Dear AI4Ed Steering Committee:

Thank you for providing us with the opportunity to address reviewers' comments on our proposal *Current Evaluation Methods are a Bottleneck in Automatic Question Generation* submitted to the AI4ED workshop at AAAI. We have found the reviewers' comments and suggestions very helpful in improving the quality of our paper. We have considered each comment and suggestion carefully and made the required revisions to our paper. Our responses to the reviewers' comments and suggestions are outlined below.

Sincerely,

The authors

Reviewer #1:

This 4-page paper gives an overview of evaluation methods for automatic question generation (AQG). This covers human evaluation with experts, human evaluation through crowdsourcing, ablation studies, post-hoc analysis of learner responses, and automated metrics borrowed from machine translation. The authors claim that the bottleneck in AQG remains because no single existing evaluation method is entirely satisfactory — human evaluation is too slow, and automatic metrics are not sensitive enough. This is an interesting and useful review of existing methods. There's a typo in §2.1 "stables" -> "staples".

Thank you very much for your positive feedback and comments. We fixed the typo "stables" as "fundamental methods."

Reviewer #2:

I believe there could be more concrete discussions about the authors' feelings on coming up with metrics for AQG evaluation. The future directions seem vague to me. The authors could have gone through more recent literature for AQG evaluation and reported what is currently used with the advent of LLMs.

It would be good to discuss recent works that aim at human-like question generation (for example - "Improving Reading Comprehension Question Generation with Data Augmentation and Overgenerate-and-rank") and how their evaluation would differ from other methods.

In general, I believe that this is a good survey of evaluation methods for AQG systems. However, I am not very much convinced with the novelty/innovativeness of the paper.

Thanks for your suggestion regarding going through the recent literature for evaluation methods used with the advent of LLMs. We scanned recent literature using LLMs for question generation. The evaluation methods used in studies employing LLMs can be categorized using the taxonomy highlighted in the proposed paper (e.g., Elkins et al., 2023; Liang et al., 2023; Sarsa et al., 2022; Wang et al., 2022).

We also carefully read the paper suggested by the reviewer. In the paper, the authors have used the ROUGE-L method, which is a type of machine translation metric discussed in the paper. We have already indicated the limitations of the machine translation metrics in the proposal. We also cited the paper *Improving Reading Comprehension Question Generation with Data Augmentation and Overgenerate-and-*

rank in the proposal because it was a good example of studies using machine translation metrics for evaluating the quality of questions generated concerning a reference question.

Regarding the novelty and innovativeness of the paper, we argue that we intend to provide a comprehensive summary of the prominent evaluation methods used by automatic question-generation systems while underscoring their limitations. Studies have extensively focused on creating many, diverse, and human-like questions by using automatic question generation methods, yet a survey on the evaluation methods used for evaluating the quality of generated questions is missing. Evaluation methods are, in fact, one of the most fundamental aspects of automatic question generation because they allow researchers to evaluate the question quality as well as the strengths and weaknesses of the question generation system. This paper, to the best of our knowledge, is a first attempt to review the most prevalent methods used in question generation systems and highlight the limitations of current evaluation methods. We highlighted our contributions in Implications and Future Directions.