# Contents

# Developing maximum performance tests

# 2

Maximum performance tests ask the test takers to do the best they can to perform a task. These tests are used to assess a wide variety of abilities, aptitudes, knowledge, and skills.

The development of a test starts with the making of a plan. The plan specifies a number of essential elements of test development: (1) the construct of interest, (2) the measurement mode of the test, (3) the objectives of the test, (4) the population and subpopulations where the test should be applied, (5) the conceptual framework of the test, (6) the response mode of the items, and (7) the administration mode of the test. These seven elements need not to be specified in the given order, and can be considered simultaneously or in another order. However, the plan must give attention to each of these elements.

The first seven sections of this Chapter discuss these elements of the plan. The remaining sections address other essential aspects of test development: item writing guidelines, item rating guidelines, pilot studies on item quality, and compiling the first draft of the test.

## 2.1 The construct of interest

The test developer must specify the latent variable of interest that has to be measured by the test. In the context of test development the terms 'latent variable' and 'construct' are used. Here, 'latent variable' will be used as a general term, while 'construct' will be used when a substantive interpretation is given of the latent variable, for example, the general statement 'the latent variable of the test' versus the substantive statement 'the construct of verbal intelligence'. The latent variable (construct) is assumed to effect test takers' item responses and test scores.

Constructs can vary in many different ways, including the following three. First, constructs vary in content from *mental abilities* (e.g., intelligence) to *psychomotor skills* (e.g., manual dexterity) and *physical abilities* (e.g., athletic capacity). Second, constructs vary in scope, from, for example, general intelligence to multiplication skill. Third, constructs vary

from educational to psychological variables. *Achievement tests* measure constructs that result from instruction and teaching. For example, an algebra test measures the outcomes of an algebra course or curriculum. *Ability* and *aptitude tests* measure constructs that are relatively independent of instruction and teaching. For example, a numerical intelligence test tries to measure an ability that is not explicitly taught, although the development of numerical ability may be stimulated by instruction and teaching. Frequently, the terms 'ability' and 'aptitude' are used interchangeably. According to Cronbach (1990, p. 701) an aptitude test is an ability test that is used to predict future competences, such as those required to be successful in a job or educational program.

A good way to start a test development project is to define the construct that has to be measured by the test. This definition describes the construct of interest, and distinguishes it from other, related, constructs. Usually, the literature on the construct needs to be studied before the definition can be given. Moreover, frequently the definition can only be given when other elements of the test development plan, such as the objectives of the test and the conceptual framework, are specified. However, somewhere at the beginning of the test development process a definition of the construct has to be given. An example is the following definition of the reading comprehension construct.

*Example 2.1 A definition of reading comprehension*
Gorin (2007, p.186) gives the following verbal definition of reading comprehension: 'Reading comprehension questions measure the ability to read with understanding, insight, and discrimination. This type of question explores the ability to analyze a written passage from several perspectives. These include the ability to recognize both explicitly stated elements in the passage and assumptions underlying statements or arguments in the passage as well as the implication of those statements or arguments.'

## 2.2 The measurement mode

Different modes can be used to measure constructs (Fiske, 1971, Chapter 5). Here are some measurement modes of maximum performance tests.

The common measurement mode is to ask test takers to perform a mental or physical task. For example, a student is asked to solve a numerical problem, or a disabled person is asked to walk ten meters without help. This mode will be called the *self-performance mode*, but other modes can be applied as well.

Instead of performing the task themselves, the test taker could be asked to evaluate his (her) ability to perform the task. For example, a student could be asked how good he or she is at solving numerical problems, or a disabled person could be asked whether he or she could walk on his (her) own for ten meters. This mode will be called the *self-evaluation mode* (Example 2.2).

*Example 2.2 Self-evaluation of aptitudes*
The Differential Aptitude Test (DAT) consists of nine subtests to measure different aptitudes. For example, the subtest Numerical Ability presents test takers with a number of numerical problems that have to be solved. Oosterveld (1996, Chapter 4) described a self-evaluation version of the DAT. For example, a self-evaluation question to measure test takers' numerical ability is: 'If I get a 15% discount in a shop, I know what price to expect' (yes/no).

A third possibility is to ask others to evaluate a person's ability to perform a task. This mode will be called the *other-evaluation mode* (Example 2.3).

*Example 2.3 Other-evaluation of aptitudes*
The self-evaluation version of the DAT (Example 2.2) can easily be converted into an other-evaluation version. For example, a teacher could be asked to assess the students' numerical ability. An example of a question to a teacher on John's numerical ability is: 'If John gets a 15% discount in a shop, he knows what price to expect' (yes/no).

### 2.3 The objectives

The test developer must specify the objectives of the test. Tests are used for many different purposes. Here are some distinctions that are relevant for the planning of the development of a test.

Tests can be used for scientific (e.g., to study human intellectual functioning) or practical purposes (e.g., to select job applicants or to assess students' math achievements). A second distinction is between objectives at the level of individual test takers and at the level of a group of test takers. Examples of individual level objectives are the use of tests to accept or reject an applicant for a job, or to pass or fail students for an examination. Examples of group level objectives are the use of mean test scores to compare the educational achievements of 16-year old students from different countries, and the use of students' mean achievements test scores to assess the quality of schools.

