Contents

Improving the ELIXIR evaluation for both management and trainers

- Authors: Richèl Bilderbeek, Daniel Wibberg

Abstract

NBIS teaches, among others, bioinformatics courses, which are evaluated with an anonymous survey sent its learners. Part of this survey consists of mandatory questions to assess course quality. However, there is discussion in how useful these questions are in achieving their goal. Here, we describe the history of this survey, how its questions were crafted and selected, followed by an evaluation of its final form. We find that no selection criterium has been written down and evidence-based best practices by the academic literature were ignored. This paper is the first to transparently show the crafting and selection of evaluation questions to be used for NBIS evaluation, being the 'future work' as mentioned in its paper most important to this topic.

1. Introduction

On 2025-01-20 the NBIS Training Steering Group had a meeting on course evaluations. The first question was 'How are evaluations evaluated?'.

It is common practice that courses are evaluated by surveys [Brazas & Ouellette, 2016][Gurwitz et al., 2020][Jordan et al., 2018].

Although these surveys are developed with the best intentions, it does not necessarily mean that the questions in such surveys are useful. For example, 2 out of 3 teachers of one NBIS course have the shared verdict that the NBIS questions are -I quote- 'useless', where 1 is neutral and reasoning 'it is what we commonly do'.

This disagreement is used as a starting point in evaluating the NBIS course evaluation questions, where the literature is searched for how these questions came to be and by which criteria the best were selected, in the hope of establishing the usefulness of this survey, as well as suggestions for improvements.

Besides discussing the current survey question, this paper is the first to give a fully transparent process on how, with the same goals in mind, a similar set of evaluation questions were developed and how the best questions of this set were selected, with the goal of helping to make course evaluations (even) more useful.

1.1. References

- [Brazas & Ouellette, 2016] Brazas, Michelle D., and BF Francis Ouellette. "Continuing education workshops in bioinformatics positively impact research and careers." PLoS computational biology 12.6 (2016): e1004916.
- [Gurwitz et al., 2020] Gurwitz, Kim T., et al. "A framework to assess the quality and impact of bioinformatics training across ELIXIR." PLoS computational biology 16.7 (2020): e1007976. website
- [Jordan et al., 2018] Jordan, Kari, François Michonneau, and Belinda Weaver. "Analysis of Software and Data Carpentry's pre-and post-workshop surveys." Software Carpentry. Retrieved April 13 (2018): 2023. PDF

2. Research questions

- RQ1: What is the history of the ELIXIR evaluation questions? How were they developed? By which criteria where the best questions selected?
- RQ2: How does the academic literature relate to the ELIXIR evaluation questions?
- RQ3: Which ELIXIR evaluation questions are concluded from a fully transparent process?
- RQ4: How different are the newly suggested questions from the current ones?

3.1. Methods 1

A literature search is performed to find out the history of the current ELIXIR evaluation questions, with a focus on answering the following sub-questions:

- How were these evaluation questions developed?
- By which criteria where the best questions selected?

The results can be found at RQ1 results.

3.2. Methods 2

A literature search is performed to assess the questions in the ELIXIR SFT.

The results can be found at RQ2 results.

3.3. Methods 3

To find out which evaluation questions are concluded from a fully transparent process, we use a procedure that involves multiple phases (as shown in figure M3-F1, each having goals as shown in table M3-T1

