: the higher the accuracy, the worse the quality of the distractors. Similarly, Offerijns et al. (2020) study how the accuracy of a QA model changes when using manually-curated distractors rather than automatically generated ones: they observe that results are similar, thus claim that the generated distractors are on-par with the human-curated ones.

Guo et al. (2024) use the generated distractors to augment a dataset, which is then used to train a QA model. The quality of generated distractors is evaluated by measuring the QA accuracy on a separate test set: a better performance on the test set would indicate that the generated distractors were effective for training the model, and thus they are good distractors.

4.3 Others

Some papers use human responses for distractor evaluation, but in a setting different from traditional distractor analysis. Kalpakchi and Boye (2021) recruit participants on a crowd-sourcing platform and ask them to answer reading comprehension MCQs without providing them with reading passages. The authors claim that this approach can evaluate the *plausibility* of distractors by measuring how often they are selected. Luo et al. (2024) compare the response accuracy of three students on questions with distractors generated with different models, and claim that lower accuracy in responding to a question would indicate that there were better distractors. Yoshimi et al. (2023) evaluate distractors by measuring how the response accuracy of human annotators changes when using the original compared to generated distractors, aiming to make the accuracy as close as possible in the two settings. This is similar to the approach by Offerijns et al. (2020) but using humans rather than QA models.

5 Static Approaches

Static approaches evaluate distractors using only the content of the items, without considering learners' responses. Importantly, most of these approaches are not aligned *per se* with how distractors would perform in real exam settings, thus they should be validated (but often are not, in previous literature). They can be divided into *Comparative*, *Stand-alone*, and *Learned* approaches.

5.1 Comparative

Comparative approaches are based on a comparison between generated distractors and the reference ones available in the test dataset: this assumes that these reference distractors are of good quality and are the *only* distractors of good quality for a question. In other words, any generated distractor which is different from the reference ones is massively penalised. Both assumptions are somewhat problematic for distractor evaluation: experimental datasets often do not contain high-quality pretested questions (particularly the publicly available ones), and it might happen that other distractors are as effective, if not better, than the ones in the datasets. This disadvantage comes from the fact that most comparative approaches were not originally thought of for distractor evaluation, but rather for Machine Translation, and thus have fundamental issues when it comes to distractor eval-

uation (Rodriguez-Torrealba et al. (2022); Taslimipoor et al. (2024), inter alia). However, even with these major shortcomings, they are by far most commonly used approaches to evaluate new distractor generation models, due to their popularity and ease of implementation.

5.1.1 String Matching

String matching is the single most frequently used approach for distractor evaluation in the literature. Most papers used BLEU (Papineni et al., 2002) and/or ROUGE (Lin, 2004) to compare the generated distractors with reference ones in the experimental datasets (see Table 2 for the list of all papers). Other common metrics are Precision, Recall, F1-score, MRR, and NDCG (the list of papers is shown in Table 3). Notably, this distinction is also due to the fact that papers in the two tables mostly work on different types of questions: papers in 2 mainly work with reading comprehension questions with longer text answers, while papers in 3 mainly work with either cloze items or science tests with single word or named entity answers.

Paper	BLEU	ROUGE
(Gao et al., 2019)	X	X
(Zhou et al., 2019)	X	X
(Chung et al., 2020)	X	X
(Qiu et al., 2020)	X	X
(Maurya and Desarkar, 2020)	X	X
(Offerijns et al., 2020)	X	X
(Rodriguez-Torrealba et al., 2022)	X	X
(Xie et al., 2022)	X	X
(Qu et al., 2023)	X	X
(Login, 2024)	X	X
(Zhou and Li, 2024)	X	X
(Qu et al., 2024)	X	X
(De-Fitero-Dominguez et al., 2024)	X	X
(Luo et al., 2024)		X
(Lin et al., 2024)	X	X
(Taslimipoor et al., 2024)	X	
(Wang et al., 2025)	X	X

Table 2: List of papers using BLEU and/or ROUGE.

Other papers evaluated generated distractors using metrics based on string matching, but different from the metrics mentioned above. Liang et al. (2018) and Bitew et al. (2022) use Mean Average Precision, Luo et al. (2024) use Accuracy, and Kalpakchi and Boye (2021) measures the fraction of MCQs for which at least one generated distractor matches one of the reference ones.