A third distinction of test objectives is between description, diagnosis, and decision-making. *Description* means that the test is only used to describe performances. For example, psychotherapists may apply tests to better understand their clients, and experimental psychologists, may apply tests to see whether the participants of their experimental and control groups are comparable. *Diagnosis* goes a step further than description by adding a conclusion to a description. For example, a teacher may conclude from a student's arithmetic test score that he or she is weak in multiplication, and a government may conclude from a national assessment study that students' math performance is alarming low. Finally, *decision-making* means that decisions are based on tests. For example, the teacher may decide to give remedial teaching to the student, the government may decide to take measures to improve students' math performance, and a company may use tests to decide which of the applicants will be hired.

## 2.4 The population

The *target population* of a test is the set of persons to whom the test has to be applied. The test developer must define the target population, and must provide criteria for the inclusion and exclusion of persons. Usually, a general description of the population is too vague, and has to be supplemented by inclusion and exclusion criteria. For example, the population description 'Dutch adults above 18 years' leaves the test developer with many open questions, for example: Persons who can read Dutch or also persons who cannot read Dutch? Persons who can speak Dutch or also persons who cannot speak Dutch? Dutch citizens or inhabitants of the Netherlands? Mentally healthy persons only or also mentally handicapped persons and psychiatric patients? Etcetera.

A target population can be split into distinct subpopulations, for example, males and females, majority and minority groups, age groups, and so on. The test developer must specify whether subpopulations need to be distinguished, and, if so, they need to define the subpopulations, and to provide criteria to include persons in subpopulations. For example, which minority groups are distinguished, and who belongs to each of these groups?

## 2.5 The conceptual framework

Test development starts with a definition or description of the construct that has to be measured by the test. However, the definition or description is usually not concrete enough to write test items. For example, Gorin's (2007, p. 186) definition (Example 2.1) clearly describes the reading comprehension construct, but an item writer needs more specific information, such as the length and difficulty of the reading passages.

A *conceptual framework* gives the item writer a handle to write items. In the literature, examples of conceptual frameworks are available. Two examples from the history of test development are discussed in brief.

The first example is a conceptual framework for the writing of achievement test items.

*Example 2.4 A taxonomy of educational objectives for writing achievement test items*

Bloom, Engelhart, Furst, Hill and Krathwohl, (1956) proposed a taxonomy of educational objectives for writing achievement test items. The taxonomy has two dimensions: (1) the subject matter that has to be tested, and (2) the cognitive domain that has to be assessed. The elements of the subject matter dimension are the topics of the study program. Bloom et al. (1956) used six elements of the cognitive domain dimension (knowledge, comprehension, application, analysis, synthesis, and evaluation). Research showed that the taxonomy has some imperfections, but simplified and modified versions of Bloom et al.'s taxonomy have proven their value for item writing (Haladyna, 2004, Chapter 2). Table 2.1 shows a taxonomy for item writing on a statistics course for psychology freshmen. The elements of the subject matter dimension are the topics that are part of the course (mean, variance, etc.). The cognitive domain dimension has only three elements, that is, recall, understanding, and application.

Table 2.1 *A taxonomy for writing items on a statistics course for psychology freshmen.*

| | Dimension | | |
|---|---|---|---|
| | Cognitive | | |
| Topic | Recall | Understanding | Application |
| 1. Mean | 0 | 0 | 1 |
| 2. Variance | 0 | 0 | 1 |
| 3. Correlation | 1 | 1 | 2 |
| Etc. | | | |

Each combination of a topic and a cognitive domain element specifies an educational objective. Moreover, the table specifies the number of test items per objective, for example, the test must contain one recall, one understanding, and two application items on correlation. A recall item on correlation asks, for example, about the definition of the product moment correlation coefficient, an understanding item on the interpretation of correlation, and an application item on the use of the product moment correlation coefficient in empirical research.

The second example is a conceptual framework for the writing of intellectual ability test items.

*Example 2.5 Guilford's Structure-of-Intellect model for ability test items*
Guilford (1967) used three dimensions to classify intellectual abilities: first, operations, which is the activity that has to be executed to solve the task. This dimension has five elements: cognition, memory, divergent production, convergent production, and evaluation. Second, content, which is the type of material of the task. This dimension has four elements: figural, symbolic, semantic, and behavioral. Finally, the product dimension, which is the way the information of the task is organized. This dimension has six elements: units, classes, relations, systems, transformations, and implications. The combination of five operations, four contents, and six products yields $6 \times 5 \times 4 = 120$ different abilities. For example, the combination of the cognitive operation, behavioral content, and transformation product yields the Cognition of Behavioral Transformations (CBT) ability, which is the ability to understand and recognize changes of behavioral information. For each of these 120 abilities, test items can be constructed. For example,

Rombouts-ten Hove Jansen (1978) gives the following example of an item to test the CBT ability:

**Item 2.1 A CBT item, adapted from Rombouts-ten Hove Jansen (1978, p.86)**
Director says to secretary 'please'.
Which of the following sentences has a *different* meaning?
a) Beggar says to unknown 'please'.
b) Child says to grandma 'please'.
c) Chauffeur says to boss 'please'.

The item is on the understanding (cognition) of the change in meaning of the word 'please' (transformation) in a conversation between two persons (behavior).

The literature reports a number of test development projects that are based on psychological and educational theories. For further information, the reader is referred to Embretson (1985), Millman and Green (1989), Wilson (2005), and Yang and Embretson (2007).

## 2.6 The item response mode

The *item response mode* needs to be specified before item writing starts. The main distinction is between the *free-* or *constructed-response* and *choice* or *selected response* modes. Item 2.2 is an example of the free-response mode, while Item 2.3 is the same item in choice mode.

