A Guidline for Creating Assessments in Machine Learning Education

Kerem Cakici
Supervisor(s): Gosia Migut, Marcus Specht
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Even though machine learning field is growing rapidly, research on education of machine learning is scarce. In this paper a research about creating assessments in the machine learning's context is presented. The aim of the research is to answer how to design assessments that reliably show progress on a module in machine learning. Learning outcomes and Bloom's taxonomy are used to make the research reproducible, and draw conclusions. One of the main conclusions drawn in this paper is that verbs that are used in learning outcomes can also be used to find the appropriate question type (e.g. open ended, multiple-choice) to assess that learning outcome. Additionally, this paper concludes there is no strict procedure of creating assessment questions. Therefore, a guideline is created by the researcher and presented in the paper. Lastly, four questions are created using this guideline and evaluated with interviews with three machine learning professors.

1 Introduction

Machine learning increased its place in business life over the years, and as a consequence now there is a high demand for people who know machine learning techniques and can use them efficiently [14]. This is why it is essential to do research on the education of machine learning.

This research aims to enhance the quality of assessments in machine learning education by creating a guideline for designing assessments and evaluating them to enhance their reliability and validity. In addition, in this research, it is claimed that verbs used in learning outcomes can play an important factor in choosing assessment methodologies. To test this claim four sample questions were created from the module "Non-parametric density estimation" in the machine learning course from the bachelor of computer science and engineering program at TU Delft. These questions are evaluated by structured interviews with four machine learning professors from Delft University of Technology. To make this study as systematic and reproducible as possible, the study is using learning outcomes, and Bloom's taxonomy to classify the learning outcomes.

Bloom's taxonomy has been chosen because it is widely known, and therefore this makes this research paper easier to compare with other papers [5]. There are three main domains in this taxonomy: cognitive, affective, and psychomotor [10]. Bloom thought that it would be best if each field has its own taxonomy as some levels or domains in this taxonomy may not apply to each field [5]. However, creating a taxonomy for the machine learning field is out of the scope of this research due to time constraints, and is only discussed in the Future Work section.

Learning outcomes have been used in the present research in order to follow a "student-centered" approach and the international trend in education [6]. The learning outcomes are only focused on the cognitive domain. Because in machine learning context there are no psychomotor skills, and even though there are affective skills, for simplicity reasons these are out of the scope of this research. As it can be seen in figure 1, cognitive domain is composed of 6 different levels, which build upon each other [6]. However, the cognitive level of the learning outcomes from the undergraduate machine learning program of TU Delft, reaches up to the analysis level. Therefore creation and evaluation levels are not included in this research.

It is critical to use criteria while writing learning outcomes because if learning outcomes do not satisfy some qualities assessing them or even teaching them may become a huge

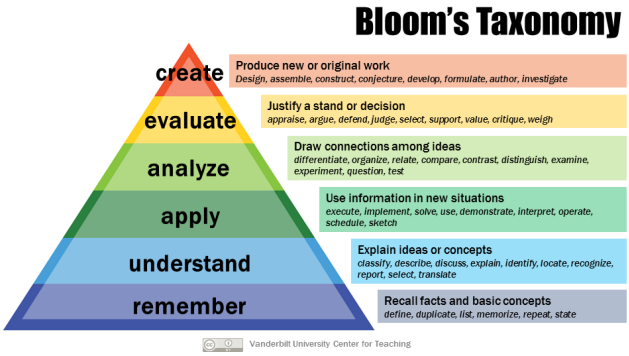


Figure 1: Cognitive levels in Bloom's revised taxonomy [7].

challenge. There are many papers on writing high-quality learning outcomes. For this paper, the practical guide for writing using learning outcomes from Declan Kennedy, Áine Hyland, and Norma Ryan is used [6]. This approach has been selected since it guarantees that the learning outcomes are realistic so there is no information overload, and quantifiable in order to allow reliable assessments.

Unfortunately, the literature for designing assessments in the machine learning context is scant, even though there are few pieces of research on constructing assessments in computer science [11, 13, 8]. The rest of the literature referenced in this paper is mainly about learning outcomes and assessments in general instead of in the machine learning context.

The research question is: How to design assessments that reliably show progress on a module in machine learning for both students and teachers using Bloom's taxonomy and learning outcomes? This is a reasonable research question since, as mentioned earlier in this section, the demand for people who can use machine learning appropriately rapidly increases. In addition, the literature on teaching and especially assessing machine learning knowledge is scarce. Therefore, it is important to experiment and research on creating assessments that reliably test students' progress in this context. The reliability of an assessment is defined in section 4.

The **Sub-Questions** of the research question are:

- What is a comprehensive set of learning outcomes of the module?
- What are the corresponding cognitive categories from Bloom's taxonomy of the learning outcomes?
- How can learning outcomes translate into sample assessment questions?
- How does Bloom's taxonomy help to find the appropriate way to assess a learning outcome?
- What are the ways of evaluating assessment questions?