McNichols et al. (2023); Feng et al. (2024); Fernandez et al. (2024), and McNichols et al. (2024) (all working on maths questions) define and use three *alignment-based* metrics: i) *partial match* evaluates whether at least one of the generated distractors matches one of the reference ones, ii) *exact match* evaluates whether all the generated distractors match the reference ones, and iii) *proportional match* measures the proportion of generated distractors which match the reference ones. In addition to these three metrics, Scarlatos et al. (2024) define *weighted proportional*, which is a reinterpretation of the proportional match: it re-weights each "match" in the proportional metric giving more importance to reference distractors which are most commonly selected by students. Notably, considering all the evaluation metrics based on string matching, this *weighted proportional* is the only one which explicitly takes into consideration how well distractors perform in real exams.

Paper	Precision	Recall	F1	MRR	NDCG
(Liang et al., 2018)	X	X		X	X
(Kalpakchi and Boye, 2021)		X			
(Ren and Zhu, 2021)	X		X	X	X
(Bitew et al., 2022)	X	X		X	
(Chiang et al., 2022)	X		X	X	X
(Panda et al., 2022)	X	X			
(Wang et al., 2023)	X	X	X		
(Yoshimi et al., 2023)			X		
(Dutulescu et al., 2024)	X		X	X	X
(Yu et al., 2024)	X	X	X	X	X

Table 3: List of papers using Precision, Recall, F1 score, Mean Reciprocal Rank, or NDCG for evaluation.

5.1.2 Semantic Similarity

Several articles evaluate generated distractors by measuring their semantic similarity to the reference ones, using diverse techniques for capturing the semantic meaning of distractors and their distance from the reference ones. While this is arguably more reliable than string matching, it still relies entirely on the quality of distractors in the experimental dataset. The most common approach is BERTScore (Zhang et al., 2019), which is used by Login (2024); Qu et al. (2024, 2023) to compute the similarity between generated distractors and the reference ones. Other embedding techniques are used in other articles: Ren and Zhu (2021) use Word2Vec (Mikolov et al., 2013), Maurya and Desarkar (2020) use BERT (Devlin et al., 2019) embeddings, and more recently Taslimipoor et al. (2024) apply Sentence-BERT (Reimers and Gurevych, 2019) to compute similarity. Notably, no one of these papers give weights to how different reference distractors perform in real exams.

5.1.3 Hybrid lexical-semantic

As a middle-ground between the purely lexical string matching approach described in §5.1 and the semantic embeddings from §5.1.2, some papers used METEOR (Banerjee and Lavie, 2005) for evaluating the similarity between generated and reference distractors. Specifically, it was used by Login (2024); Zhou and Li (2024); Maurya and Desarkar (2020); Xie et al. (2022); Lin et al. (2024). This has the same limitations as the approaches described above, as it relies entirely on the quality of the reference distractors, and implies that those are the only good distractors for a given question.

5.2 Stand-alone Approaches

Stand-alone approaches are all the evaluation techniques which are based on textual information only and do not rely on reference distractors. As such, they are meant to detect high-quality distractors even when these do not match some reference ones, and are not susceptible to low-quality distractors in the reference data. Most of these evaluation metrics are meant to capture the *plausibility* and *diversity* requirements of good distractors.

5.2.1 Estimating plausibility

Pho et al. (2015) focus on the relatedness between the distractors and the correct answer option, primarily working on questions whose responses are named entities. The semantic similarity is then measured looking at the distance between the named entities of each distractor and the correct answer option in a taxonomy of named entities.

Plausibility is modelled as the cosine similarity between each generated distractor and the correct answer option in (Rodriguez-Torrealba et al., 2022; De-Fitero-Dominguez et al., 2024). The authors state that higher similarity to the correct answer option means better distractors and use such approach for evaluating the distractors. Still, they do not study the correlation between the results obtained with their evaluation metric and an evaluation based on students' responses, thus this metric might reward distractors which are too close to the correct answer, and thus low quality.