**Item 2.2 Example of free-response mode**
$8 \times 12 = ...?$

**Item 2.3 Example of choice mode**
$8 \times 12 = ...?$
a) 84
b) 96
c) 108

Free-response items are further divided into *short-answer items* and *essay items*. Item 2.2 is an example of a short-answer item. An essay item asks the test taker to give an elaborate response. An example is Item 2.4.

**Item 2.4 Example of an essay item**
Which test types are distinguished in the first chapter of this book?

Different types of choice modes are used in achievement and ability testing; an overview is given by Haladyna (2004, Chapter 4). The most popular choice mode is the conventional multiple-choice mode, the only one that is discussed in this book.

A conventional multiple-choice item consists of a *stem* and two or more *options*. The options are divided into one correct answer and one or more *distractors*. Item 2.5 shows the structure of a three-choice item.

**Item 2.5 Structure of a three-choice item**

| | |
|---|---|
| Stem: | What is the capital of Spain? |
| | |
| Options: | |
| Distractor: | a) Lisbon |
| Correct answer: | b) Madrid |
| Distractor: | c) Rome |

The stem of Item 2.5 is a direct question. The stem can also be formulated as a partial sentence: 'The capital of Spain is'. Empirical research has not given conclusive evidence that one type of stem should be preferred above the other (Ascalon, Meyers, Davis, & Smits, 2007).

Usually, choosing the correct option of a multiple-choice item indicates that test takers' ability or skill is sufficiently high to solve the item, whereas choosing a distractor indicates that their ability is too low. However, distractors can be constructed to contain specific information on the reasons why the test taker failed to solve the item correctly. The choice of a distractor indicates which deficiency the test taker has and as such can be used for diagnosing specific deficiencies (Gorin, 2007). Example 2.6 demonstrates this type of distractor construction in mathematics testing.

*Example 2.6 Diagnostic distractor construction*
Timmer (1969) described the construction of multiple-choice items to assess high school students' skill to solve linear equations. He distinguished a number of subskills that are needed for this skill. Two of these subskills are:

Subskill 1: The ability to multiply both sides of an equation by the same number. For example, a student who masters this subskill will give the correct solution of the equation $4 = \frac{x}{2}$ (i.e., $x = 8$).

Subskill 2: The ability to eliminate parentheses. For example, a student who masters this subskill will give the correct elimination of parentheses from 5(x + 1) (i.e., 5x + 5).

The subskills can be combined for item writing. Item 2.6 shows an item that combines the multiplication and parentheses elimination subskills.

**Item 2.6 Item that has diagnostic distractors (Timmer, 1969, Section 10.2.4)**

$6(x+1) = \frac{1}{2}$ implies that

a) $3x + 1 = 1$
b) $3x + 3 = 1$
c) $12x + 1 = 1$
d) $12x + 12 = 1$

The choice of the c-distractor indicates that the student masters the first subskill ($6(x + 1) = \frac{1}{2}$ implies that $12(x + 1) = 1$), but does not master the second subskill ($12(x + 1)$ is not equal to $12x + 1$). The choice of the b-distractor indicates that the student does not master the first subskill ($6(x + 1) = \frac{1}{2}$ does not imply that $3(x + 1) = 1$), but masters the second subskill ($3(x + 1)$ is equal to $3x + 3$). The choice of the a-distractor indicates that the student does not master either of the two subskills, and the choice of the correct d-option indicates that the student masters both subskills.

Example 2.6 demonstrates that the choice of distractors can be used for diagnosis and remedial teaching. For example, a student who chooses the c-distractor does not know how to eliminate parentheses, and needs remedial teaching on this subskill.

The options of Item 2.6 appear to be partially ordered. The choice of the d-option indicates a higher skill level than the choice of the b- or c-distractor, and the choice of the b- or c-distractor indicates a higher skill level than the choice of the a-distractor. However, the choices of the b- and c-distractors cannot be ordered because the choice of each of these two distractors indicates a deficiency in one of the two subskills. Therefore, the options of Item 2.6 are partially ordered with respect to the skill: d above b and c, and b and c above a.

Briggs, Alonzo, Schwab, and Wilson (2006) described *ordered multiple-choice items*, where the options are completely ordered (Example 2.7).

*Example 2.7 Ordered multiple-choice item*
Briggs et al. (2006) constructed items to assess students' understanding of the 'Earth in the Solar System'. They distinguished four ordered levels of understanding this system by fifth grade students:

Level 1: Student does not recognize the systematic nature of the appearance of objects in the sky.

Level 2: Student recognizes that (1) the Sun appears to move across the sky every day, and (2) the observable shape of the Moon changes every 28 days.

Level 3: Student knows that (1) the Earth orbits the Sun, (2) the Moon orbits the Earth, and (3) the Earth rotates on its axis.

Level 4: Student knows that (1) the Earth is both orbiting the Sun and rotating on its axis, (2) the Earth orbits the Sun once per year, (3) the Earth rotates on its axis once per day causing the day/night cycle and the appearance that the Sun moves across the sky, and (4) the Moon orbits the Earth once every 28 days, producing the phases of the Moon.

Item 2.7 has options that are ordered, according to these four levels of understanding.

**Item 2.7 Options ordered according to four levels of understanding (Briggs et al., 2006)**

It is most likely colder at night because

a) the Earth is at the further point in its orbit around the Sun. (Level 3).
b) the Sun has traveled to the other side of the Earth. (Level 2).
c) the Sun is below the Earth and the Moon does not emit as much heat as the Sun. (Level 1).
d) the place where it is night on Earth is rotated away from the Sun. (Level 4).