2 Methodology

In this section, the methodology of this research is explained. The main aim is to make this methodology as systematic and repeatable as possible. Therefore, this section can be used

as a procedure to create assessments that reliably show progress in the machine learning context in the future.

The "Non-parametric density estimation" module has been chosen to be used in this study as the amount of content of the module is appropriate when the time limitations of this research are considered. First, problems are identified in the initial learning outcomes from the mentioned module using Bloom's taxonomy and the practical guide from Kennedy [6]. Even though using this guide makes it easier to identify problems in the learning outcomes, it does not benefit from pointing out missing ones. Therefore, to identify missing ones, exams, and course material is investigated.

After the problems are identified with the initial learning outcomes, in this step they are improved, and the missing ones are added. To test if they cover all the course material needed, each question about non-parametric density estimation from all the past exams within 3 years is matched with at least one learning outcome. Afterward, these revised learning outcomes were evaluated by an interview with a machine learning course staff from the Delft University of Technology. This evaluation is used to improve their quality further and finalize them.

Subsequently, finalized version of each learning outcome is classified to their corresponding cognitive level in Bloom's taxonomy. This step is critical for creating questions to assess these learning outcomes, as the guideline suggested in this research uses Bloom's taxonomy both for creating and evaluating assessment questions.

After the classification, assessment methodologies were determined, and sample questions were created. These sample questions were evaluated by interviews with machine learning course staff from the Delft University of Technology to demonstrate how this evaluation strategy explained in section 5 can be applied.

3 Adapting Learning outcomes

In this section, a case study is presented on identifying problems in learning outcomes and revising them accordingly. The case study uses non-parametric density estimation module from the machine learning course for undergraduate students from the Delft University of Technology.

Initial learning outcomes have been explicitly stated in the course material, and they are as follows:

1. Explain The Difference Between Parametric And non-parametric density estimation.
2. Explain Parzen, k-Nearest Neighbour, and Naïve Bayes density estimation and classification in detail.
3. Explain the advantages and disadvantages of those methods.
4. Implement k-nearest neighbor classifier in Python.

If the practical guide for learning outcomes by Kennedy is used, one of the problems with these learning outcomes can already be identified [6]: the second learning outcome is not specific as it is not clear for the students what "in detail" covers. Therefore, this learning outcome needs to be more explicit and if needed it should be divided.

To find other problems with these learning outcomes, the course material and previous exams from the last three years have been investigated. To make this in a systematic way,

this research matched all the information in each slide from the course material with at least one of the learning outcomes. Again for the exams, each question related to this module has been matched with the learning outcomes. If the matching fails that is noted as a missing learning outcome. In addition to this, the information on the lecture slides that have been matched with the second learning outcome has also been noted as they are useful to make the learning outcome satisfy specific criteria.

The identified problems are as follows:

1. Second learning outcome is too broad.
2. It is not obvious that k-nearest neighbors contain distance metrics.
3. The window function (kernel) should be separated from Parzen.
4. There is a missing learning outcome about identifying these methods (Parzen, k-nearest neighbor, and Naïve Bayes) as parametric or non-parametric.
5. There is a missing learning outcome in identifying these methods (Parzen, k-Nearest Neighbour, and Naïve Bayes) as linear or non-linear.
6. There is a missing learning outcome about the effect of the parameter "k" in the k-nearest neighbor algorithm.
7. There are missing outcomes about implementing kernel and Parzen in Python.

To fix all these identified problems, there is a need of writing new learning outcomes as there are many missing. To write them, this research mainly uses the Practical Guide from Kennedy et al. [6]. This well-written guide explains what are learning outcomes, and gives a description of how to write them. In this research, the definition from this guide is used.

The practical guide for writing and using learning outcomes mentioned above also suggests that specific verbs match with specific categories in Bloom's taxonomy [6]. For example, the "Compute" verb is associated with the Application layer in Bloom's taxonomy [6]. Associating verbs with levels is important, as it makes it easier to identify the level of a learning outcome, and adds cohesion to each level in this taxonomy. Additionally, this research suggests these verbs can be used to choose the assessment methodology.

Using above mentioned guide, and a case study for assessing learning outcomes by Diab Abuaiadah et al. the initial learning outcomes have been revised as follows [11]:

1. Explain the difference between parametric and non-parametric density estimation. **(Analysis)**
2. Explain Parzen density estimation and the purpose of window functions in detail. **(Understanding)**
3. Know which parameter needs to be optimized in Parzen density estimation. **(Remembering)**
4. Find Parzen probability density function estimates at a given point using (Gaussian, box, tri, triweight) window function. **(Application)**
5. Explain k-nearest neighbors, the influence of parameter k, how to optimize parameter k, and how to break ties in detail. **(Understanding)**

6. Compute Euclidean, Manhattan, and Hamming distance. **(Application)**
7. Apply k-nearest neighbors with a specific k parameter, and with one of the above distance metrics. **(Application)**
8. Implement the k-nearest neighbor classifier in Python. **(Application)**
9. Explain Naïve Bayes, and the effects of choosing Gaussian or Parzen as model per feature in detail. **(Understanding)**
10. Explain the advantages and disadvantages of the above-mentioned methods. **(Analysis)**

The "Implement" verb from the 8th learning outcome cannot be found in the practical guide used [6]. Although, in the machine learning context the "Implement" verb is necessary to have in learning outcomes as other verbs do not fit for implementing or coding of the algorithms, which is in the course material. Therefore, in this research "implement" verb is also going to be used, and it is associated with the application level in Bloom's taxonomy.

