Quiz Generation Experiment

Hypotheses

As previously discussed, we want to demonstrate the following hypotheses:

1. **LLMs can generate different types of questions from a given context that teachers can use to create a quiz more efficiently and of comparable quality to a handwritten version.**
2. LLMs can generate different types of questions from a given context that teachers can use to create a quiz more efficiently and of higher quality than a handwritten version.

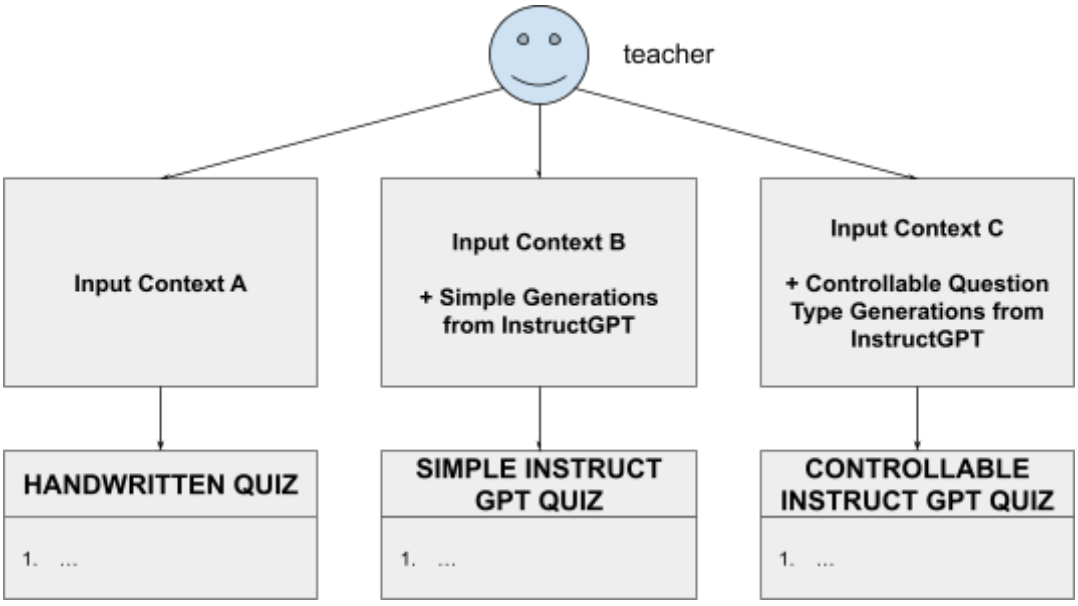
Experimental Setup

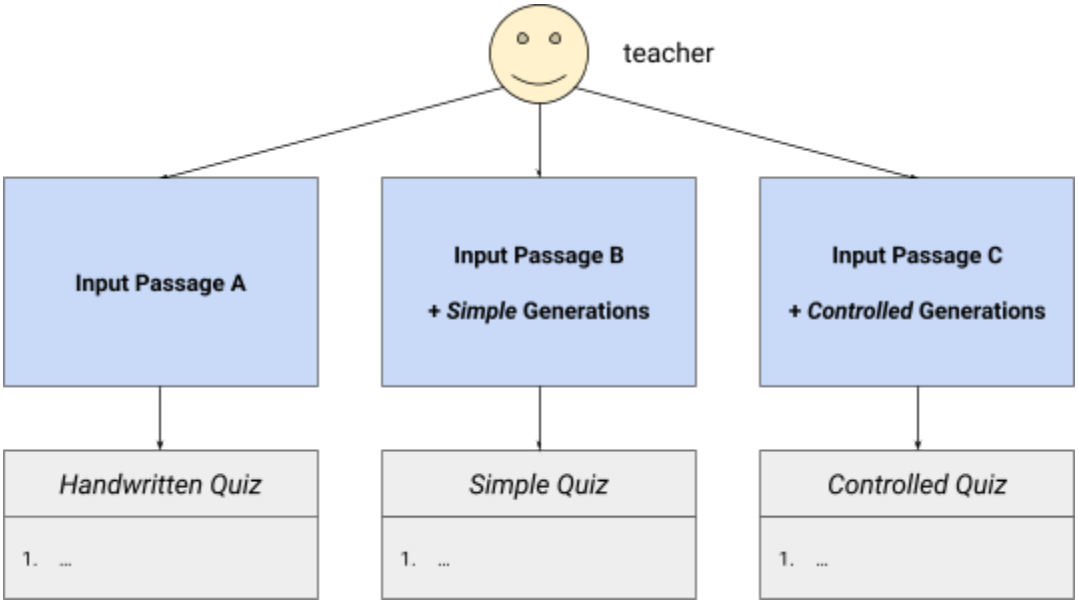
Similar work: Quiz Design Task: Helping Teachers Create Quizzes with Automated Question Generation by Laban et al 2022 → the difference here is the *controllable* generation and assessment of the quizzes as a whole, not just individual questions