The four options of Item 2.7 are completely ordered according to the levels of understanding: d above a, a above b, and b above c.

For the analysis of item response data it is convenient to distinguish different types of *item response scales*. Test takers' answers to free-response items are assessed in a dichotomous or ordinal-polytomous scale. A dichotomous item response scale has two ordered categories, that is, an answer is correct or incorrect. For example Item 2.2 has a dichotomous (incorrect/correct) item response scale. An ordinal-polytomous scale has more than two ordered categories, for example, answers to Item 2.4 can be

correct, partly correct, or incorrect. Usually, the test taker's choice of an option of a conventional multiple-choice item, such as Item 2.5, is assessed on a dichotomous scale: the choice of the correct option is correct and the choice of a distractor is incorrect. However, the scale of conventional multiple-choice items can also considered to be partially ordered: the correct option is ordered above the distractors, but the distractors are not ordered among themselves. Diagnostic multiple-choice items may also have a partial ordinal-polytomous response scale. For example, option d of Item 2.6 is ordered above Options b and c, and Options b and c are ordered above Option a. Finally, ordered multiple-choice items, such as Item 2.7, are assessed on an ordinal-polytomous scale because the options are completely ordered. To conclude, three item response scales are distinguished to assess test takers' responses to maximum performance items: the (1) *dichotomous*, (2) *partial ordinal-polytomous*, and (3) *ordinal-polytomous item response scales.*

## 2.7 The administration mode

A test can be administered to test takers in different ways. The main modes are: (1) oral, (2) paper-and-pencil, (3) computerized, and (4) computerized adaptive test administration.

*Oral administration* means that the test is presented orally by a single test administrator to a single test taker. The use of this administration mode is needed when test takers have difficulty in reading the test, such as, young children, non-native speakers, visually handicapped, etc.

*Paper-and-Pencil (P & P) administration* means that the test is presented in the form of a booklet: test takers read the instruction, answer the items, and report their free responses or choices. The test can be administered in groups, for example, a class of students, but it can also be administered individually, for example, by sending the test booklet to the test taker's home. The test and the order of the items is the same for each of the test takers.

*Computerized Test (CT) administration* resembles P & P administration in the sense that the test and the order of the items is the same for each of the test takers. However, the difference with P & P administration is that the test is presented on a computer instead of a booklet, and the test takers report their free responses or choices by the computer.

*Computerized Adaptive Test (CAT) administration* is the most advanced administration mode. The test is adaptive in the sense that the computer program searches for the items that best fit the test taker (e.g., difficult items for high ability test takers, and easy items for low ability test takers).

Consequently, different items are presented to different test takers, and the item administration order may differ per test taker. CAT is appropriate for advanced applications and is discussed further in Section 14.3 of this book.

## 2.8 Item-writing guidelines

Item writing is a hard job that requires experience in both the content of the ability or achievement and item writing techniques. The first draft of an item must be reviewed by several experts. Haladyna (2004, p. 97) reports that it is common that about 50% of the concept items do not survive a critical review procedure. Moreover, items that survive are mostly revised several times before they can be admitted to a test.

Haladyna, Downing, and Rodriguez (2002) reviewed educational measurement textbooks and research studies on guidelines for item writing. Most of these guidelines also apply to the writing of ability test items. Some important guidelines for item writing are discussed below.

### *2.8.1 Focus on one relevant aspect*

The test constructor has made a specification of the achievement or ability that will be measured by the test. Each item should focus on a single relevant aspect of the specification in order to guarantee good coverage of the important aspects of the achievement or ability. Moreover, only a single aspect of the specification needs to be measured to guarantee that test takers' item responses are unambiguously interpretable. An example of an item that disregards this guideline is Item 2.8.

**Item 2.8 Testing two aspects in one item**
What are the capitals of Italy and Spain?
a) Lisbon and Madrid
b) Lisbon and Rome
c) Madrid and Rome

The correct option is c (Madrid and Rome). However, the choice of Option c is ambiguous. An examinee could choose Option c because he or she knows that Madrid is the capital of Spain and Rome is the capital of Italy, but he or she could also choose Option c because he or she thinks that Madrid is the capital of Italy and Rome is the capital of Spain.

### *2.8.2 Use independent item content*

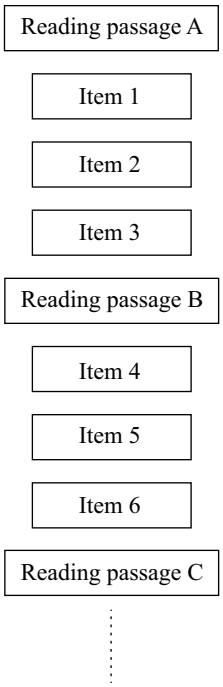
In general, it is recommended that the content of different items is independent. Items 2.9 (1) and 2.9 (2) disregard this guideline.

**Items 2.9 (1) and (2) Dependent item content**
(1) What is the product of 6 and 24?
(2) What is the square root of the product of 6 and 24?

An examinee can only give a correct answer to Item 2.9 (2) if he or she can give a correct answer to Item 2.9 (1).

In some testing situations it is more efficient to use dependent items than independent items. An example is the testing of reading comprehension. A reading passage is given and a set of items asks questions on the same passage. A schematic representation is given in Figure 2.1.