To evaluate the revised learning outcomes, feedback is asked from one of the course professors. A focus group that consists of professors, and students would be a better fit as it combines two main stakeholders' perspectives, allowing them to interact with each other's ideas. Although, due to time constraints a focus group was not feasible to be put into practice. More information on focus groups for evaluating learning outcomes can be found in the Future Work section.

The received feedback suggests that there is a need for another learning outcome about defining the characteristics above-mentioned methods. Therefore, the eleventh learning outcome has been added as follows:

11. Identify the characteristics of the above-mentioned methods (i.e. linear / non-linear, parametric / non-parametric, etc.). **(Remembering)**

4 Creation of the Assessment Questions

Creating assessment questions is a difficult task, and unfortunately, there are no exact steps that work in every case. Therefore, in this section first, it is explained what needs to be considered while creating assessments. Subsequently, it is discussed how verbs that are used in the learning outcomes affect choosing the assessment methodology. Lastly, the sample assessment questions created taking into consideration this section is presented, which are evaluated in section 5.

To be able to thoroughly discuss what needs to be considered while creating assessment questions, first "reliability" and "validity" on assignments need to be defined. In this research, for both terms, the definitions from The Student Assessment Handbook by Dunn are going to be used [3]. Briefly, if an assessment question is reliable, it should produce the same result in the same conditions consistently, and these results should be matching with the "real" level of the test taker [3]. Although, for an assessment question to be valid, it needs to measure what needs to be measured [3]. For example for this research valid assessment questions are the ones that align with a specific learning outcome, and show the level of a student for that learning outcome.

To make reliable and valid assessment questions, there are a few things to consider while creating an assessment. Although considering these things is not a solid procedure, therefore,

the most essential part of assessment creation is evaluating the questions created, which is explained in section 5.

In the "constructive-alignment" model, the learning outcomes and the assessments need to be closely aligned with each other [3]. To do this, the first thing to do before creating a question is to choose which learning outcome to assess. If the learning outcomes are well-made then there should not be any issues with choosing. Although, if there is something to assess that does not match with any learning outcomes, then it means that either the question does not fit the course material or learning outcomes are not well-made.

After, choosing what to assess, the second step is to classify the learning outcome in Bloom's taxonomy. For example, if the chosen learning outcome is the sixth one from section 3, it is classified as application level. This classification is essential as the assessment created needs to be able to measure if the student's progress on this outcome reaches this level. For example, "What is the formula for the hamming distance?" only reaches the first level in the cognitive domain while the application level is the third level. Therefore, it is crucial for the validity to take in consideration the cognitive level of the learning outcome while writing a question to assess it.

The Achievability of the questions is also significant for the reliability of the whole assessment, whether it is summative or formative. For example, assessments need to be able to finish in the given time considering the student's knowledge. Otherwise, the assessment will remarkably fail to assess the student's knowledge level. Another example can be the complexity of the question. These factors should be taken into consideration by the question's author. Although only considering is not sufficient, therefore in section 5, is going to be discussed how to evaluate the achievability of a question created.

For summative assessments, it is not generally feasible to evaluate all learning outcomes. Therefore, it is important to choose and prioritize them. Since the focus of this research is on formative assignments, prioritizing between learning outcomes is out of scope.

Lastly, the most momentous thing to consider is which question type to use mostly due to feasibility [12]. For example, while multiple-choice questions are very feasible, in most cases it is very difficult to assess application-level learning outcomes [5]. Therefore, one of the aims of the research is to find a systematic way for choosing the most suitable question type. First, matching cognitive levels in Bloom's taxonomy with question types has been tried. However, this led to some issues. The main problem was that all cognitive levels covered in this research theoretically can be tested with most types of questions. Various studies explain that multiple-choice questions can test levels that do not require creativity [2]. For example, the sixth learning outcome which is, to compute Euclidean, Manhattan, and Hamming distance, from section 3 is at the application level, and it can be tested with this multiple choice question:

What is the euclidean distance between (2, 5) and (-4, 5)?

A. 4

B. 6

C. 7

D. 8

E. 9

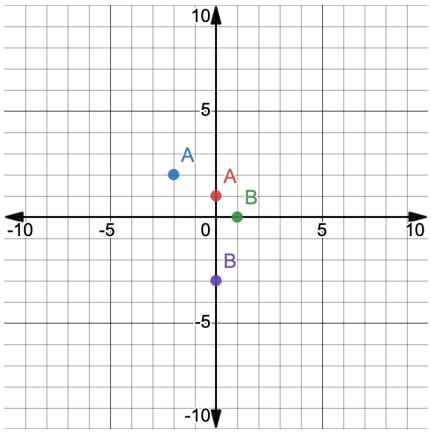
Figure 2: A multiple-choice question for an application-level learning outcome.