Initial brainstorming for this experiment is here: Experiment 3 Brainstorm - V3 This includes a visual of what each quiz should look like, slides 15 to 24.

In brief, we want to compare 3 quiz types:

1. **Handwritten Quiz:** quiz written by teachers from scratch.
2. **Simple Generations Quiz:** quiz written by teachers who are given an unordered list of generations from InstructGPT (zero-shot and no control elements).
3. **Question Type Generations Quiz:** quiz written by teachers who are given a list of question types and corresponding generations from InstructGPT.





Question Generation Process

See https://arxiv.org/pdf/2304.06638.pdf → same generation process (model, prompt, etc.)

Key difference: I moved to *jointly* generating Bloom's taxonomy questions (i.e., generating them all at once with one prompt) in order to address the independence assumption seen in the paper above. Proposal to generate jointly is here: April 2023 - Ekaterina/Jackie/Iulian/Sabina

I investigated if generating Bloom's taxonomy questions in a different order affected the results, here: Generating in Bloom's Taxonomy → Answer is: the differences between level-ordering and random-ordering do not appear significant (slightly better when one uses does the typical level-ordering).

New generation examples (in domain): Generation Few Shot Examples

Question Type	Definition
Remembering	The question should ask students to retrieve from memory a fact, term, concept, etc..
Understanding	The question should ask students to demonstrate their understanding of material by describing, explaining, comparing, interpreting, etc.
Applying	The question should ask students to use the presented concepts to solve problems, or explain ideas in a different way.
Analyzing	The question should ask students to break material into parts, and/or show how different ideas relate to one another.
Evaluating	The question should ask students to give opinions on, make judgments about, or interpret meaning from material.
Creating	The question should ask students to combine material together in a different way than it was presented.

Quiz-Writing Annotator's Process

Each teacher makes a quiz of each type on 3 *different* context inputs. They must be different to avoid bias of them looking for their original questions in the generations.

1. STEP 1 - Consent, information and training.
   https://docs.google.com/forms/d/14zYX0ZFIKIerTfsxNnD99uvHyOSuQvZkpvcr1Xw7mA4/edit
   - I assess the results of their training with this rubric before moving on to the rest of this experiment (especially considering that this task is more difficult than our previous experiment).

Note that these next 3 steps will be in a random order (to reduce bias - e.g.). Note that they have already seen all types during annotator training.

2. Template - Handwritten Quiz
   - Note that here and below, we ask the participants to record their screen so that I might later extract the most information possible about their quiz writing process (i.e., which are copied, which are changed, which are handwritten, etc.).
3. Template - Simple Generations Quiz
   - **There will be 9 simple questions** as there are 9 question types used in step 4 (and we want to minimize the differences EXCEPT for the controllable nature of the generations)
4. Template - Question Type Generations Quiz

5. STEP 5 - Post Quiz/Self Evaluation - At the end, we get the quiz-writing participants to compare the quiz writing experience across the 3 types.
   https://docs.google.com/forms/d/1j4WZqqijHKo3OrO8RZTwor3VoAfzxf6BEOBvYj5vRaM/edit
   Pilot version (includes questions about the set up of the study):
   https://docs.google.com/forms/d/1BoOeDzt39sNPjyuiLBmsots0zrvkEGP3XBOMa1PRphc/edit

Note: I considered and drafted a few over versions (including excel, google forms, and Django) but decided that the above version is (a) most similar to the actual process a teacher might use, and (b) eliminates bias introduced by UI by using tools teachers are already familiar with/not adding design components to the teaching experience.