Figure 2.1. *Schematic representation of a reading comprehension test that has passage-dependent items.*



The inter-item dependence is more subtle than the dependence of Items 2.9 (1) and 2.9 (2), but the items are dependent because they refer to the same reading passage (In Figure 2.1, Items 1, 2, and 3 refer to Passage A, Items 4, 5, and 6 refer to Passage B, etc.). A reading passage set of items (e.g., Items 1, 2, and 3 of Figure 2.1) is an example of a testlet. Wainer, Bradlow, and Wang (2007, pp. 52-53) defined a *testlet* as 'a group of items that may be developed as a single unit that is meant to be administered together'. The usual psychometric methods do not apply to testlet data, and these data have to be analyzed by special psychometric methods; an overview of testlet methods is given by Wainer, Bradlow, and Wang (2007).

### *2.8.3 Avoid overly specific and overly general content*

The disadvantage of overly specific item content is that the content may be trivial (Item 2.10), and the disadvantage of overly general content is that the content may be ambiguous (Item 2.11).

**Item 2.10 Overly specific item content**
Which is the first guideline for item writing that was discussed in Section 2.8.1 of this book?

This item is trivial because it is not relevant that an examinee recalls the order in which the guidelines are discussed in this book.

**Item 2.11 Overly general item content (Haladyna, 2004, p. 103)**
Which is the most serious problem in the world?

a) disease
b) hunger
c) lack of education

The item is ambiguous because it is too general.

### *2.8.4 Avoid items that deliberately deceive test takers*

The item deliberately distracts the test takers' attention from the problem that they have to solve (Item 2.12).

**Item 2.12 Item that deceives test takers (Roberts, 1993)**
For finding the Y' regression equation, the standard deviation of X is 23.34, the standard deviation of Y is 18.27 and the correlation between X and Y is .394. What is the b or slope value for finding the X' value given a Y value?

a) .308
b) .503
c) .783
d) 1.278

The item asks for the regression of the X variable on the Y variable ('finding the X' value given a Y value'). However, the beginning of the stem misleads the test taker by focusing on the regression of Y on X ('finding the Y' regression'). The test taker may fail this item because he or she misreads the stem. Therefore, the item is not only measuring statistics, but also close reading.

### *2.8.5 Keep vocabulary simple for the population of test takers*

A rule of thumb is that items for a population of adult native speakers should not require reading skills beyond that of a twelve year old. For other populations (e.g., young children, non-native speakers, etc.) the required reading skill will be lower, and must be adapted to the population of test takers.

### *2.8.6 Put item options vertically*

The horizontal order of options saves space, but may be harder to read than the vertical option order. Therefore, options should be placed in vertical order as was done in Item 2.5.

### *2.8.7 Minimize reading time and avoid unnecessary information*

Verbosity obscures the item content and unnecessary information distracts test takers from the problem that they are asked to solve (Item 2.13 (1)).

**Item 2.13 (1) Unnecessary information (Haladyna, 2004, p. 110)**
High temperatures and heavy rainfall characterize a humid climate. People in this kind of climate usually complain of heavy perspiration. Even moderately warm days seem uncomfortable. Which climate is described?

a) Savanna
b) Tropical rainforest
c) Tundra

Item 2.13 (2) is the same item where the unnecessary information is removed.

**Item 2.13 (2) Unnecessary information of Item 2.13 (1) removed (Haladyna, 2004, p. 110)**
Which term below describes a climate with high temperatures and heavy rainfall?

a) Savanna
b) Tropical rainforest
c) Tundra

### *2.8.8 Use correct language*

The language usage (capitalization, grammar, punctuation, spelling) of items has to be correct.

### *2.8.9 Use non-sensitive language*

The wording of an item should be non-sensitive to test takers of the population. According to the Educational Testing Service (ETS), which constructs a large number of achievement tests, this means that an item (1) should not foster stereotypes, (2) should not contain ethnocentric or gender-based underlying assumptions, (3) should not be offensive from a test taker's point of view, (4) should not contain controversial material, and (5) should not be elitist or ethnocentric. Moreover, the ETS criteria state that the whole test should be balanced with respect to multicultural material (Ramsey, 1993). The ETS uses trained reviewers to detect concept items that may be sensitive to some of the test takers.

### *2.8.10 Use a clear stem and include the central idea in the stem*

The stem should be clear. Moreover, the central idea of the problem should be in the stem instead of the options (Item 2.14).

**Item 2.14 Central idea not in the stem (Ebel, 1965, p. 178)**
Physiology teaches us that

a) The development of vital organs is dependent on muscular activity.
b) Strength is independent of muscle size.
c) The mind and body are not influenced by each other.
d) Work is not exercise.

Changing the stem to 'What does physiology teach us?' does not improve this item because physiology teaches many different things (Ebel, 1965, p. 178).

### *2.8.11 Word the item positively, and avoid negatives*

The stem and options of an item should be worded positively. Negatively phrased items (e.g., not, except, etc.) are harder to understand, and may confuse test takers (Haladyna, 2004, p. 111). Item 2.15. is an example of a negatively phrased item.

**Item 2.15 Negatively phrased item (Ebel, 1965, p. 178)**
In the definition of a mineral, which of the following is incorrect?

a) It was produced by geologic processes.
b) It has distinctive physical properties.
c) It contains one or more elements.
d) Its chemical composition is variable.

A test taker could easily misread the word 'incorrect' as 'correct'.