On the contrary, the eighth learning outcome which is to implement the k-nearest neighbor classifier in Python, cannot be assessed with a multiple-choice question. Even though these two learning outcomes are at the application level, the eighth learning outcome requires an open-ended question, possibly an implementation question. Therefore, in this research, it is concluded that it is not practical to match cognitive levels to question types. This being the case, a more pragmatic approach has been taken, which is matching the verbs of learning outcomes with question types. This approach enlarges the number of possible inputs significantly. Therefore, in the current research, this approach is tested on a small scale.

While open-ended questions are generally the most reliable ones to test all cognitive levels, in a university with lots of students, this question type is not always feasible as the grading process would necessitate additional resources. Three machine learning professors interviewed approved that feasibility prevents them from using all types of questions. This is why it is vital to choose more feasible question types like multiple-choice for the learning outcomes when it is possible. Therefore, as the example above, while an application-level learning outcome can be tested in a multiple-choice question, an open-ended question should not be used. Because the resources to grade that question could be used for another question that assesses a learning outcome that cannot be assessed more feasibly.

To test if verbs used in learning outcomes are a better indicator for choosing the question type than the cognitive levels, there are four questions created. The first two questions have different question types as the learning outcomes they assess use two different verbs, although they are both at the application level. For the last two questions that have been created, their corresponding learning outcomes use the same verb but are on different cognitive levels. More specifically, one of the learning outcomes is on the understanding level while the other is on the analysis level. Moreover, these questions have been specially designed to have the same question type because the learning outcome they asses use the same verb. Even though verbs are correlated with the cognitive levels, some of the different levels in Bloom's taxonomy may share verbs as can be seen with the tenth and fifth learning outcomes [6]. Both of them use the verb "explain" while the former one is in the analysis layer, and the latter one is in the understanding (comprehension) layer.

The four questions with the learning outcomes they assess is presented below:



1. **(Assesses learning outcome no. 6)** Mark wants to choose which distance metric to use for his k-nearest neighbors algorithm with k = 2 . To do this he first decides

to plot a part of the data as the data is two-dimensional. He writes the labels of the points next to them. After he plots a partition of data, he wants to choose the distance metric that will classify a point in the origin as B. Which distance metrics should he choose?

A. Manhattan Distance

B. Euclidean Distance

C. Either of the metrics classifies a point in the origin as A

D. Either of the metrics classifies a point in the origin as B

2. **(Assesses learning outcome no. 8)** Implement the following methods for k-nearest neighbors with k = 1:

```
# Points are stored in a list of tuples: For example x=1 y=4 label=B as (1, 4, 'B')
# This method should return the classified label for the new point in k nearest neighbor algorithm with k=1
def classify_using_closest_point(points: list, new_point: (int, int)):
    # implement here
    return

# This method should return euclidean distance between two two-dimensional points
def calculate_euclidian(x1, y1, x2, y2):
    # implement here
    return
```

3. **(Assesses learning outcome no. 10)** Can you explain one advantage and one disadvantage k-nearest neighbor algorithm?

4. **(Assesses learning outcome no. 5)** Can you explain two different ways to break ties in k-nearest neighbor algorithm?

These four questions' types' appropriateness to assess their corresponding learning outcomes are evaluated by interviews with three professors that are experts on machine learning. The first two questions are to show that two different question types can be used for assessing learning outcomes that are at the same cognitive level, while they do not share the same verb. While the last two questions are to show if two learning outcomes share the same verb, even though they are at different cognitive levels, the same question type can be used to assess them. Combining both propositions, it can be concluded that verbs used in learning outcomes are a better fit to choose the question types than the cognitive level of the learning outcomes.

Even though the amount of interviewed data is not enough to prove the propositions explained in the above paragraph, it can give insights. Blue bars from figure 3 shows averaged ratings for appropriateness of the question type for each question. From the qualitative analysis, it is concluded that question types were fitting to assess their corresponding learning outcome, which can be seen quantitatively from the plot in figure 3 as all the question type ratings are higher than 4.