Quiz-Quality Annotator

Finally, after all quizzes are made, we get different teachers to analyze quiz quality with the metrics explained below. While ideally the controllable generations quizzes will be of higher quality, the main goal here is to ensure that none of the quizzes/annotators are poor quality, and should be ruled out,

Training:
https://docs.google.com/forms/d/1FelfGg_KhVB6Z0r3TrpzTImm_QHXNOXLAI0vl01k-x4/edit

Annotation: Pilot 2 - Quiz Quality Annotation

Outdated - V1 of Quality Annotation Scheme: Quiz Quality Annotation

Limitations of the Experimental Setup

1. VARIANCE: Due to the design of the hypothesis and experimental setup, this experiment has two key variables that are changing at the same time: the annotator and the input context.

I can see no way around this…

A. Keeping the same context for each annotator's 3 quizzes generates inherent variance because a teacher will be looking for their handwritten questions (or whichever quiz type they did first) on the generation quiz writing steps.
B. Keeping the same annotator for all of the contexts we include limits the assertion that teachers *generally* find this useful.

This might be avoided by calculating *annotator agreement* on a task. **HOWEVER**, this is an open ended task where we want the teachers to have freedom to create a quiz as they really would, making assessments of their 'inter-annotator agreement' kind of redundant.

The best solution I can see - and have designed into the experiment - is to have each annotator try each quiz type on a different context, but have each context seen for each quiz type (see the table in the Contexts section if this sentence explanation is not clear). The results can then be compared along two axes:

1. An individual annotator's 3 quizzes with the context as a potential variance source.
2. An individual context's 3 quizzes with the annotator as a potential variance source.

In an eventual written work, I plan to emphasize the **teacher's opinions** on the three quiz writing processes (see step 5 below) to show that they find the automatic generations to be useful - separate from their quizzes' quality. I will also explain the objective metrics and the quiz quality metrics - but mention these potential sources of variance and explain the reasoning behind how we chose to set up the experiment.

2. CONSTRAINED SETTING: Writing a quiz in this manner is a hugely constrained version of what a teacher would really have to do - it limits the scope to a single article and removes entirely their need to choose the material/combine it with other resources.

As such, any published work would NEED to have a limitations section outlining that this experiment only shows a preliminary result about how automatic generations can be useful in a classroom setting (because only a very constrained version of the setting is explored - and there is a huge missing element of these quizzes being placed in front of students).

Self-Pilot - Setup and Results

To validate the idea for this experiment, I conducted a self-pilot. In this, I followed steps 2 through 4 (I skipped training/consent) outlined in the experimental setup above.

Cumulated Results and Analysis: **Self-Pilot Results**

Complete Results

STEP 1: annotator training/consent → skipped for now
STEP 2: handwritten quiz Self-Pilot - STEP 2 - Handwritten Quiz video results
STEP 3: simple generations Self-Pilot - STEP 3 - Simple Generations Quiz video results
STEP 4: CTG Self-Pilot - STEP 4 - Question Type Generations Quiz video results
STEP 5: post quiz STEP 5 - Responses

Comments/Observations

- With the multiple tables of generated questions in step 4, I noticed that I lost my place on the page a few times (i.e. wasn't sure which table of generations I should be looking back at) → made some changes to make it clearer
- I think teachers should be able to write **5 to 10** questions → i.e., number_of_paragraphs to 2*number_of_paragraphs as this lets teachers ask about multiple key concepts in a paragraph (even more than 2 if they ask 3 about 1 paragraph and only 1 about the next) but keeps the final quiz at a manageable length.
- In this experiment I tried using three different sources for contexts - wikipedia, a textbook, and an article.
    - After discussing with Ekaterina, since we want to maintain style/complexity across all the articles included we decided to stick with wikipedia (similar format of presentation, style and level, open access and replicable).

Pilot 1 - Setup and Results

See progress of the pilot here: Pilot Tracker
Contexts and Generations used: Pilot Contexts and Generations

The pilot is running as explained in the experimental setup section. Ekaterina and I acted as the quiz quality annotators. The only difference is in the post-quiz where there are additional questions about the experimental setup.