A different take on plausibility is taken by Raina et al. (2023): they define *plausibility* as the sum of the confidence scores of a multiclass QA model for each of the distractors. This approach assumes that the confidence of a MCQA model is a good proxy of the confidence of real students, and evaluates this assumption by using a dataset which provides statistical information about how often distractors in the dataset are selected by real students (Mullooly et al., 2023); this is one of few works validating the metrics used for distractor evaluation.

5.2.2 Estimating diversity

More papers focused on studying the diversity of generated distractors, using Pairwise-BLEU, Distinct (Li et al., 2016), or other techniques. Pairwise-BLEU is used by Qu et al. (2023) and Wang et al. (2025), while Distinct is used by Qu et al. (2024) and Qu et al. (2023). Two different approaches are used by Raina et al. (2023), who use the BERT Equivalence Metric (BEM) (Bulian et al., 2022), and Taslimipoor et al. (2024), who use Sentence-BERT to measure the semantic similarity between different generated distractors. In all these papers the authors claim that high diversity is desirable, hence similarity between distractors should be low.

5.2.3 Others

Kalpakchi and Boye (2021) propose a set of evaluation metrics, including several stand-alone approaches different from all the approaches used by other papers. Most of them are filters which could actually be implemented within a DG model itself, and include measures such as i) the fraction of MCQs with two or more generated distractors which are equal, ii) the fraction of MCQs for which generated distractor match the correct answer, and others (we refer to the paper for the full list).

5.3 Learned Approaches

Learned evaluation metrics are machine learning models – with different architectures – specifically trained to evaluate the quality of generated distractors. Several approaches have been used in the literature, and they try to capture different characteristics that good distractors are expected to have. Notably, these approaches are on average the most recent of all the papers surveyed.

The first learned metric to evaluate generated distractors was proposed by Ghanem and Fyshe (2023), which is one of the few papers exclusively focusing on the evaluation of generated distractors. The proposed approach consists in automatically generating *bad distractors*, and training a model to estimate whether a distractor is good or bad (i.e. binary classification); the metric is validated with manual evaluation. A similar approach is used by Raina et al. (2023) and Qu et al. (2024). In the first paper, a model is trained to distinguish between

the correct answer option and the distractors, in a binary classification setting; the probability that such trained model assigns to each distractor (more specifically, $1 - P$) indicates *how incorrect* each distractor is. [1] In the latter, an Alberta model is trained to predict whether a given distractor is a correct answer to the corresponding question, and return a classification score in the range $[0, 100]$; the authors refer to this as *faithful score*.

Three papers focused on learned approaches to estimate the plausibility of generated distractors. In two of them (McNichols et al., 2023; Feng et al., 2024) the authors, who define plausibility as the likelihood of a distractor being selected by real students, compute it by training a BERT-based machine learning model on real students' responses to predict the fraction of students selecting each distractor. The trained model assigns a probability score to each distractor, and these scores are then combined in two ways: i) by summing the selection probability of all distractors, and ii) by computing the entropy among them (to make sure that all are selected with reasonable frequency by students). In the third (Lee et al., 2025), the authors train a pairwise ranker to select, given a pair of distractors, the more plausible. Ground truth plausibility is estimated from students' responses, thus this metric is aligned distractor performance in exam settings.

Finally, in one paper which performs distractor generation via reinforcement learning from preference feedback (Wang et al., 2025), the authors leverage the same reward model that was used in training during the reinforcement learning phase to then evaluate the generated distractors.