### *2.8.12 Use three options, unless it is easy to write plausible distractors*

Test developers prefer a large number of distractors per item because the probability of randomly guessing the correct answer is smaller when the

number of options is large. For example, the probability of randomly guessing the correct answer of a three-choice item is 1/3, whereas the probability of randomly guessing the correct answer of a six-choice item is 1/6. However, theoretical and empirical research indicates that generally three options (the correct option and two distractors) is preferable (Rodriguez, 2005). The reason is that it is frequently hard to write more than two plausible distractors, that is, distractors that look like a correct answer to test takers who do not know the correct answer. Exceptions to this guideline are areas where it is easy to write plausible distractors, for example, for some mathematics and numerical items.

#### *2.8.13 Use one option that is unambiguously the correct or best answer*

Item 2.16 disregards this guideline.

**Item 2.16 Two correct options**
Where was the cradle of democracy?

a) Athens
b) Egypt
c) Greece
d) Rome

Athens is in Greece, and, therefore, both Options a) and c) are correct answers.

#### *2.8.14 Place the options in alphabetical, logical, or numerical order*

Examples are Items 2.3 and 2.5, where the options are in numerical and alphabetical order, respectively.

#### *2.8.15 Vary the location of the correct option across the test*

For example, vary the correct options of a three-choice test so that about 1/3 of the first options is correct, about 1/3 of the second options is correct,

and about 1/3 of the third options is correct. Usually, this guideline is automatically satisfied if the guideline of the previous section (options in alphabetical, logical, or numerical order) is applied.

### *2.8.16 Keep the options homogeneous in length, content, grammar, etc.*

Inexperienced item writers tend to make the correct option longer than the distractors because they need more words for a correct formulation than an incorrect one. Therefore, test wise test takers, who do not know the correct answer, will choose the longest option.

Options should be homogeneous with respect to, among others, content and grammar (Item 2.17).

**Item 2.17 Options heterogeneous with respect to grammar (Haladyna, 2004, p. 119)**

The most effective strategy in playing winning tennis for a beginner is

a) more pace in ground strokes.
b) to keep the ball in play.
c) volley at the net as often as possible.
d) hit the ball as hard as possible.

Only Option b) is grammatically consistent with the stem of the item that suggest that b) is intended to be the correct option.

Ascalon et al. (2007) studied the effect of content similarity of item options. They found that items that have options of similar content are more difficult than items that have options of dissimilar content. They prefer options of similar content because dissimilar distractors may be implausible (see Section 2.8.18).

### *2.8.17 Avoid 'all-of-the-above' as the last option*

The all-of-the-above option is correct if one of the options is the best answer, but the other options are not perfectly incorrect. Item 2.18 demonstrates the disadvantage of the all-of-the-above option.

**Item 2.18 Disadvantage of all-of-the-above option (Ebel, 1965, p. 182)**
What does the term 'growth' mean?

a) Maturation
b) Learning
c) Development
d) All of these

The best answer is 'maturation', but the other options are not perfectly incorrect. Therefore, it could be argued that 'all of these' is also a correct answer.

### *2.8.18 Make distractors plausible*

Each of the distractors should look like a possible correct answer to test takers who do not know the correct answer. The analogy Item 2.19 disregards this guideline.

**Item 2.19 Implausible distractor**
DOOR is to ROOM as COVER is to

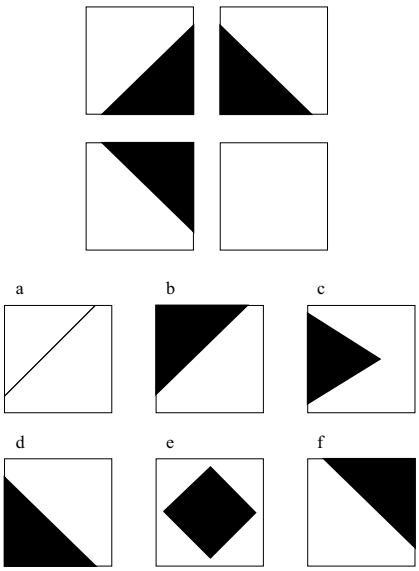
a) box
b) content
c) cow

The last option ('cow') is implausible. Test takers who do not know the correct answer will eliminate the last option and guess between the first two options.

### *2.8.19 Avoid giving clues to the correct option*

Van der Maas (2007) gives examples of hidden clues in intelligence tests. He tried to derive the correct option from the options by looking only at the options and ignoring the stem of the items, and found the correct answer to many of the items. Item 2.20 from Raven's Progressive Matrices test demonstrates his strategy to find the correct option from hidden clues in the options.

**Item 2.20 Hidden clues in options to correct answer**




Van der Maas looked at the characteristics of the options, and their frequency of occurrence:
(1) The color black occurs in 5 out of the 6 options;
(2) The triangle in the edge occurs 4 out of 6 times;
(3) The triangle is in the left upper corner twice.
Combining these characteristics, Van der Maas concluded that the correct option is a black triangle at the left corner (Option b), which is indeed the correct answer.

## 2.9 Item rating guidelines

The responses to free- (constructed-) response items have to be graded by raters. Hogan and Murphy (2007) reviewed textbooks on psychological and educational measurement, and derived, among others things, guidelines for rating free responses. Some important guidelines are described below.

### *2.9.1 Rate responses anonymously*

Preferably, items should be rated without knowing the identities of the test takers.

### *2.9.2 Rate the responses to one item at a time*

If the test contains more than one free-response item, the responses of all test takers to one item should be rated, before moving on to the responses to the next item.

### *2.9.3 Provide the rater with a frame of reference*

Raters should be given rating instructions, schemes, or ideal responses that they can use as a frame of reference.