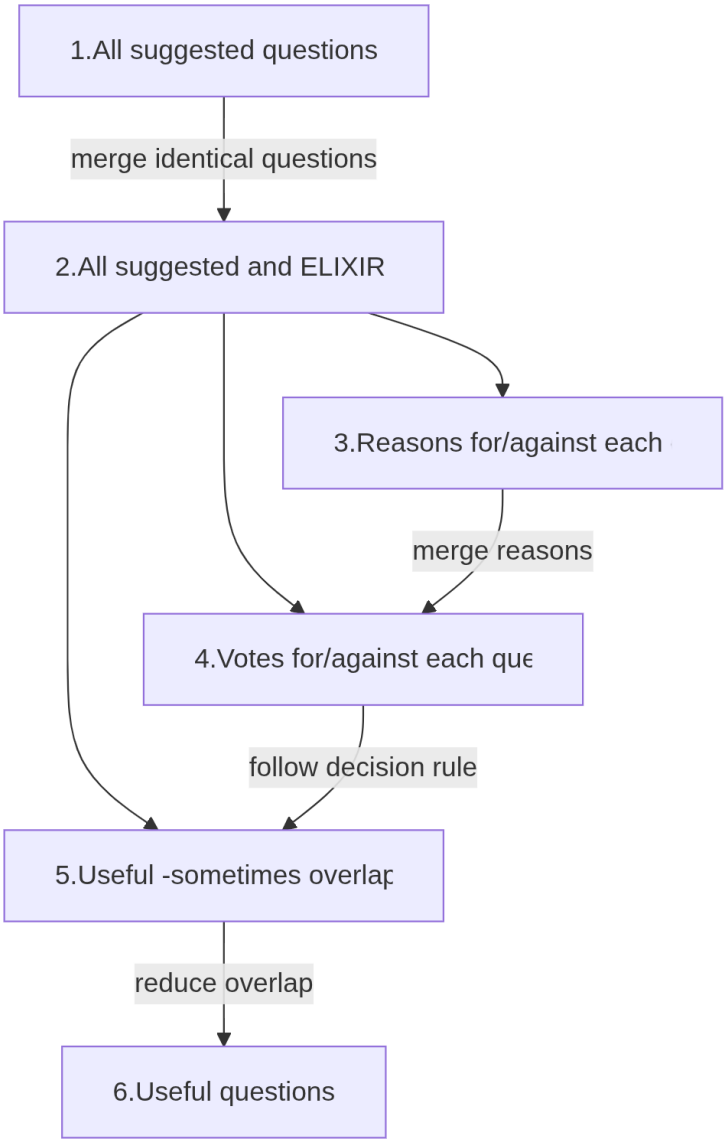

Figure 1: Figure M3-F1. Overview of the procedure

Figure M3-F1. Overview of the procedure

Phase	Goal
1	Collect all questions that are considered 'good' by at least 1 NBIS trainer
2	Collect all questions that are considered 'good' by NBIS and ELIXIR
3	Collect all reasons for and against each question
4	Rate all questions
5	Select the questions that are considered good by the NBIS community
6	Merge overlapping questions

Table M3-T1: goals of each phase in the procedure

Here each step of the procedure is described.

3.3.1. Phase 1

The goal of phase 1 is to collect all questions that are considered 'good' by at least 1 NBIS trainer.

To do so, trainers need to

- be aware of this experiment
- know the goals of ELIXIR
- be invited to submit their questions
- do this before a deadline

At an NBIS Training Liaison meeting, introduce this procedure to the people involved in training, as well as advertise in the relevant communication channels. Present, or share an online presentation online that shows the rationale behind this experiment, as well as the goals of ELIXIR.

In an online anonymous survey, repeat the rationale of this experiment, as well as the ELIXIR goal of the evaluation.

Set a deadline of several weeks. Remind trainers to submit 1 week before the deadline ends.

Collect all questions that teachers think are useful anonymously, creating data_set_1_raw.csv.

If less than 10 questions are collected, this experiment is cancelled. If more than 10 questions are collected, the authors of this paper are allowed to add their favorite questions too.

As there may be duplicates in the data set, remove the duplicates transparently, creating data_set_1.csv and describe the process to do so in data_set_1_merge.md.

3.3.2. Phase 2

Combine `Data Set 1` with the current NBIS questions. Shuffle these questions randomly, creating data_set_2.csv

3.3.3. Phase 3

- Per question, as the teachers anonymously for reasons why they would be for or against each question. The collection of reasonings per questions results in data_set_3.csv

3.3.4. Phase 4

- Per question, and its pros and cons, vote anonymously if the question is useful enough to be included in a survey. Allow 'no', 'yes' and neutral data_set_4.csv

3.3.5. Phase 5

From the questions and votes, select the set of questions that had more 'yes' than 'no' votes: these are the questions that this NBIS community thinks are useful.

The results can be found at data_set_5.csv.

3.3.6. Phase 6

From the questions that had more 'yes' than 'no' votes, merge potential overlap in questions.

The results can be found at data_set_6.csv.

3.4. Methods 4

The results can be found at RQ4 results.

4.1. Results of RQ1: What is the history of the ELIXIR evaluation questions?

4.1.1. What is the ancestry of the NBIS questions?

The paper where these questions were described first in [Gurwitz et al., 2020] . In that paper, one can read that these questions are based on [Jordan et al., 2018] and [Brazas & Ouellette, 2016]. These last two papers do not reference any academic papers on where their questions originated from.

4.1.2. Development of the questions

ELIXIR developed these evaluation questions to, as quoted from [Gurwitz et al., 2020]:

- describe the audience demographic being reached by ELIXIR training events
- assess the quality of ELIXIR training events directly after they have taken place'

The resulting metrics can be found at https://training-metrics-dev.elixir-europe.org/all-reports.