6 Discussion

6.1 Alignment with exam performance

Considering all the evaluation approaches described above, the only ones which are by definition aligned with how distractors perform in real exam settings are the techniques from traditional distractor analysis (§4.1), since they evaluate distractors based on the responses of real students. We argue that these approaches should be used whenever possible. Unfortunately, in most cases, that is not feasible, and some alternative approaches have to be used. In all these cases, it is important to validate the evaluation approach to ensure that they align with the exam performance of distractors, but this is rarely done in the literature. The main reason for this is that most of the publicly available datasets – e.g., RACE (Lai et al., 2017), SQuAD (Rajpurkar et al., 2016), or the MCQ dataset by Ren and Zhu (2021) – do not provide such information, thus it is impossible to properly validate the evaluation metrics on them and all evaluations are built upon weak foundations. One notable exception is the Cambridge MCQ Reading Dataset (Mullooly et al., 2023), which contains an indication of how often distractors are selected by students in real exam settings: the dataset contains both *good* and *bad* distractors, and can thus be used to validate different evaluation metrics. Similarly, private datasets, such as the Eedi dataset used by Scarlatos et al. (2024) and others, likely contain statistics about students' responses, and thus provide the information needed to validate the evaluation metrics (as it is done for the weighted proportional metric described in §5.1.1). However, they are inaccessible for the wider research community.

6.2 Evaluating individual distractors and distractor sets

The taxonomy proposed in §3 categorises evaluation metrics based on the information used for evaluating generated distractors. However, another relevant dimension to consider is whether evaluation metrics work on individual distractors or distractors as a set of options. Indeed, distractors should ideally be evaluated with both, since they capture different aspects in relation to designing a good question item. The number of papers that evaluate distractors individually is an overwhelming majority in the literature, and only few use metrics that consider distractors as a set, as shown in Table 1.

All the comparative approaches in §5.1 focus on evaluating individual distractors. While this is very relevant, as it can help detect distractors which are too close to or too far from the correct answer option, it is a suboptimal evaluation. Indeed, in real exam settings distractors are shown to students in a set of (usually) four items (one being correct), and distractor evaluation metrics should also consider the similarity and differences between the distractors – thus evaluating *sets* of distractors. Notably, even considering the papers which perform a manual evaluation of the distractors, these are evaluated individually (e.g., annotators are asked to classify each of them as *acceptable* or *not acceptable* (although out of the main scope of this survey paper, we include an analysis of manual evaluation in the

[1] The metric is validated using student response data from a publicly available dataset (Mullooly et al., 2023).

appendix §A). From our analysis, a total of 15 papers (out of the 40 doing automated evaluation) use automated metrics that evaluate distractors as a set rather than as individual items.

6.3 Educational domains and question types

In the context of distractor generation and evaluation for MCQs, question types and educational domains play a crucial role in designing effective evaluation metrics. These factors influence the characteristics of distractors and the criteria used to assess their quality. The subject or educational domain influences the complexity, language, and knowledge required for distractor evaluation. For instance, in science and mathematics, evaluation metrics should check for scientific validity or in language learning, like in reading comprehension questions, evaluation should assess linguistic similarity and conceptual relevance. These aspects of evaluation have not been investigated explicitly in the literature, however we can see that for example almost all papers experimenting with the RACE dataset for reading comprehension, evaluate distractors using metrics from machine translation (see Table 2) while most distractor generations in the domain of science (Liang et al., 2018; Ren and Zhu, 2021; Bitew et al., 2022; Dutulescu et al., 2024) or with Cloze-style questions (where answers and distractors are single words or named entities) (Chiang et al., 2022; Panda et al., 2022; Wang et al., 2023; Yoshimi et al., 2023; Yu et al., 2024) are mainly evaluated using ranking based statistical measures (see Table 3).

6.4 About manual evaluation

Although not discussed in this survey, since our focus is on automated metrics which could be used in an automated generation and evaluation pipeline, manual evaluation is still used by the majority of papers (see Appendix A), sometimes in addition to the automated metrics and in other cases as the single evaluation approach. Annotators are domain experts, or the authors themselves, or recruited from crowd-sourcing platform – thus leading to annotations of varying reliability.

7 Conclusions

In this survey paper we have performed a comprehensive study of the metrics and techniques which are used to automatically evaluate generated distractors in the context of Multiple-Choice Tasks, and have proposed a taxonomy to categorise them. We have seen that there is not a commonly agreed metric in the literature, and different authors and research groups tend to use different evaluation techniques. Most importantly, the metrics which are most commonly used in the literature (e.g., BLEU and ROUGE) are sub-optimal and arguably not aligned with how distractors actually perform in exams: indeed, they evaluate newly generated distractors by comparing them with some reference ones assuming that the references are i) of high quality and ii) the only distractors of high quality that can be created for the given question. Both assumptions are very strong, and not really supported by previous research, especially for publicly available datasets such as RACE (which is one of the most commonly used datasets).