With the created sample questions and the conclusions from the interviews, three verbs can already be mapped with their "efficient" assessment methodology; "explain", "compute", and "implement" map to open-ended, multiple-choice, and implementation questions respectively. As the phrase "efficient assessment methodology", it is meant the question type that is most feasible but can still assess the corresponding learning outcome with validity and reliability. There are three verbs remaining from the revised learning outcomes;

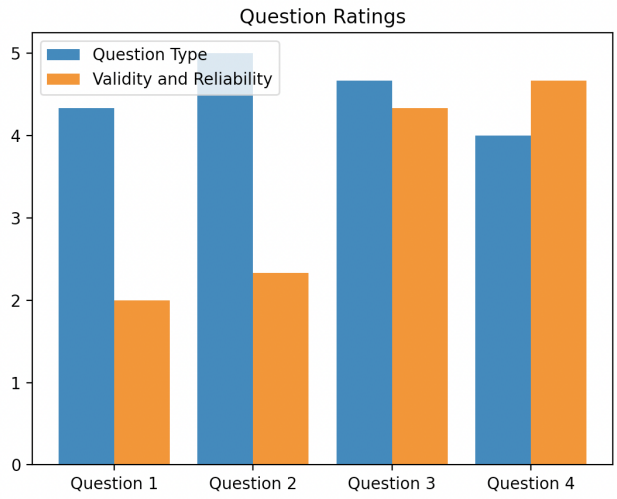


Figure 3: Averaged ratings for question type and validity & reliability for each question from the conducted interviews with machine learning professors.

"Know", "Find" and "Apply". These verbs mapped to multiple-choice questions too, because questions from past exams from the machine learning bachelor course at TU Delft have used multiple-choice question types for assessing similar learning outcomes. By putting all this information into a figure, table 1 has been created.

5 Evaluation

In this research paper, evaluation strategies will be divided into two different categories; pre-assessment evaluation and post-assessment evaluation to clearly indicate if the evaluation strategy should be conducted before or after the assessment has been carried out. Pre-assessment evaluation should be conducted before the assessment is used, and post-assessment evaluation should be performed after students have taken the assessment. As they are mutually exclusive, it essential to implement both of them for the success of the assessments. In this section, both types of evaluations are explained, and the results of the pre-assessment evaluation for the four questions created in section 4 is presented.

5.1 Assessment Evaluation Strategies

Pre-assessment evaluation is essential to having questions with high reliability and validity, and there are few ways of conducting a pre-assessment evaluation. The first way is by peer-reviewing questions that are created by other professors in the course [3]. By conducting this, another expert than the creator of the questions is reading the question and trying to understand. From the interviews conducted with three machine learning professors from Delft University of Technology (TU Delft), it can be concluded that this method is already applied as a pre-assessment evaluation in the machine learning bachelor course. Although, in this peer-review process only the questions are given to the other professors, which means even though problems about the reliability of the questions are uncovered, problems about

Learning Outcome	Verb Used	Cognitive Level	Assessment Methodology
Explain the difference between parametric and non-parametric density estimation.	Explain	Analysis	Open-ended Question
Explain Parzen density estimation and the purpose of window functions in detail.	Explain	Understanding	Open-ended Question
Know which parameter needs to be optimized in Parzen density estimation.	Know	Remembering	Multiple-choice Question
Find Parzen probability density function estimates at a given point using (Gaussian, box, tri, triweight) window function.	Find	Application	Multiple-choice Question
Explain k-nearest neighbors, the influence of parameter k, how to optimize parameter k, and how to break ties in detail.	Explain	Understanding	Open-ended Question
Compute Euclidean, Manhattan, and Hamming distance.	Compute	Application	Multiple-Choice Question
Apply k-nearest neighbors with a specific k parameter, and with one of the above distance metrics.	Apply	Application	Multiple-Choice Question
Implement the k-nearest neighbor classifier in Python.	Implement	Application	Implementation Question
Explain Naïve Bayes, and the effects of choosing Gaussian or Parzen as model per feature in detail.	Explain	Understanding	Open-ended Question
Explain the advantages and disadvantages of the above-mentioned methods.	Explain	Analysis	Open-ended Question

Table 1: Learning outcome to Efficient Assessment Method Table.

validity may be missed as the reviewer may not know what is meant to be assessed by that question. This process can be improved by also sharing the learning outcome that is being assessed with the question [3]. If only a specific part of a learning outcome is meant to be assessed by the question that is being reviewed, then that part can be highlighted and shared with the reviewer. Hauer et al. have created a rubric for peer-reviewing assessments in "Twelve tips for assessing medical knowledge with open-ended questions", this rubric can also be used to make the pre-assessment evaluation process more systematic [12].

The second but more "feasible" way to pre-evaluate assessment questions is the review process in which the reviewer is a teaching assistant instead of a professor that is an expert on the topic. From the interviews conducted, it is learned that this procedure is also applied in the machine learning bachelor course at TU Delft. Again, the improvement suggested in the previous paragraph applies here: Teaching assistants (reviewers) should also be supplied with the learning outcomes for each question for evaluating the validity. An advantage of this pre-assessment evaluation methodology is to get a more realistic perspective, as assessments are going to be taken by students, an evaluation that is done by teaching assistants will uncover problems that are more likely to emerge. For example, a phrase in a question may be confusing for a student, while an expert on the matter can easily understand the phrase. This issue with wording is more likely to be discovered by a teaching assistant than a professor. While the second method is more feasible in an ideal case, both pre-assessment evaluation methodologies should be combined for more diverse and accurate conclusions.