This is what is written about how the ELIXIR short-term evaluation questions came to be (quote from [Gurwitz et al., 2020]):

> We were interested in participant satisfaction as a reflection on training quality in order to be able to inform best practice for ELIXIR training. We acknowledge that training quality is more complex than solely participant satisfaction and that the community would benefit from future work to obtain a fuller picture on training quality.

This paragraph shows that this ELIXIR group took the liberty of adding questions besides its two primary sources.

Again from [Gurwitz et al., 2020] we read:

> These metrics were developed out of those already collected by ELIXIR training providers, as well as from discussions with stakeholders, external training providers, and literature review [Brazas & Ouellette, 2016][Jordan et al., 2018]

There are no references to the literature that was reviewed besides these two papers.

Neither does the referred literature:

- [Brazas & Ouellette, 2016] shows the results of surveys from bioinformatics workshops. The survey questions where taken from other sources (i.e., the Society for Experimental Biology and the Global Organisation for Bioinformatics Learning, Education and Training), without any reference to the literature. It is not described how the evaluation questions came to be and with which reasoning the best were selected
- [Jordan et al., 2018] shows the results of surveys from Data Carpentry workshops. Also here, it is not described how the evaluation questions came to be and with which reasoning the best were selected: this paper has zero references to the literature

Taking a closer look at the evaluation questions of [Jordan et al., 2018], we see that some questions of its evaluations were not copied to the ELIXIR evaluation. The reasoning why some questions were copied and some not is unpublished.

4.1.3. References

- [Ang et al., 2018] Ang, Lawrence, Yvonne Alexandra Breyer, and Joseph Pitt. "Course recommendation as a construct in student evaluations: will students recommend your course?." Studies in Higher Education 43.6 (2018): 944-959.
- [Brazas & Ouellette, 2016] Brazas, Michelle D., and BF Francis Ouellette. "Continuing education workshops in bioinformatics positively impact research and careers." PLoS computational biology 12.6 (2016): e1004916.
- [Darling-Hammond et al., 2010] Darling-Hammond, Linda, Xiaoxia Newton, and Ruth Chung Wei. "Evaluating teacher education outcomes: A study of the Stanford Teacher Education Programme." Journal of education for teaching 36.4 (2010): 369-388.
- [Gurwitz et al., 2020] Gurwitz, Kim T., et al. "A framework to assess the quality and impact of bioinformatics training across ELIXIR." PLoS computational biology 16.7 (2020): e1007976. website
- [Jordan et al., 2018] Jordan, Kari, François Michonneau, and Belinda Weaver. "Analysis of Software and Data Carpentry's pre-and post-workshop surveys." Software Carpentry. Retrieved April 13 (2018): 2023. PDF
- [Liaw et al., 2012] Liaw, Sok Ying, et al. "Assessment for simulation learning outcomes: a comparison of knowledge and self-reported confidence with observed clinical performance." Nurse education today 32.6 (2012): e35-e39.
- [Roxå et al., 2021] Roxå, Torgny, et al. "Reconceptualizing student ratings of teaching to support quality discourse on student learning: a systems perspective." Higher Education (2021): 1-21.

- [Raupach et al., 2011] Raupach, Tobias, et al. “Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness.” Medical teacher 33.8 (2011): e446-e453.
- [Plaza et al., 2002] Plaza, Cecilia M., et al. “Curricular evaluation using self-efficacy measurements.” American Journal of Pharmaceutical Education 66.1 (2002): 51-54.
- [Uttl et al., 2017] Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. “Meta-analysis of faculty’s teaching effectiveness: Student evaluation of teaching ratings and student learning are not related.” Studies in Educational Evaluation 54 (2017): 22-42.

4.2. Results of RQ2

With the goal of the SFT ('to improve the course and its materials') in mind, here we go through the mandatory questions that resulted from the process described in the results of Research Question 1. The relevant questions are found in `Section 3 - Quality Metrics` of the NBIS short-term evaluation. Here, we go through each of these questions in detail.

4.2.1. Question 5

```
5. Have you used the tools/resource(s) covered in the course before?

- Never - Unaware of them
- Never - Used other service
- Occasionally
- Frequently
```

Question 5 is an interesting way to evaluate the quality of a course, because it is about something learners have done **before** the course took place. Searching the literature for 'using previous experience in course evaluations' (and sentences alike) resulted in zero hits.