Ideally, distractors should be evaluated with Traditional Distractor Analysis (i.e., with real learners) but, when this is not possible, the evaluation metrics used in its place should aim at being more aligned with how distractors perform in real exam settings and with the requirements that good distractors are expected to satisfy (according to vast literature from Education and Assessment), such as being consistent and coherent with the question and the correct option, and being plausible enough to *distract* learners. This highlights the need for validating the evaluation metrics which are used in distractor generation and evaluation settings and developing new, more aligned, ones. The development of such metrics should also take into consideration the differences between different educational domains, as the requirement might be different depending on the specific application scenario.

Limitations

When collecting the papers to review, we have performed several searches and used snow-balling to collect all the relevant publications which we could find. However, there is always a possibility that we might have missed some relevant research works. Also, we have highlighted the limitations of the current approaches to distractor evaluation, and this survey paper serves as motivation to focus more on the evaluation of distractors but, at this stage, we do not have an alternative approach to propose that might target these issues (yet).

Acknowledgments

This paper reports on research supported by Cambridge University Press & Assessment.

References

Itziar Aldabe and Montse Maritxalar. 2010. Automatic Distractor Generation for Domain Specific Texts. In *Advances in Natural Language Processing*, Lecture Notes in Computer Science, pages 27–38, Berlin, Heidelberg. Springer.

Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. 2024. Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14437–14458, Miami, Florida, USA. Association for Computational Linguistics.

Samah AlKhuzaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2021. A Systematic Review of Data-Driven Approaches to Item Difficulty Prediction. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, volume 12748, pages 29–41. Springer International Publishing, Cham.

Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating Questions and Multiple-Choice Answers using Semantic Analysis of Texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1125–1136, Osaka, Japan. The COLING 2016 Organizing Committee.

Halim Wildan Awalurahman and Indra Budi. 2024. Automatic distractor generation in multiple-choice questions: A systematic literature review. *PeerJ Computer Science*, 10:e2441.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. Association for Computational Linguistics.

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2022. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, page 3556538.

Semere Kiros Bitew, Amir Hadifar, Lucas Sterckx, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. Learning to reuse distractors to support multiple-choice question generation in education. *IEEE Transactions on Learning Technologies*, 17:375–390.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305. Association for Computational Linguistics.

Darryl J Chamberlain and Russell Jeter. 2020. Creating diagnostic assessments: Automated distractor generation with integrity. *Journal of Assessment in Higher Education*, 1(1):30–49.

Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. 2022. CDGP: Automatic Cloze Distractor Generation based on Pre-trained Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5835–5840, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A BERT-based Distractor Generation Scheme with Multi-tasking and Negative Answer Training Strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4390–4400, Online. Association for Computational Linguistics.

David De-Fitero-Dominguez, Eva Garcia-Lopez, Antonio Garcia-Cabot, Jesus-Angel Del-Hoyo-Gabaldon, and Antonio Moreno-Cediel. 2024. Distractor Generation through Text-to-Text Transformer Models. *IEEE Access*, pages 1–1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186. Association for Computational Linguistics.

Andreea Dutulescu, Stefan Ruseti, Denis Iorga, Mihai Dascalu, and Danielle S. McNamara. 2024. Beyond the Obvious Multi-choice Options: Introducing a Toolkit for Distractor Generation Enhanced with NLI Filtering. In *Artificial Intelligence in Education*, pages 242–250, Cham. Springer Nature Switzerland.

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.

Nigel Fernandez, Alexander Scarlatos, Simon Woodhead, and Andrew Lan. 2024. DiVERT: Distractor Generation with Variational Errors Represented as Text for Math Multiple-choice Questions. *Preprint*, arXiv:2406.19356.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, pages 6423–6430, Honolulu, Hawaii, USA. AAAI Press.

Bilal Ghanem and Alona Fyshe. 2023. DISTO: Evaluating Textual Distractors for Multi-Choice Questions using Negative Sampling based Approach. *Preprint*, arXiv:2304.04881.