### *2.9.4 Separate irrelevant aspects from the relevant performance*

Concentrate on the relevant performance, and disregard irrelevant aspects. For example, when measuring numerical ability, disregard writing (spelling, grammar, etc.), or rate writing separately from the numerical ability rating.

### *2.9.5 Use more than one rater*

If possible, more than one rater should rate free responses to the items. The raters have to work completely independently, which means that they should not confer with each other, and that they should not know each other's ratings. The ratings of independent raters can be used to study the agreement between different raters (see Section 2.10.3 of this chapter).

### *2.9.6 Re-rate the free responses*

A rater can change his (her) rating standards in time. Therefore, free responses, or parts of them, should be re-rated. The same free responses are rated on more than one occasion. The time between the occasions and the size of the rating task are chosen so that, on the second occasion, the rater cannot remember his (her) first ratings.. The ratings of the same rater taken on different occasions can be used to study the rater's consistency (see Section 2.10.3 of this chapter).

#### *2.9.7 Rate all responses to an item on the same occasion*

A rater could change his (her) standards in time. Therefore, all responses to one item should be rated on the same occasion.

#### *2.9.8 Rearrange the order of responses*

A rater could change his (her) standards in a sequence of ratings. Therefore, the order of the responses should be rearranged when going from the responses to one item to the responses of another item.

#### *2.9.9 Read a sample of responses*

Read a sample of responses before the start of the rating. The best procedure is to arrange training sessions, where a sample of responses is read and rated by a group of raters. The rating procedure, the frame of reference, and the ratings are discussed, and raters' comments are used to improve the rating procedure.

### 2.10 Pilot studies on item quality

Standard practice is that item writers produce a set of concept items and pilot studies are done to test the quality of these concept items. Generally, at least half of the concept items do not survive the pilot studies, and items that survive are usually revised several times.

Experts' and test takers' pilot studies need to be done for both free-response and multiple-choice items. Additionally, for free-response items pilot studies need to be done on the ratings of test takers' responses to the items.

#### *2.10.1 Experts' pilots*

The concept items have to be reviewed before they are included in a test. Items are reviewed on their content, technical aspects, and sensitivity.

The content and technical aspects are assessed by experts in both the field of the test and item writing. For example, math teachers are needed

to assess whether the concept items of a mathematics test are correct from substantive and technical points of view, and whether the items meet the educational objectives of the curriculum. The teachers need to have a sufficient background in mathematics and sufficient experience in teaching, and they need to be trained in item writing, item reviewing, and applying item writing guidelines (see Section 2.8). Each of the concept items is discussed by a panel of experts. A good start for the discussion of a multiple-choice item is to look for distractors that panel members could defend as (partly) correct answers. The reviewing of the items yields qualitative information that is used to rewrite items or to remove concept items that cannot be repaired. Revised items should be reviewed again by experts until further rewriting is not needed.

The sensitivity of items also needs to be reviewed. Usually, the panel for the *sensitivity review* of the items consists of persons not on the panel reviewing the content and technical aspects of the items. The sensitivity review panel is composed of members of different groups, for example, males and females of minority groups. The panel has to be trained to detect aspects of the items and the test that may be sensitive to subpopulations, for example, females or minority group members. Ramsay (1993) gives a case study of sensitivity analysis of the ETS. The sensitivity review provides qualitative information that also could lead to rewriting or removal of concept items.

### *2.10.2 Test takers' pilots*

The concept items are individually administered to a small group (10 to 20) test takers from the population of interest. Each of the test takers is interviewed on their thinking while working on an item. Two versions of the interview can be applied (Leighton & Gierl, 2007). Using the *concurrent interview* the test taker is asked to think aloud while working on the item, and using the *retrospective interview* the test taker is asked to recollect his (her) thinking after completing the item. Protocols of the interviews are made, and the information is used to rewrite or remove concept items.

### *2.10.3 Raters' pilots*

Raters must grade the responses of test takers to free-response items. Usually, two types of rating scales are applied. First, the *dichotomous scale* is used, where test takers responses are graded in two ordered categories, that is, a

correct or an incorrect answer. Second, the *ordinal-polytomous scale* is applied, where test takers' responses are graded in more than two ordered categories, for example, a correct, partly correct, or incorrect answer (three ordered categories), and A, B, C, D, and E (five ordered categories).

The ratings of free-responses raise two different questions on the quality of the ratings. First, how consistent are the ratings of a given rater? A rater may be inconsistent in the sense that he or she gives a different rating when rating a test taker's free response to an item for the second time. Second, to what extent do different raters agree among themselves? Different raters, who rate the same free response, may disagree in their rating. Therefore, two types of rating pilot studies are needed on (1) the consistency of each of the raters, and (2) the agreement between different raters.

The ratings of a given rater are completely consistent if the rater gives identical ratings on different occasions. An *intrarater consistency* study is done to assess the rating consistency of a given rater. The raters are adequately trained to do the rating job. The same free responses are rated on two or more occasions by the same raters. The time between two occasions and the number of free responses that are rated are chosen so that, on the second occasion, the raters cannot remember their ratings on the first occasion. Table 2.2 gives the (hypothetical) frequencies and proportions of a given Rater A's ratings of the free responses of 100 test takers to the same item on two occasions.