4.2.2. Question 6

```
6. Will you use the tools/resource(s) covered in the course again?

- Yes
- No
- Maybe
```

Question 6 is another interesting way to evaluate the quality of a course, because it is about the usefulness of the topic being taught, combined with predicting the future. Searching the literature for 'using self-predicted future use of content in course evaluations' (and sentences alike) resulted in zero hits.

4.2.3. Question 7

```
7. Would you recommend the course?

- Yes
- No
- Maybe
```

Question 7 attempt to measure course quality by asking the learner if he/she would recommend the course. This question originates from one of the two evaluations that this ELIXIR evaluation is based on (`[Jordan et al., 2018]`).

Searching the literature for 'using course recommendation in evaluation' (and sentences alike) resulted in one relevant hit. This paper, `[Ang et al., 2018]`, shows that using this question may indeed be a valid way to asses course quality `[Ang et al., 2018]`.

4.2.4. Question 8

```
8. What is your overall rating for the course

- Poor (1)
- Satisfactory (2)
- Good (3)
- Very Good (4)
- Excellent (5)
```

Question 8 too attempts to measure course quality by asking the learner to rate it. This question is absent from the two questionnaires (i.e. those described in `[Brazas & Ouellette, 2016]` and `[Jordan et al., 2018]`) this questionnaire is based one.

Searching the literature for 'using course satisfaction in evaluations' (and sentences alike) resulted in many relevant papers. The most important paper is a meta-analysis, which concluded that there is no relation between training quality and participant satisfaction [Uttl et al., 2017] and this meta-analysis gives some examples how problematic this metric is.

4.2.5. Question 9

```
9. A. May we contact you by email in the future for more feedback?

- Yes
- No
```

Question 9 is an interesting way to measure the course quality, based on the learner being willing to answer questions on the future. It seems more likely that question should be placed outside of the section `Section 3 - Quality Metrics`.

Searching the literature for 'using future contact in course evaluation' (and sentences alike) resulted in zero relevant hits.

4.2.6. References

- [Ang et al., 2018] Ang, Lawrence, Yvonne Alexandra Breyer, and Joseph Pitt. "Course recommendation as a construct in student evaluations: will students recommend your course?." Studies in Higher Education 43.6 (2018): 944-959.
- [Brazas & Ouellette, 2016] Brazas, Michelle D., and BF Francis Ouellette. "Continuing education workshops in bioinformatics positively impact research and careers." PLoS computational biology 12.6 (2016): e1004916.
- [Jordan et al., 2018] Jordan, Kari, François Michonneau, and Belinda Weaver. "Analysis of Software and Data Carpentry's pre-and post-workshop surveys." Software Carpentry. Retrieved April 13 (2018): 2023. PDF
- [Uttl et al., 2017] Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related." Studies in Educational Evaluation 54 (2017): 22-42.

4.3. Results of RQ3

On 2025-02-17:

- a presentation was given at the NBIS TrSG, with 5 people attending. The presentation lasted around 10 minutes, after which there was 15 minutes of questions
- after the meeting, a video of that presentation was recorded
- NBIS teachers were invited to participate

4.4. Results of RQ4: How different are the newly suggested questions from the current ones?

```
No results yes
```

5. Conclusion

5.1. RQ1: What is the history of the ELIXIR evaluation questions?

The ELIXIR student evaluation is based on two evaluations. Neither evaluation describes (1) how its questions were developed, (2) by which criteria the best questions were selected.

From these two questionnaires, some (but not all) questions were selected to be put in the ELIXIR evaluation. Additionally, ELIXIR added new questions, regarding student satisfactions. The criteria for neither not copying questions for the earlier questionnaires, not for adding adding questions were not written down.

As none of these evaluations describe the process on how questions were developed, it comes as no surprise that neither evaluation refers to evidence-based best practices in the literature.

5.2. RQ2: How does the academic literature relate to the ELIXIR evaluation questions?

The mandatory ELIXIR evaluation questions with the goal of evaluating course quality have the following relation to the academic literature

Question	Relation to the literature
5	None found
6	None found
7	One paper which states that this question is suitable to assess course quality
8	Many papers, including a meta-analysis that this question is unsuitable to assess course quality
9	None found

5.3. RQ3: Which ELIXIR evaluation questions are concluded from a fully transparent process?

`TODO`

5.4. RQ4: How different are the newly suggested questions from the current ones?

`TODO`

6. Discussion

First and foremost, one can argue that a questionnaire is simply a questionnaire. That the development of a questionnaire is 'just' a group effort. That the people in such groups are constrained by time. And that this is a good enough excuse. However, this overlooks the fact that at least sixteen thousand participants have taken the time to fill in this questionnaire. If we care about our participants, maybe we should care about the usefulness of a questionnaire we send to each of them.