Results:

- Annotator's training results: Pilot - Annotator Training
- Overall results (from videos and coverage metric): Pilot Results
- Quality annotation results: Pilot - Quiz Quality Annotation

The analysis of the results:
https://drive.google.com/file/d/1aiD3wAXE4My_74eTS1BhLwN2SqVRd7Dt/view?usp=sharing
Summary of the key takeaways/action points: April 2023 - Ekaterina/Jackie/Iulian/Sabina

Pilot 2 - Setup and Results

After the first pilot, we deemed the proposed changes to be sufficient to warrant another pilot experiment.

See progress of the pilot here: Pilot 2 Tracker
Contexts and Generations used: Pilot 2 - Contexts and Generations
In-domain few-shot examples: New ML Generation Few Shot Examples

Updated annotator training form:
https://docs.google.com/forms/d/1zEGLMmwctVmuvKtP5j9P04ZkSHsap2Bj2Advtd8yW0w/edit
Updated post-quiz form:
https://docs.google.com/forms/d/1IfMQCUHD82F_d68pP7J7n_7W_rPvwPy-cMO3K1E-Fs0/edit?usp=drive_web
Quiz Writing Tasks, raw sheets:
https://drive.google.com/drive/u/0/folders/1HpJTZr07sziYqeg6iU0D23_A8Kp2Z4WB?ths=true

Quiz Quality Annotation training:
https://docs.google.com/forms/d/1FeIfGg_KhVB6Z0r3TrpzTImm_QHXNOXLAI0vl01k-x4/edit
Quiz Quality Annotation: Pilot 2 - Quiz Quality Annotation

Training Results (both annotator types): Pilot 2 - Assessment of Training
Results: Pilot 2 - Quiz Writing Annotator Results Pilot 2 - Quiz Quality Annotation
Analysis: May 2023 - Ekaterina/Jackie/Sabina

Full Experiment - Setup and Results

Second pilot worked out the kinks, now we're ready for the full experiment!

See progress of the pilot here: Annotator Tracker
Contexts and Generations used: Contexts and Generations

Quiz Writing Annotator training: ML and BIO
Post-Quiz Assessment: ML and BIO

Quiz Quality Annotation training:

Quiz Quality Annotation:

Training Results (both annotator types): Quiz Writing Training Results

Quiz Writing Results:

Quiz Writing Annotator Results - ML Quiz Writing Annotator Results - BIO

Quiz Quality Results:

Analysis:

Draft messages for Quiz-Writing Annotators:

Case	**Message**
Asking to Join - Prev. Annotator	Hey **X**! I have been working on a follow up experiment to the one you annotated recently. Any interest in participating? The work involved would be (a) a consent and training module, (b) 3 quiz writing steps (should take no more than 30 minutes each), and (c) a post-quiz. The goal is to assess if the generated questions are useful in actual educational material, instead of the isolated way we were looking at them before. Overall, I project about 2.5 hours of work. This time the work will be split up and we will communicate in between each of the 5 steps (training, 3 x quiz writing, and post quiz) - which just means it will be smaller chunks of work, as opposed to one long google form. Again I can offer a compensation rate of 20$/hour. Please let me know if you are interested or have any questions! :))
First Step	Here is the first step of the pilot study: **INSERT LINK** It consists of a consent form, some demographics about you, and a training module. Please fill out the form and let me know when you're done. It shouldn't take more than 30 minutes - but please let me know how long it takes you so I can accurately pay you for your time. When you are done, I will take a look at your responses and if all is well then I will send the necessary info for the rest of the study.
Subsequent Steps	Okay! I just took a look at your responses - they're all looking good. :)) The next steps are essentially to perform each of the quiz writing tasks you saw in the training - (a) writing by hand/from scratch, (b) writing with generated questions, and (c) writing with generated questions of particular types. This will happen in a random order. Note that these will be in the format of a google doc instead of a form in order to facilitate reading/writing a longer quiz.

	Please do the following quiz writing tasks in order (n.b., you don't have to do all three at once, but each task should be done in one sitting): 1. **INSERT LINK** 2. **INSERT LINK** 3. **INSERT LINK** Don't forget to screen record yourself and record the time it takes so that I might fairly compensate you! After these quizzes, there is a short post-quiz where I can get your feedback on the quiz writing tasks: **INSERT LINK** Do you have any questions?
Payment	…

Resource Usage Projections

TL:DR; For the pilot, I think we will need 5 annotators, over a total of 12.5 hours, costing $250. I want to start here, and adjust the full experiment accordingly (but some initial projections can be seen below).