From the analysis of the interviews conducted with machine learning professors, it can be concluded that post-assessment evaluation is as essential as pre-assessment evaluation. The methodologies that are explained for post-assessment evaluation are all acquired from the interviews. First, in the grading process of the open-ended questions, even great answers may not fit rubrics and that is why "rubrics are never final" for these typed questions according to one of the professors. Rubric editing is a post-assessment evaluation methodology, and it increases the reliability of the question [12]. Although, when a rubric is changed, it is critical to start over for the grading of the question to be consistent. After the grading process, a lot of statistics can be used to post-evaluate questions, especially when there are a "big amount" of exam takers this data can be used to identify questions with high reliability and low [12]. For example, if one question has a high mean of grades, but the students that scored high on the exam scored low in the question, that may mean the question has low reliability as "good" students scored low while the average score is high. Therefore, this question needs a further check to see if there is a problem with it. On the contrary, if one question has a low mean, and only "good" students scored high, then this means the question is hard but reliable. If there are questions with a significantly low mean, then these questions may have problems with the wording or may be too difficult for the level of the students. If these problems are not applicable, then this can be an indicator of an issue in the teaching of the corresponding learning outcome according to one of the professors interviewed.

As explained in the above paragraph statistics is a good approach to analyzing the reliability and difficulty of the questions but only using statistics may not be comprehensive. This is why in this paragraph additional methods are going to be briefly explained. The most direct method to learn about what confused or disappointed students is to ask for their feedback [3]. There are several ways this can be done. First, questionnaires can be put after assessments but in summative assessments, students may not have time to fill it, therefore it is important these questionnaires can be reached after the summative too. If there is a need for further evaluation focus groups can be considered [3]. The participators of this focus group may be only one stakeholder group or it can consist of different stakeholders of the assessment [3].

5.2 A Case Study of Pre-Assessment Evaluation

An evaluation has been conducted for the four questions created and presented in section 4 by interviewing three machine learning teachers to simulate a peer-review process. The analysis methods used to analyze interviews are explained in detail in section 6.

The professors are supplied with the questions and the corresponding learning outcomes and asked to rate the question 1 to 5 (5 being the best) considering the validity and the reliability of the question. After their answers, they are asked to justify their rating. The averaged ratings for each question can be seen from the orange bars in figure 3.

The first question rated lowest of all of the questions even though the question type was appropriate according to professors. The problem with the question that is deduced from the analysis of the interviews is the validity of the question. While the learning outcome is only about the application of the distance metrics, the question was requiring comprehension of k-nearest neighbors, which lowers the validity of the question. For example, a student who can apply distance metrics correctly may fail to solve the question if he is lacking knowledge of how to classify a point using k-nearest neighbors. The use of the figure is appreciated by one of the professors. Therefore, the validity of this question can be increased by using a figure but instead of diving into k-nearest neighbors, the question may ask directly which

distance metric origin is closer to point A than point B with a new figure with only two points.

The second question's rating is even higher than the first question's rating, it is clear from figure 3 that there is at least one problem with the reliability or the validity. After the analysis of the interviews, it is concluded that the problem with the question is once again on validity. Two out of three professors indicated that this question is not testing the "Implement the k-nearest neighbor classifier in Python" learning outcome with a high validity as there are no evaluation or training or testing steps of the classifier. Additionally, one professor commented on data structures used in the question may confuse students, which lowers the reliability of the question. Thus with this pre-assessment evaluation conducted, it can be concluded that this question is not a well-made question due to the reasons explained in this paragraph.

The third and the fourth question's ratings are noticeably higher than the first two questions'. Having said that two out of three professors agreed on the third question is excessively open-ended, and therefore the reliability of the question may fall due to the grading of the question. As some students' answers may not completely fit to the rubric created. Even though, these "rubrics are never final" for open-ended questions, if possible it can be beneficial to keep the question as specific as possible.

6 Responsible Research

Being responsible while doing scientific research is critical. If researchers are irresponsible and without integrity, then the academic community may stop believing their research as the data may be fabricated data, or someone else's work may be shown as their own work. Therefore, this section is dedicated to explaining the ethical implications of this research and the effort on making this research reproducible.

Reproducibility is essential in academic research. This is one of the reasons for using learning outcomes in this research. Since learning outcomes defined using Bloom's taxonomy are widely known and used, using this technique of writing course material makes this research easily comparable with other research [5]. Additionally, even though learning outcomes have a very specific format, this research chose the Practical Guide from Kennedy et al. for writing learning outcomes for increasing reproducibility [6].

Identifying problems in learning outcomes can be done in different ways, thus it is important for this research to clearly explain how this has been done to make it reproducible. This is the reason that this research suggests a structured way of identifying problems as it can be found in section 3.

A structured interview has been conducted with three machine learning professors from the Delft University of Technology to support sections; 4 and 5. A structured interview is an interview type that which the interviewer sticks to a set of questions as much as possible, and it is ideal for comparing the results of the interviews [9]. A detailed guideline for how to conduct a structured interview can be found in Wilson's book named "Interview Techniques for UX Practitioners: A User-Centered Design Method" [9]. To inform and get the consent of the professors a consent form has been created and asked for them to fill out. This consent form can be found in Appendix A.2. The questions used in the interview can be found in Appendix A.1.