Mark J. Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. 2017. Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, 87(6):1082–1116.

Hugo Gonçalo Oliveira, Igor Caetano, Renato Matos, and Hugo Amaro. 2023. Generating and ranking distractors for multiple-choice questions in portuguese. In *12th Symposium on Languages, Applications and Technologies (SLATE 2023)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Hubbard C Goodrich. 1977. Distractor efficiency in foreign language testing. *Tesol Quarterly*, pages 69–78.

Norman Edward Gronlund. 1968. *Constructing Achievement Tests*.

Qi Guo, Chinmay Kulkarni, Aniket Kittur, Jeffrey P Bigham, and Emma Brunskill. 2016. Questimator: Generating Knowledge Assessments for Arbitrary Topics. page 3726–3732.

Yingshuang Guo, Jianfei Zhang, Junjie Dong, Chen Li, Yuanxin Ouyang, and Wenge Rong. 2024. Optimization Strategies for Knowledge Graph Based Distractor Generation. In *Knowledge Science, Engineering and Management*, pages 189–200, Singapore. Springer Nature.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20, Florence, Italy. Association for Computational Linguistics.

Thomas M Haladyna and Steven M Downing. 1993. How many options is enough for a multiple-choice test item? *Educational and psychological measurement*, 53(4):999–1010.

Ronald K Hambleton and Hariharan Swaminathan. 2013. *Item response theory: Principles and applications*. Springer Science & Business Media.

John Brian Heaton. 1988. *Writing English language tests*. Longman.

Shu Jiang and John SY Lee. 2017. Distractor generation for Chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148. Association for Computational Linguistics.

Dmytro Kalpakchi and Johan Boye. 2021. BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 387–403, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Dmytro Kalpakchi and Johan Boye. 2024. Generation and Evaluation of Multiple-choice Reading Comprehension Questions for Swedish. *Northern European Journal of Language Technology*, 10(1).

Girish Kumar, Rafael E Banchs, and Luis Fernando D'Haro. 2015. Revup: Automatic gap-fill question generation from educational texts. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–161.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *Preprint*, arXiv:1704.04683.

John Lee, Donald Sturgeon, and Mengqi Luo. 2016. A CALL System for Learning Preposition Usage. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 984–993, Berlin, Germany. Association for Computational Linguistics.

Yooseop Lee, Suin Kim, and Yohan Jo. 2025. Generating Plausible Distractors for Multiple-Choice Questions via Student Choice Prediction. *Preprint*, arXiv:2501.13125.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C Lee Giles. 2018. Distractor Generation for Multiple Choice Questions Using Learning to Rank. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 284–290, New Orleans, Louisiana. Association for Computational Linguistics.

Chen Liang, Xiao Yang, Drew Wham, Bart Pursel, Rebecca Passonneaur, and C. Lee Giles. 2017. Distractor Generation with Generative Adversarial Nets for Automatically Creating Fill-in-the-blank Questions. In *Proceedings of the 9th Knowledge Capture Conference*, K-CAP '17, pages 1–4, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Association for Computational Linguistics.

Runfeng Lin, Dacheng Xu, Huijiang Wang, Zebiao Chen, Yating Wang, and Shouqiang Liu. 2024. DGRC: An Effective Fine-Tuning Framework for Distractor Generation in Chinese Multi-Choice Reading Comprehension. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 815–820.

Nikita Login. 2024. Wrong Answers Only: Distractor Generation for Russian Reading Comprehension Questions Using a Translated Dataset. *Journal of Language and Education*, 10(4):56–70.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-Exemplar: Enhancing Distractor Generation for Multimodal Educational Question Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. 2020. Learning to Distract: A Hierarchical Multi-Decoder Network for Automated Generation of Long Distractors for Multiple-Choice Questions for Reading Comprehension. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 1115–1124, New York, NY, USA. Association for Computing Machinery.

Hunter McNichols, Wanyong Feng, Jaewook Lee, Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2023. Exploring Automated Distractor and Feedback Generation for Math Multiple-choice Questions via In-context Learning. *Preprint*, arXiv:2308.03234.