Pilot Projections

1. Number of annotators: 5 annotators total
    a. Number of Quiz-Writing Participants: **4** → **12 quizzes** ⇒ 4 quizzes/type
    b. Number of Quiz-Grading Participants: **1**
2. Time Required by the Participants
    a. Per writing-annotator:
        0.5 hour training + 1.5 hours quiz writing + 0.5 hour post quiz = 2.5 hours
    b. Per grading-annotator:
        0.5 hour training + 12 * 10 minutes per quiz = 2.5 hours
    c. Across all annotators = 2.5 * 4 + 2.5 * 3 = **12.5 hours**
3. Cost Projections ⇒ assuming same rate of 20$/hour, **total cost = $ 250**

Full Experiment Projections

1. Number of annotators/annotations
    a. Cohorts: 2 (two domains, as per the last experiment)
    b. Number of Quiz-Writing Participants per Cohort:
        12 → **36 quizzes** ⇒ 12 quizzes/type
    c. Number of Quiz-Grading Participants per Cohort: **3**
2. Time Required by the Participants
    a. Per writing-annotator:
        0.5 hour training + 1.5 hours quiz writing + 0.5 hour post quiz = 2.5 hours
    b. Per grading-annotator:
        0.5 hour training + 12 * 10 minutes per quiz = 2.5 hours
    c. Across all annotators = 2.5 * 12 + 2.5 * 4 = **40 hours**
3. Cost Projections ⇒ assuming same rate of 20$/hour,
    total cost per cohort = $ 800
    total cost = $ 1600 → ***upper bound***

Metrics

Objective Metrics

The following objective metrics can be measured from the annotation videos (screen recordings by the teachers of their quiz writing tasks).

1. Time taken to write the quiz
2. Number of questions edited by the teacher
3. Number of questions fully written by hand
4. % coverage of the context (depending on sample size, I can annotate by hand a map from questions to related context spans)
    - Pyramid method: https://aclanthology.org/N04-1019.pdf

Subjective Metrics

Post Quiz: https://docs.google.com/forms/d/1IfMQCUHD82F_d68pP7J7n_7W_rPvwPy-cMO3K1E-Fs0/edit

Metrics included in this :

- For each quiz type:
    - Rating of how efficient the quiz writing experience was
    - Rating of how frustrating the quiz writing experience was
    - Rating of if the generations were helpful (for the later two quizzes)
    - Free-form answer of pain points with any of the quiz writing experiences
- Which quiz was the most efficient, difficult, enjoyable, and frustrating
- Teacher's preferred quiz writing experience (and free-form answer as to why)

Original Idea for Quiz Quality Metrics

~~Metric~~	~~Scale~~	~~Definition~~
~~Coherence~~	~~3-point Scale~~	~~Does the ordering of questions flow well/make sense?~~ ~~N.b. it does not necessarily need to follow the ordering of the input context, but just needs to make sense.~~
~~Answerability~~	~~3-point Scale~~	~~Are the questions answerable from the input context?~~ ~~N.b. you do not need to be able to find a passage from the input that is an answer to the question. It is enough if a student could reasonably answer the question from the context (for example, applying logic explained in the passage to a new situation can be 'answerable').~~
~~Coverage~~	~~3-point Scale~~	~~Do the questions together cover the breadth of the input context?~~ ~~N.b. not every little detail of the input context needs to be covered, but major ideas should be.~~
~~Correctness~~	~~2-point~~	~~Do any of the questions contain logical or grammatical errors?~~

	~~Scale~~	~~N.b. any grammatical error counts, including punctuation and capitalization errors.~~
~~Quality~~	~~5-point Scale~~	~~Overall, in your opinion, is this a good quiz? Would you use it in a classroom setting (assuming you wanted to teach the provided context)?~~ ~~*Importantly,* this need not follow the ratings for the previous metrics. A quiz given all 'yes' ratings might lack originality, and be rated low quality. Alternatively, a quiz with 'somewhat' and 'no' ratings might be high quality, if its questions are diverse and engaging to a potential student.~~

~~Ask the annotators to explain why they chose this rating if any of these binary metrics are 'no', and also ask that they give an explanation about the quiz quality rating they give.~~

New Quiz Quality Annotation Scheme Proposal

Unfortunately, in all of the related work I found - in both QG and education - there was no full annotation scheme appropriate for assessing the quality of a quiz in the way we want to. 😕

Here are some of the different approaches I found to assessing a set of questions, often using automatic metrics, or actual students' success.