6.1. RQ1: What is the history of the ELIXIR evaluation questions?

Although this is not formally published, maybe the current survey (or the surveys it is based on) have been developed by evidence-based best practices, yet in unpublished and more informal documents. However, it seems unlikely that references to the literature are used in informal communication, yet subsequently omitted when a formal academic paper is written.

To us, it seems more likely that a too short amount of time was allocated to researching the academic literature, resulting in missing the papers mentioned in this paper.

6.2. RQ2: How does the academic literature relate to the ELIXIR evaluation questions?

The majority of the ELIXIR mandatory evaluation questions, in the section to assess course quality have little connection to the academic literature. Three out of five questions (i.e. questions 5, 6 and 9) resulted in zero papers being written on their effectiveness.

To us, these questions simply seem to be in the wrong session. Why it was chosen to put these questions in the section called 'quality metrics', instead of a (new) section with a better fitting name is unknown.

Regarding the two questions that seem to be in the correct section, however, only one is supported by the literature ('Would you recommend the course?') by one paper, where the other ('How satisfied are you with the course?') has strong support **against** its effectiveness.

Regarding these two questions, they seem to be measuring the same thing: course satisfaction (on its own), and recommending a course (because the learner is satisfied with the course). It is unknown why these two similar questions are both in the evaluation and it would be interesting to see how strong the correlation is between the answers on these two questions.

6.3. Epilogue

We know that a teacher reflecting on his/her work is one of the best ways to increase his/her teaching quality. Or: 'student ratings can only become a tool for enhancement when they feed reflective conversations about improving the learning process and when these conversations are informed by the scholarship of teaching and learning `[Roxå et al., 2021]`. The other best way for teachers to improve is to do peer observations. Note that neither practice needs an evaluations.

If we really care about teaching quality, shouldn't we encourage doing the things that actually work?

6.4. References

- `[Roxå et al., 2021]` Roxå, Torgny, et al. "Reconceptualizing student ratings of teaching to support quality discourse on student learning: a systems perspective." Higher Education (2021): 1-21.

Appendix

A1. NBIS Short Term Feedback (STF)

```
Besides this text, this page shows the current
NBIS Short Term Feedback form. Except for changes in layout
and the numbering of sections,
the content is unmodified.
```

A1.1.Core question set information

The intention of the STF survey is to find out how participants have used the skills and knowledge they gained through participating in the NBIS course.

The STF survey aims to provide data back to NBIS from course participants.

The survey should preferably be given by the course leader to the participants on the last day of the course. Some of the questions below are CORE Questions and needs to always be included in the survey. There are also room for ADDITIONAL questions that can be modified for respective course.

- Contents
- Important Information
- Core Question Set
- Demographic Information
- Quality Metrics
- Additional Questions - Training content/information
- Additional Questions - Training logistics

A1.2. Important Information

Below are the core questions for NBIS short term feedback (STF), which are required to be captured for all NBIS training events from August 2018 onwards, most typically in an end-of-training-event feedback survey (i.e. exit survey). The information and Core questions are extracted from the ELIXIR and ELIXIR-EXCELERATE courses. Additional questions are free to be modified to suit the course needs. The format for collecting the data is up to each training provider, although results should be exportable to Excel format. The core questions may be divided into two categories and will by and large be analysed separately - both categories are required to be captured:

- Demographic information
- Quality metrics

For the demographic information questions specifically, these may be captured either in the exit survey OR in the registration form. The exit survey should be administered as close as possible to the end of the training event, preferably on the last day of the course. Please add the result of the survey to the course folder in Google Drive (NBIS Course Catalogue).

The core question set is followed by a set of Additional (suggested) questions that training organisers might also like to ask. Please note: while the core question set is compulsory, Course leader(s) are encouraged to ask any additional questions for their own collection and data analysis, should they wish.

Data formatting: Preferred column headers for each core metric are in 'red'. It would be very helpful for analysing the data if everyone used these column headings when exporting the results. Please note: these descriptors are case sensitive (e.g. use `advertised` not `Advertised`). Also, the underscores are important! (e.g. `career_stage` is NOT the same as `career stage`).