Lastly, to analyze qualitative questions from the structured interview conducted with the machine learning professors systematically to make it reproducible and more reliable, conventional content analysis is used. This methodology is an inductive method as codes

used in the analysis process are created from the data, and it is a beneficial method when the literature is scarce for the context of the interview. Briefly, first, the recordings of the interviews are transcribed. Subsequently, transcripts are read by the researcher several times to create the codes. The created codes are feasibility, evaluation, peer-review, reliability, insufficiency and isolation. Afterward, these codes are merged to create categories, and lastly, findings have been created. For more detail, "Three approaches to qualitative content analysis" by Hsiu Fang Hsieh and Sarah E. Shannon can be read [4].

7 Future Work

Due to limitations and cohesion of the research, several things were kept out of this research's scope. In this section, these are explained for future studies to build up on this research relatively easily.

First of all, while using Bloom's taxonomy has its advantages like making the research easier to compare with other studies, it also has disadvantageous [5]. For example, some domains or levels in these domains may not apply to each field. To give an example in the machine learning context, the psychomotor domain does not apply in machine learning education. This is why Benjamin Bloom himself said that it is best for each field to have its taxonomy [5]. Therefore, a new optimized taxonomy for machine learning education can be very beneficial to further studies on this topic.

In this study, learning outcomes have been revised from the module "Non-parametric density estimation" and evaluated by a review from a machine learning professor instead of a focus group due to time constraints. This evaluation process could have been more reliable if it would have done in a focus group. This is why, the researcher suggests it to the Delft University of Technology undergraduate machine learning education team to revise all of the learning outcomes in each module, and evaluate them with a focus group that consists of representatives of all stakeholder groups. In The Student Assessment Handbook by Lee Dunn et al., there is a detailed explanation of how to conduct a focus group experiment to evaluate assessment questions but it can also be used to evaluate learning outcomes [3].

After the sample assessment questions were created, and the investigation of the past exams from TU Delft's undergraduate machine learning course was done, table 1 has been filled. However, to prove that verbs used in learning outcomes are a better parameter to choose the appropriate question type than using Bloom's taxonomy classification of the learning outcome, more data needs to be collected. Therefore, in future research assessment questions can be created for all learning outcomes from all modules in TU Delft's machine learning course, and statistical tests can be used to prove the statement.

Lastly, one thing that could not be included in this research due to limitations is returning feedback to students. Feedback is essential, especially in formative assessments, and makes students grow [1]. Therefore, there is a need for various researches on giving reliable feedback to students.

8 Conclusions

This research's aim was to suggest a systematic procedure for creating an assessment that reliably shows progress. To be able to create and evaluate assessments, it is essential to have the course material in a specific format. In this research learning outcomes methodology is used. In section 3, the initial learning outcomes of a module from Delft University of

Technology's Bachelor's machine learning course have been investigated. First, the problems have been identified with these initial learning outcomes. Subsequently, these problems have been fixed by revising learning outcomes. In this step, they also get classified using Bloom's taxonomy to further use the classification while creating assessments. Overall, in section 3 the first sub-question has been answered by finalizing the learning outcomes, and the second sub-question by classifying them using Bloom's taxonomy.

Section 4 is dedicated to suggesting a guideline for creating assessments using learning outcomes, and showing that verbs used in learning outcomes are better to choose assessment methodology (Question type) than using Bloom's taxonomy classification of the learning outcomes. To test this theory, four questions have been created in a way that if the question types proved to be appropriate, then it supports the theory. With the interviews conducted with three machine learning professors, it is concluded that the questions have appropriate types to assess their corresponding learning outcomes. Even though, enough amount of data to prove the theory is not collected, the first insights were supporting it. Even though Bloom's taxonomy was not part of choosing the assessment method, it is most definitely correlated, and it is used while writing the question. For creating the guideline, first, a procedure with solid steps is meant to be produced but later it is concluded that assessment creation has no solid steps but it is a creative process while the creator needs to take into consideration multiple aspects like feasibility. Therefore, instead of a procedure, a guideline has been suggested. By that means, "How can learning outcomes translate into sample assessment questions?" has been answered.

Lastly, in section 5, the last sub-question is answered, which is the ways of evaluating assessment questions. It is concluded that there are two types of assessment evaluation; pre-evaluation and post-evaluation. pre-evaluation should be conducted before the assessment is used while post-evaluation is after. A suggestion is made to how pre-evaluation is applied in the Delft University of Technology's bachelor machine learning course, to improve evaluation of the validity of the assessment questions. Subsequently, conclusions of the pre-evaluation that has been performed by interviewing three machine learning professors to evaluate four questions created are stated. That being said, this research can be used for revising course material, using the guideline suggested to write assessment questions to assess the course material, and for gathering brief information on how to evaluate assessment questions.