- Here, the annotators assess if the questions are human or machine written, and then ask them if they are high enough quality that they would use the quiz in their class.
- In this quiz design task, they only assess individual questions' acceptance into the quiz as a measure of 'acceptability', not the quality of the quiz as a whole (it is implicitly assumed because there are teachers writing the quizzes).
- Here, they ask Amazon Mechanical Turk workers to evaluate questions for clarity and grammatical correctness, as well as 'whether this question would help them remember or understand the meaning of the sentence' (from which the question was generated).
- Here, they use automatic metrics to assess individual question quality, and gather human judgements on fluency, relevancy, and answerability.
- Here is a human evaluation of automatic question generation for reading questions, where the authors assess a large collection of metrics: answerability, usefulness ('would-you-use-it'), grammar, etc.
- Assessment of quizzes/lessons/questions using students - a variety of tactics such as their *performance, engagement, success, etc.*:
    - Gamage et al. 2019: https://link.springer.com/article/10.1186/s40594-019-0181-4#Sec11
    - Qu et al. 2021: https://arxiv.org/pdf/2109.05179.pdf
    - Van Campenhout et al. 2021: https://ceur-ws.org/Vol-2895/paper06.pdf
    - Van Campenhout et al. 2022: https://doi.org/10.1007/978-3-031-11644-5_28
- This survey paper on AQG covers a lot of individual question metrics. In particular, figure 5 has a list of the most common question human assessment metrics.

Instead, I propose the following annotation scheme. It uses a collection of metrics to assess the quality of the whole quiz. Many of them are adapted from works evaluating individual questions.

Hopefully, being able to reference the existing works below, and explain why/how they were adapted to a whole quiz, will be enough of an argument to validate the metrics we're using to assess quiz quality.

Metrics per Question:

Metric Name	Possible Values	Definition/Explanation	Source(s)
Relevancy	**[yes, no]**	Measures whether the set of questions are semantically relevant to the input context.	Qu et al. 2019 (Edu. QG) with scale [1,3] Du et al. 2018 (Edu. QG) with scale [1,5] Steuer et al. 2021 (Edu. QG) as binary Marked as common metrics in this survey paper (no scale given)
Fluency	**[yes, no]**	Measures whether the set of questions are grammatical and clear (a.k.a. understandable).	Qu et al. 2019 (Edu. QG) with scale [1,3] Du et al. 2018 (Edu. QG) with scale [1,5] Mazidi & Nielsen 2014 (Edu. QG) as binary Marked as common metrics in this AQG survey paper (no scale given)
Answerability	**[yes, no]**	Measures whether the set of questions can be answered from the input context.	Qu et al. 2019 (Edu. QG) with scale [1,3] Du et al. 2018 (Edu. QG) with scale [1,5] Steuer et al. 2021 (Edu. QG) as binary Marked as common metrics in this AQG survey paper (no scale given)

Metrics to be rated for the quiz as a WHOLE:

Metric Name	Possible Values	Definition/Explanation	Source(s)
Structure	**[great structure,** **acceptable structure,** **poor structure]**	Measures whether the set of questions 'make sense' together → i.e., are they intuitively linked together with a natural/understandable flow (e.g., from easy to difficult, or from start to finish of the context). The possible values are as follows: **Great structure** = The quiz has an intuitive order of questions and the presentation of concepts is logical. You would not make any changes. **Acceptable structure** = The quiz has an understandable order of questions. You would only make minor changes (e.g. reorder a few questions, adding a single missing concept, etc.). **Poor structure** = The quiz has no	Chhun et al. 2022 use the same idea of coherence to evaluate textual generation storytelling (and reference at least 4 other papers that do the same) → We might argue that we can borrow this metric because the flow/narrative seen in this generation task is desirable for quiz writing, also. Similar to coherence in this AQG survey paper as a common metric for *conversational* QG. → Further argument for a set of questions

