HO 1c Number of questions, alternatives, and available time

Introduction

Apart from giving insight into the distribution of test questions across *content* and *target levels*, the test matrix (see handout 1b) also provides insight into the *number* of questions that is needed to assess the learning goals in a valid and reliable way. In this handout, we cover the factors that play a role in determining the number of questions of a test (i.e., test length). The length of the test depends, of course, on the nature and the magnitude of the learning content. Furthermore, the guess-chance (the probability randomly guessing the correct answer) is also a factor to take into account for MC tests. The guess-chance depends on the number of alternatives that belong to the test questions. The available time of the staff is often a restrictive factor for the number of questions. After all, constructing sound MC questions can take a lot of time. Finally, the number of questions that can be used in a test depends on the available time for administration of the test, which is usually scheduled in advance.

Questions that we answer in this handout are:

- How many questions should I include in a test? And, what happens when the test questions have 3 or 4 alternatives?
- How do I determine beforehand how many time students will need to finish the test?

Test length

Five factors determine the length of a test, namely: the function of the test (*high stakes* or *low stakes*), the intended validity, the intended reliability, the number of answer alternatives, and the available time for test administration.

1. **The function of the test:** is it a practice test, a formative test, or a summative test?

The aim of a practice test is to give students an idea of what type of questions they can expect in the summative test, so that they can prepare themselves for the summative test. The criterion for the number of questions in the practice test is, in this case, as many questions as are needed to give a representative idea of the variation in type of questions, that will be used in the summative test. Generally, a set of 10 questions will already be enough to fulfil the practice function.

Formative tests (*low stakes* test), go one step further than practice tests. It is not about passing or failing yet, but you do want to be able to make a reasonably reliable and valid evaluation of the extent to which students master the learning contents, and where improvement is needed. Students can draw their own conclusions on the basis of the results of the test(elements), but teachers can also provide targeted feedback and directions to the student on the basis of the formative test. The number of questions of a formative test depends on the feedback that you want to give. In addition, a formative test will also create expectations regarding the summative test. The formative test should thus be comparable to the summative test, so that students can improve themselves on the parts on which they will be assessed in the summative test. Depending on the philosophy of the degree programme, and how formative feedback is viewed, it is possible that a summative test is merely a sample from the (many and extensive) formative tests that have preceded the summative test. In our education, however, it is usually the other way around: the formative test is a representative subset of the more extensive summative test.

The goal of a summative test (high stakes test) is that important pass/fail decisions can be made on the basis of the test results. Factors 2 to 5 play a role in determining the length of this test.

2. **The validity:** The test covers the course goals and is congruent with the study-activities of the students. The extent and nature of the goals will be determining factors for the number of questions, and the table of specifications is the instrument that is used to construct a test that covers the goals. In handout 1b, more information about the table of specifications can be found.

3. **The reliability** of the test: the number of questions is sufficient to make a reliable judgement. If the test is constructed in a uniform manner (using a table of specifications), then it is possible to make a reasonable judgement of the reliability of the test on the basis of experience of the teacher and colleagues with previous tests. More certainty about the reliability of the test can be obtained after the test is administered and the test analysis is available.

The test analysis will show if the test actually contained enough questions to be able to uncover differences between students. The rule of thumb is: the more questions the test contains, the more reliable (consistent) the test is. The relationship between test length and reliability is shown in Table 1. Suppose that the measured reliability (α) of a test is 0.60, whereas the intended reliability is 0.75, then the test needs to contain twice as many (2K) similar questions. Obviously, when the choice is made to add questions, the time for test administration also needs to be adjusted. Before any adjustments are made to the number of questions, it should be decided what the value of adding extra questions is in comparison to the extra time that is needed. Formulating more questions will take a lot of time, whereas it often only leads to a small increase in reliability. If the test is not *very* unreliable, one should ask oneself if adding questions is worth the extra time investment.