If possible, please name the dataset file as follows to assist with data handling: `YYYY-MM-DD_L/STF_Location_CourseName`, e.g. `2018-06-11_STF_Visby_RaukR`

A1.3. Core Question Set

A1.4. Section 1 - Template: NBIS Short Term Feedback (STF) survey `COURSE NAME, LOCATION, YYYY-MM-DD`

Thank you for filling the questionnaire. It is really important to us in order to continually improve the course and the materials we deliver. In filling the questionnaire, please keep in mind that your comments - which are not mandatory - are especially precious. We may share anonymised information with course presenters and developers as well as for wider quality/impact analyses.

A1.5. Section 2 - Demographic Information

1. Where did you see the course advertised? `advertised`
    - a. NBIS website
    - b. SciLifeLab website
    - c. Social Media (e.g. NBIS twitter)
    - d. Host Institute website
    - e. Colleague
    - f. TeSS
    - g. Email
    - h. Internet search
    - i. Other (comments)
2. What is your career stage? `career_stage`
    - a. PhD candidate
    - b. Postdoctoral researcher
    - c. Senior researcher/Principal investigator
    - d. Staff scientist
    - e. Industry scientist
    - f. Other (comments)
3. What is your host university? `host_university`
4. Gender `gender`
    - a. Male
    - b. Female
    - c. Prefer not to say
    - d. Other (please specify)

A1.6. Section 3 - Quality Metrics

5. Have you used the tools/resource(s) covered in the course before? `have_used_resources_before`
    - 1. Never - Unaware of them
    - 2. Never - Used other service
    - 3. Occasionally
    - 4. Frequently
6. Will you use the tools/resource(s) covered in the course again? `will_use_resources_future`
    - 1. Yes
    - 2. No
    - 3. Maybe
7. Would you recommend the course? `would_recommend_course`
    - 1. Yes
    - 2. No
    - 3. Maybe
8. What is your overall rating for the course*. `overall_satisfaction`
    - a. Poor (1)
    - b. Satisfactory (2)
    - c. Good (3)
    - d. Very Good (4)
    - e. Excellent (5)

(*please include both numeric and categorical scale for this question.)

9. A. May we contact you by email in the future for more feedback? `contact_future`
    - 1. Yes
    - 2. No

9 B**. If you answered 'yes' to the above question, please enter your email address, below. email (** Information for question 9B must be collected and stored by each Node/Institution, but should NOT be shared with the Q&I subtask or any other third party due to GDPR considerations.)

A1.7. Additional Questions - Training content/information

These are suggested questions that may be of interest (not compulsory):

1. What part of the training did you enjoy the most? `enjoy`

2. What part of the training did you enjoy the least? `to_improve`

3. The balance of theoretical and practical content was `theoretical_practical`

   - `a`. Too practical
   - `b`. About right
   - `c`. Too theoretical

4. How do you rate the pre-course information given? pre_course_information

   - Linear scale 1-5
   - `1`. (Very unsatisfactory/Not useful)
   - `5`. Very good/Very useful

5. What other topics would you like to see covered in the future? `future_topics`

6. Any other comments? Comments

7. PLEASE RATE EACH SESSION OF THE COURSE `satisfaction_per_session_YYYY_MM_DD_am/pm`

   - `a`. Did not attend
   - `b`. Poor (1)
   - `c`. Satisfactory (2)
   - `d`. Good (3)
   - `e`. Very Good (4)
   - `f`. Excellent (5)

8. Comments on teaching staff `teaching_staff` Help our teaching staff to improve by providing constructive feedback Paragraph text answer

9. Was the course held at a teaching level matching your training? teaching_training_level

10. STATEMENTS REGARDING WHAT PARTICIPANTS COULD DO before TRAINING (customised to a specific training) skills_before

11. STATEMENTS REGARDING WHAT PARTICIPANTS CAN DO after TRAINING (customised to a specific training) skills_after

12. What other topics would you like to see covered in the future? future_topics

13. Any other comments? Comments_1

A1.8. Additional Questions - Training logistics

These are suggested questions that may be of interest (not compulsory):

1. What would be the preferred length of the course? `preferred_length`

   - Linear scale 1-5 Days

2. How did you like the facilities/localities of the course (rooms and surroundings)? `course_localities`

   - Linear scale 1-5
   - `1`. Not at all
   - `5`. Very much

3. How did you like the lunch(es) and "fika(s)"? lunch_fikas

   - Linear scale 1-5
   - `1`. Not at all
   - `5`. Very much

4. Any other comments? `Comments_2`

It was a great experience and we are working hard to make it even better. Now go make something great!