		discernable order of questions and the presentation of concepts is illogical. You would have to make a lot of changes, or even rewrite the whole quiz.	`containing some coherence/structure.
Redundancy	**[no redundancy,** **minor redundancy,** **major redundancy]**	Measures if there is redundancy/repetition within the quiz, e.g. two questions that ask for the same answer without any different perspective or thought process required from the student. The possible values are as follows: **No redundancy** = The quiz has no redundant or repetitive questions. You would not make any changes. **Minor redundancy** = The quiz has a small amount of redundant or repetitive questions. You would only make minor changes (e.g. remove or alter a single question). **Major redundancy** = The quiz has many redundant or repetitive questions. You would have to make a lot of changes, or even rewrite the whole quiz.	
Usefulness	**[useful,** **useful with minor changes,** **useful with major changes,** **not useful]**	We ask the annotators to "evaluate whether they would use the [quiz] in [an] assessment they create for their class". In other words, we would ask the annotators to answer: *Assuming you wanted to teach the provided context, would you use this quiz in your own classroom?* The possible answers to this question are as follows: **Not useful** = The core content of the question is not useful to teach context X at all. For example, the candidate might be off topic, have logical issues, simply not a useful question to teach context X, or be otherwise unacceptable. **Useful with major edits** = The core content of the quiz is useful, but the phrasing or presentation is not, and would require changes that take more than a minute. For example, the quiz might contain questions with confusing sentence structure that would need to be completely	Wang et al. 2022 (Edu. QG) as binary Bhat et al. 2022 (Edu. QG) as binary Steuer et al. 2021 (Edu. QG) as binary Essentially the usefulness metric from the previous work, so we can cite that also. (N.b. all of these are on *individual* questions so we would need to argue the jump to a full quiz.)

		re-written. **Useful with minor edits** = The core content of the quiz is useful, but the phrasing or presentation has some minor issues (e.g. grammatical errors, ordering issues, a redundant question to remove) that could be fixed in less than a minute. **Useful with no edits** = The quiz is useful as is, and can be used directly without making any changes. Note that the quiz does not necessarily need to be entirely answerable from the context in order to be considered useful.	
Notes		Anything they want to add!	

The hypothesis we want to answer is “LLMs can generate different types of questions from a given context that teachers can use to create a quiz more efficiently and of comparable quality to a handwritten version.”. ⇒ **Assuming we can demonstrate comparable *relevancy*, *fluency*, and *answerability* of the questions in each quiz type, as well as comparable *usefulness* across the quiz types, then I would argue that we show ‘comparable quality’.**

Contexts

For the self-pilot, see the 3 contexts collected here: Self-Pilot Contexts

For the pilots, we will need 4 contexts. Pilot Contexts and Generations
Pilot 2 - Contexts and Generations

	Handwritten Quiz	*Simple Generation Quiz*	*Question Type Generation Quiz*
Annotator 1	C1	C2	C3
Annotator 2	C2	C1	C4
Annotator 3	C3	C4	C1
Annotator 4	C4	C3	C2
Mean Time/# Question Metrics	-	-	-

Assuming we have 12 participants per cohort (see above) we will need 12 contexts per cohort in order to have each context be used to make a quiz of each type - organized like so:
Contexts and Generations

	Handwritten Quiz	*Simple Generation Quiz*	*Question Type Generation Quiz*
Annotator 1	C1	C2	C3
Annotator 2	C2	C1	C4
Annotator 3	C3	C4	C1
Annotator 4	C4	C3	C2
Annotator 5	C5	C6	C7
Annotator 6	C6	C5	C8
Annotator 7	C7	C8	C5
Annotator 8	C8	C7	C6
Annotator 9	C9	C10	C11
Annotator 10	C10	C9	C12
Annotator 11	C11	C12	C9
Annotator 12	C12	C11	C10

Constraints on the contexts themselves:

- Length (5 paragraphs)
- Topic (depends on the cohort)
- Difficulty (how to measure this?)
    - Ideas: TFIDF, length, …
- Source (educational resources) → Ekaterina and I discussed using Wikipedia articles because they all have a similar format of presentation, style and level. Also, they are open access and make the experiment more replicable.