References

[1] James W Pellegrino. "Knowing What Students Know". In: *Science and Technology* 19 (2 2002), pp. 48–52.

[2] Lambert W.T. Schuwirth and Cees P.M. Van Der Vleuten. "Written assessment". In: *BMJ* 326 (7390 Mar. 2003), p. 643. ISSN: 14685833. DOI: 10.1136/bmj.326.7390.643.

[3] Dunn Lee et al. *The Student Assessment Handbook : New Directions in Traditional and Online Assessment.* Routledge, 2004. ISBN: 9780415335300.

[4] Hsiu Fang Hsieh and Sarah E. Shannon. "Three approaches to qualitative content analysis". In: *Qualitative Health Research* 15 (9 Nov. 2005), pp. 1277–1288. ISSN: 10497323. DOI: 10.1177/1049732305276687.

[5] Mark G Simkin and William L Kuechler. "Multiple-Choice Tests and Student Understanding: What Is the Connection?" In: *Decision Sciences Journal of Innovative Education* 3 (2005).

[6] Declan Kennedy, Aine Hyland, and Norma Ryan. *Implementing Bologna in your institution C 3.4-1 Using learning outcomes and competences Planning and implementing key Bologna features Writing and Using Learning Outcomes: a Practical Guide.* 2007. URL: http://www.eua.be.

[7] Patricia Armstrong. *Bloom's Taxonomy.* https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/. Accessed: 2022-06-18. Vanderbilt University Center for Teaching, 2010.

[8] Shuhaida Shuhidan, Margaret Hamilton, and Daryl D'Souza. "Instructor perspectives of multiple-choice questions in summative assessment for novice programmers". In: *Computer Science Education* 20 (3 Sept. 2010), pp. 229–259. ISSN: 08993408. DOI: 10.1080/08993408.2010.509097.

[9] Chauncey Wilson. *Interview Techniques for UX Practitioners : A User-Centered Design Method.* Accession Number: 516200; OCLC: 865332107; Language: English. Morgan Kaufmann, 2013. ISBN: 9780124103931.

[10] Brown et al. *The Essentials of Instructional Design.* Routledge, 2016. ISBN: 978-1-138-79705-5.

[11] Diab Abuaiadah et al. "Assessing Learning Outcomes of Course Descriptors Containing Object Oriented Programming Concepts". In: *New Zealand Journal of Educational Studies* 54 (2 Nov. 2019), pp. 345–356. ISSN: 21994714. DOI: 10.1007/s40841-019-00139-y.

[12] Karen E. Hauer et al. "Twelve tips for assessing medical knowledge with open-ended questions: Designing constructed response examinations in medical education". In: *Medical Teacher* 42 (8 Aug. 2020), pp. 880–885. ISSN: 1466187X. DOI: 10.1080/0142159X.2019.1629404.

[13] Rozita Kadar et al. "Students' Assessments in Learning Programming based on Bloom's Taxonomy". In: *Journal of Computing Research and Innovation (JCRINN)* 6 (3 2021), pp. 13–21. URL: https://jcrinn.com:eISSN:2600-8793%20https://creativecommons.org/licenses/by-nc-sa/4.0//.

[14] Amy J. Ko. *We need to learn how to teach machine learning.* URL: https://medium.com/bits-and-behavior/we-need-to-learn-how-to-teach-machine-learning-acc78bac3ff8. (accessed: 24.04.2022).

A Appendix

A.1 Interview Questions

1. How do you choose the question type for a question? (Critical reasoning, short answer, multiple choice, implementation)
2. Do you evaluate the assessment questions you have created? If yes, how?
3. Can you rate the quality of these questions 1 to 5 (5 is the best) considering the learning outcome they are testing, and give any feedback you have?
4. Can you rate these questions' types 1 to 5 (5 is the best) considering the learning outcome they are testing, and give any feedback you have?

A.2 Consent Form

Consent Form for interview conducted as part of the "We need to learn how to teach Machine Learning" Research Project

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study...		
I have read and understood the study information dated 07/06/2022, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
I understand that taking part in the study involves being part of a interview session in which researchers will ask questions and they will take notes on the answers that I give.	☐	☐
I give my permission to be audio-recorded throughout the interview session.	☐	☐
I give my permission that the audio-record will be analyzed and the outcome will be used in a paper that will be made public.	☐	☐
I give my permission to the research team to use this interview in their study and publish it anonymously in their paper.	☐	☐
Risks associated with participating in the study I understand that taking part in the study involves the risk of being identified since there is a limited number of Machine Learning teachers at TU Delft.	☐	☐
Use of the information in the study I understand that the information I provide will affect the outcomes of the research that will further provide an idea for the structure for the CSE2510 Machine Learning course.	☐	☐
I understand that the only personal information that will be used is my profession ("Teacher of the TU Delft's Machine Learning course"), and no other collected information that can identify me, such as my name, will be shared beyond the study team.	☐	☐

Signatures

Signature of the participant
Date:

Signatures of Researchers
Date: