

Level-2 Evaluation Toolkit – Frequently Asked Questions

1. What is a Level-2 evaluation?

The term “Level-2 evaluation” comes from Donald Kirkpatrick’s four Levels of evaluation, where the second level **assesses what and how much participants in a course learn** from it.¹

The Evaluation Group of the World Bank Institute has measured participant learning by testing course takers at the very beginning of the course, testing them on an equivalent test at the end of the course, and computing the “**learning gain**” by deducting the average class pre-test score from the average class post-test score.

The evaluation aims to assess the **course effectiveness in imparting knowledge to its participants in order to improve the course**. It does not aim to test individual course takers for accreditation.

2. If I do a Level-2 evaluation, do I still need to evaluate the course at Level-1?

Yes, Level-1 and Level-2 evaluations measure different aspects of the course. They **complement** each other.

A Level-1 evaluation asks participants their opinions of the course. It is best used to refine the course design.

A Level-2 evaluation tests participants’ knowledge of the course. It is best used to refine the course contents.

3. How many items do we need on a test?

We recommend that you ask a minimum of 20 test questions on the pre-test and 20 on the post-test.

This means that collectively the item writers should write a **minimum of 40 items, or 20 pairs of items**.

From an evaluative viewpoint, the more items the better. However, testing participants takes time away from the course, so the right amount of items is a trade-off between testing enough parts of the course and having not too many items to avoid taking away too much time from the course.

Please note that for technical reasons the Toolkit can only process 50 items on the pre-test and 50 on the post-test (100 in all).

4. Should easy items be on the pre-test and difficult items on the post-test?

No. The pre-test and post-test should have the **same level of difficulty**. The purpose of the evaluation is to assess the effectiveness of the course in imparting knowledge to its participants. This can only be achieved with test of equivalent difficulty. Using a pre-test easier than the post-test would defeat the purpose of the evaluation.

5. How to make the pre-test as difficult as the post-test?

The Toolkit explains how to pair items, pilot the tests, and randomly assign one item in each pair to the pre-test and one to the post-test. In combination, these techniques help to make the tests equivalent in difficulty.

¹ Kirkpatrick, Donald L., Evaluating Training Programs, The Four Levels, 2nd edition, Berrett-Koehler Publishers, Inc., 1998.

6. How difficult should be the test items?

Ideally, each item should have an **average** difficulty level. In writing an item, the content expert should expect less than half of the participants to know the answer if the item happened to be asked on the pre-test and more than half to answer correctly if it is asked on the post-test.

7. What does “pairing” items mean?

Pairing items consist of writing two items so they test the same concept, but being phrased differently. The aim is to create two items of even level of difficulty without triggering pure participants recall between the pre- and post-test.

8. Why not use the same items on the pre- and post-tests?

Using the same items of both tests can trigger recall (pre-test effect, as an internal threat to validity of the evaluation).

9. What should be the passing score for participants on the post-test?

There is **no passing score** using the Toolkit. The aim is to evaluate the course, not individual participants.