Hunter McNichols, Wanyong Feng, Jaewook Lee, Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Automated Distractor and Feedback Generation for Math Multiple-choice Questions via In-context Learning. *Preprint*, arXiv:2308.03234.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Ruslan Mitkov and Le An Ha. 2003. Computer-Aided Generation of Multiple-Choice Tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 17–22.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark J. F. Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, and Shiva Taslimipoor. 2023. The Cambridge Multiple-Choice Questions Reading Dataset. Technical report, Cambridge University Press and Assessment.

J.C. Nunnally and I.H. Bernstein. 1994. *Psychometric Theory*.

Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. 2020. Better Distractions: Transformer-based Distractor Generation and Multiple Choice Question Filtering. *Preprint*, arXiv:2010.09598.

Subhadarshi Panda, Frank Palma Gomez, Michael Flor, and Alla Rozovskaya. 2022. Automatic Generation of Distractors for Fill-in-the-Blank Exercises with Round-Trip Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 391–401, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. Large Language Models are Students at Various Levels: Zero-shot Question Difficulty Estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177, Miami, Florida, USA. Association for Computational Linguistics.

Van-Minh Pho, Anne-Laure Ligozat, and Brigitte Grau. 2015. Distractor Quality Evaluation in Multiple Choice Questions. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, volume 9112, pages 377–386. Springer International Publishing, Cham.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2020. Automatic Distractor Generation for Multiple Choice Questions in Standard Tests. *arXiv:2011.13100 [cs]*.

Fanyi Qu, Hao Sun, and Yunfang Wu. 2024. Unsupervised Distractor Generation via Large Language Model Distilling and Counterfactual Contrastive Decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 827–838, Bangkok, Thailand. Association for Computational Linguistics.

Fanyi Qu, Che Wang, and Yunfang Wu. 2023. Accurate, Diverse and Multiple Distractor Generation with Mixture of Experts. In *Natural Language Processing and Chinese Computing*, Lecture Notes in Computer Science, pages 761–773, Cham. Springer Nature Switzerland.

Vatsal Raina, Adian Liusie, and M.J.F. Gales. 2023. *Assessing Distractors in Multiple-Choice Tests*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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*. Association for Computational Linguistics.

Siyu Ren and Kenny Q. Zhu. 2021. Knowledge-Driven Distractor Generation for Cloze-Style Multiple Choice Questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4339–4347.

Ricardo Rodriguez-Torrealba, Eva Garcia-Lopez, and Antonio Garcia-Cabot. 2022. End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. *Expert Systems with Applications*, 208:118258.

Alexander Scarlatos, Wanyong Feng, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Improving Automated Distractor Generation for Math Multiple-choice Questions with Overgenerate-and-rank. *Preprint*, arXiv:2405.05144.

Jinnie Shin, Qi Guo, and Mark J. Gierl. 2019. Multiple-Choice Item Distractor Development Using Topic Modeling Approaches. *Frontiers in Psychology*, 10:825.

Katherine Stasaski and Marti A Hearst. 2017. Multiple choice question generation utilizing an ontology. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312. Association for Computational Linguistics.

Shiva Taslimipoor, Luca Benedetto, Mariano Felice, and Paula Buttery. 2024. Distractor Generation Using Generative and Discriminative Capabilities of Transformer-based Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5052–5063, Torino, Italia. ELRA and ICCL.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2024. Question Difficulty Prediction Based on Virtual Test-Takers and Item Response Theory. In *Proceedings of the EvalLAC'24: Workshop on Automatic Evaluation of Learning and Assessment Content*.

Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. 2023. Distractor Generation based on Text2Text Language Models with Pseudo Kullback-Leibler Divergence Regulation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12477–12491, Toronto, Canada. Association for Computational Linguistics.

Ruofan Wang, Yuru Jiang, Yuyang Tao, Mengyuan Li, Xia Wang, and Shili Ge. 2025. High-Quality Distractors Generation for Human Exam Based on Reinforcement Learning from Preference Feedback. In *Natural Language Processing and Chinese Computing*, pages 94–106, Singapore. Springer Nature.

Jiayuan Xie, Ningxin Peng, Yi Cai, Tao Wang, and Qingbao Huang. 2022. Diverse Distractor Generation for Constructing High-Quality Multiple Choice Questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:280–291.