Table 1: Reliability (α) when shortening / lengthening the test, with K-questions, bij K-vragen

K	1.5 K	2 K	3 K
α	α'	α''	α'''
0.20	0.27	0.33	0.40
0.40	0.50	0.57	0.60
0.60	0.69	0.75	0.77
0.80	0.86	0.89	0.91

4. **The guess-chance and the number of and quality of the alternatives.**

When determining the number of questions needed in an MC test, the guess-chance should be taken under consideration. The guess-chance is the chance that a student chooses the correct answer purely on the basis of guessing. Generally speaking, the more alternatives a question has, the lower the guess-chance is. Therefore, a test that consists of true/false questions should contain more questions to come to a reliable measurement than a test with questions with four alternatives. Table 2 gives an indication of the number of questions that is needed in tests with different numbers of alternatives, to obtain a similar measurement range.

Table 2: Relationship between the number of alternatives, test length and measurement range for MC-tests
Reference: Milius (2007)

Number of alternatives(A)	Test length (N)	Guess-chance (R=N:A)	Measurement range (N-R)
Four alternatives	40	10	30
Three alternatives	45	15	30
Two alternatives	60	30	30

Here, we are talking about the theoretical and average guess-chance on the basis of a wild guess. Fortunately, students will not make wild guesses, but make their choice on the basis of the knowledge they possess and the information that is provided in the question. The addition of an alternative also adds extra information, which sometimes makes it easier instead of more difficult for students to find the right answer. The quality (and difficulty) of the question will affect the establishment of a well-reasoned guess-chance. Below, you will find three examples of questions which each have a theoretical guess-chance of 50%.

Example a ***True/false question*** *(guess-chance* 50%)	**Example b** ***Question with two alternatives*** (guess-chance 50%)	**Example c** ***Question with two alternatives*** (guess-chance 50%)
Montevideo is the capital of Uruguay A. True B. False	Montevideo is the capital of: A. Uruguay B. Paraguay	Montevideo is the capital of: A. Uruguay B. France

The guess-chance of example a and b may be the same, however in example b, more information is given. The same is true for example c, which is the easiest question of the three. At the same time, example c is possibly an excellent question in the context of primary education (e.g., 'Countries and capitals of the world). In short, the level of difficulty is connected to the goal, and affects the guess-chance. A well-reasoned guess-chance /question is directly related to the cut score (pass mark. See handout 5a for a simple procedure for how the level of difficulty, and the level of mastery can be accounted for in the pass mark.

Whether or not adding an alternative makes a question more difficult, depends on the quality of that alternative. Imagine that at example b, the answer 'France' would be added. The theoretical guess-chance of the question would then decrease from 50% to

33,3%. However, for most students, it will be easier to determine that France is a wrong answer. The quality of this alternative, is thus low and the students will eliminate this answer directly. Therefore, the actual guess-chance of this question is close to 50% and the question has not become more difficult.

5. **The available time for test administration**
Once a decision is made on the number of questions (see 1 to 4), the next step is to make a realistic assessment of the time needed for test administration. Preferably, the allotted time is not too long, because this sends out a wrong signal to the students. If students get, for example, 3 hours for a MC-test with 20 questions, this can damage the credibility. The time student get for the test is thus a good reflection of the estimated time that is needed to answer the questions, and is more than adequate. In our education, an important principle is that students should receive enough time to take a test. The speed with which students solve problems should, therefore, not play a significant role. A rule of thumb is that an 'average student' can finish the test in ⅔ of the available time. If this is, for example, 60 minutes, then 90 minutes is a reasonable duration for all students. Apart from these considerations, an indication can be given for the time that is needed to solve questions with 2, 3, 4 or 5 answer alternatives.

Rule of thumb for the required time and the number of alternatives.
As the number of alternatives increases, students need more time to solve a question. For a question with two alternatives, students need, on average, 50 seconds, while they need 60 seconds for a question with three alternatives, and 75 seconds for a question with four or five alternatives (Van Berkel & Bax, 2013). These are only indications, based on the average, and they obviously do not hold for questions that ask students to do a calculation, or to read and interpret a case with raw data.

Further reading

Berkel, H.J.M. van, & Bax, A.E. (2014). Toetsen met gesloten vragen (In: Toetsen in het hoger onderwijs. 3e druk. Houten: Bohn Stafleu VanLoghum)

Dousma, T., Horsten, A., & Brants, J. (1997). Tentamineren. Groningen: Wolters-Noordhoff.

Milius, J.J. (2007). Schriftelijk tentamineren: een draaiboek voor docenten in het hoger onderwijs. IVLOS, Universiteit Utrecht.