Chak Yan Yeung, John SY Lee, and Benjamin K Tsou. 2019. Difficulty-aware distractor generation for gap-fill items. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164. Australasian Language Technology Association.

Nana Yoshimi, Tomoyuki Kajiwara, Satoru Uchida, Yuki Arase, and Takashi Ninomiya. 2023. Distractor Generation for Fill-in-the-Blank Exercises by Question Type. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 276–281, Toronto, Canada. Association for Computational Linguistics.

Han Cheng Yu, Yu An Shih, Kin Man Law, KaiYu 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.

Lishan Zhang and Kurt VanLehn. 2021. Evaluation of auto-generated distractors in multiple choice questions from a semantic network. *Interactive Learning Environments*, 29(6):1019–1036.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Hao Zhou and Li Li. 2024. Qadg: Generating question–answer-distractors pairs for real examination. *Neural Computing and Applications*.

Xiaorui Zhou, Senlin Luo, and Yunfang Wu. 2019. Co-Attention Hierarchical Network: Generating Coherent Long Distractors for Reading Comprehension. *arXiv:1911.08648 [cs]*.

A On Manual Evaluation

Even though manual evaluation is not scalable to large amounts of distractors and cannot be used in a fully-automated content generation pipeline, it is still the most commonly used approach to evaluated distractors in distractor generation papers. From our analysis, a total of 23 papers out of 40 use manual evaluation in addition to automated evaluation; in addition to these, we also find 9 papers where the manual annotation is the only evaluation that is performed.

There are not commonly agreed guidelines on how to evaluate the distractors manually, and different papers follow different approaches and provide different labels, in some cases limiting the annotation to *good* and *bad* distractors, and in some other cases ranking on a Likert scale (e.g., from 1 to 5) some aspects of the distractors. In general, we observe that the annotators are either asked to provide an overall evaluation of the distractors (i.e., whether they are *good* distractors), or evaluate them according to the following aspects: plausibility (also referred to as distracting ability), fluency, coherence with the text (also referred to as validity), diversity (between the generated distractors), and being related to students' misconceptions. Notably, only two papers explicitly ask annotators to evaluate the diversity of the generated distractors – thus evaluating them as a set – and most of the papers perform an evaluation of individual distractors. Table 4 provides an overview of which of these aspects are considered in the different papers which perform manual evaluation of distractors.

Paper	Overall	Plausibility	Fluency	Misconception	Coherence	Diversity	Only Manual Eval
(Kumar et al., 2015)	X						X
(Guo et al., 2016)	X						X
(Lee et al., 2016)	X	X					
(Araki et al., 2016)		X			X		X
(Jiang and Lee, 2017)		X					X
(Liang et al., 2017)	X						X
(Stasaski and Hearst, 2017)	X						X
(Zhou et al., 2019)		X	X		X		
(Gao et al., 2019)		X					
(Offerijns et al., 2020)	X				X		
(Chamberlain and Jeter, 2020)				X			X
(Qiu et al., 2020)		X	X		X		
(Maurya and Desarkar, 2020)		X			X		
(Ren and Zhu, 2021)		X			X		
(Kalpakchi and Boye, 2021)	X						
(Xie et al., 2022)			X		X	X	
(Bitew et al., 2022)	X						
(Panda et al., 2022)	X						
(Ghanem and Fyshe, 2023)	X						
(Gonçalo Oliveira et al., 2023)	X						X
(Wang et al., 2023)		X			X		
(Qu et al., 2023)		X	X		X	X	
(Zhou and Li, 2024)		X	X		X		
(Feng et al., 2024)		X			X		
(Yu et al., 2024)	X	X			X		
(Dutulescu et al., 2024)	X						
(Scarlatos et al., 2024)	X	X					
(Kalpakchi and Boye, 2024)	X						X
(Luo et al., 2024)		X			X		
(Taslimipoor et al., 2024)	X						
(Wang et al., 2025)	X	X			X		
(Lee et al., 2025)			X		X		

Table 4: List of papers using **manual evaluation**, with an indication of which characteristics of distractors the annotators are asked to evaluate.