Table 2.2 *(Hypothetical) frequencies and between parentheses proportions of Rater A's ratings of the free responses of 100 test takers to the same item on two occasions.*

| **Rater A** | | **Occasion 2** | | | |
|---|---|---|---|---|---|
| | | Correct | Partly correct | Incorrect | |
| Occasion 1 | Correct | 68 (.68) | 1 (.01) | 0 (0) | 69 (.69) |
| | Partly correct | 2 (.02) | 9 (.09) | 1 (.01) | 12 (.12) |
| | Incorrect | 1 (.01) | 0 (0) | 18 (.18) | 19 (.19) |
| | | 71 (.71) | 10 (.10) | 19 (.19) | 100 (1) |

The table shows, for example, that Rater A rated 68 responses as correct on both occasions, 1 response as correct on the first occasion and partly correct on the second occasion, and 0 responses as correct on the first occasion and incorrect on the second occasions. The diagonal of the table contains the frequencies and proportions of completely consistent ratings on the two occasions: correct/correct 68 (.68); partly correct/partly correct: 9 (.09); incorrect/incorrect: 18 (.18). Rater A seems to be quite consistent because the observed proportion *(O)* of identical ratings is high:

$O = .68 + .09 + .18 = .95.$

The observed proportion of identical ratings *(O)* may give a wrong impression of Rater A's consistency. The reason is that a certain degree of consistency can be expected by chance alone. Under the assumption that Rater A's ratings on the two occasions are completely unrelated and the marginal proportions are fixed, the chance expected proportion of a diagonal cell is the product of the corresponding marginal proportions, for example, the chance expected proportion correct/correct ratings is: .69 × .71 = .49. Therefore, the chance expected proportion *(E)* identical ratings is:

$E = .69 \times .71 + .12 \times .10 + .19 \times .19 = .54.$

Cohen (1960) proposed coefficient kappa to correct for change expected consistency. The coefficient compares the gain in proportion identical ratings $(O - E)$ to the maximum possible gain, which is $(1 - E)$ because the maximum observed proportion identical ratings is $O = 1$. The definition of *Cohen's coefficient kappa* is

*Definition 2.1 Cohen's coefficient kappa*

$$Kappa = \frac{O - E}{1 - E} \quad (E < 1).$$

Kappa reaches its maximum value ($Kappa = 1$) if consistency is perfect, that is, $O = 1$, and Kappa = 0 if the observed ratings are completely unrelated, that is, $O = E$.

Using Definition 2.1, the value of kappa for Rater A's consistency (Table 2.2) is:

$$Kappa = \frac{.95 - .54}{1 - .54} = .89.$$

This value is high, which indicates that Rater A's ratings are indeed consistent on the two occasions.

Interrater agreement concerns the agreement between different raters who rate the same free responses. An *interrater agreement* study is done to assess the agreement between different raters, who rate the same free responses. The raters are adequately trained to do the rating job. The raters work completely independently, which means that they do not confer with each other, and that they do not know each other's ratings. Table 2.3 gives the (hypothetical) frequencies and proportions of two raters (A and B) who independently rated the free responses of 100 test takers to the same item.

Table 2.3 *(Hypothetical) frequencies and between parentheses proportions of two raters' (A and B) ratings of the free responses of 100 test takers to the same item.*

| | | **Rater B** | | | |
|---|---|---|---|---|---|
| | | Correct | Partly correct | Incorrect | |
| Rater A | Correct | 50 (.50) | 15 (.15) | 5 (.05) | 70 (.70) |
| | Partly correct | 10 (.10) | 9 (.09) | 1 (.01) | 20 (.20) |
| | Incorrect | 5 (.05) | 1 (.01) | 4 (.04) | 10 (.10) |
| | | 65 (.65) | 25 (.25) | 10 (.10) | 100 (1) |

The table shows, for example, that both raters rated 50 responses as correct, Rater A rated 15 responses as correct, whereas rater B rated these responses as partly correct, and Rater A rated 5 responses as correct, whereas Rater B rated these responses as incorrect. The diagonal of the table contains the frequencies and proportions of ratings, where the two raters completely agree: correct/correct: 50 (.50); partly correct/partly correct: 9 (.09); incorrect/incorrect: 4 (.04). The observed proportion of identical ratings of the two raters is:

$O = .50 + .09 + .04 = .63.$

As for intrarater consistency, the observed proportion of identical ratings may give a wrong impression of rater agreement. The chance expected proportion of identical ratings is computed in the same way as for the intrarater consistency:

$$E = .70 \times .65 + .20 \times .25 + .10 \times .10 = .515.$$

The value of coefficient kappa for interrater agreement is computed using Definition 2.1:

$$Kappa = \frac{.63 - .515}{1 - .515} = .24.$$

This value is low, and indicates low agreement between the two raters.

Coefficient kappa was demonstrated using a scale of three categories (correct, partly correct, and incorrect). The coefficient is computed in the same way for other scales, for example, two categories (correct and incorrect) or five categories (A, B, C, D, and E).

Coefficient kappa focuses on complete consistency or agreement (e.g., correct/correct, partly correct/partly correct, and incorrect/incorrect). It ignores the degree of inconsistency or disagreement, for example, correct/incorrect is a greater degree of inconsistency or disagreement than correct/partly correct. The weighted coefficient kappa (Cohen, 1968) is a coefficient that uses the information on the degree of inconsistency or disagreement. However, the main interest of most intrarater consistency and interrater agreement studies of item responses is in identical ratings. Therefore, coefficient kappa is generally the right coefficient to assess the degree of consistency and agreement.

Rating pilot studies yield important information that can be used for improving training, removing items or removing raters. Pilot studies can reveal the following:

1. Many raters are inconsistent in their ratings of many different items. An explanation is that the rating procedure is inadequate. If improvement of the procedure does not lead to sufficient consistency, the items are probably unsuitable.
2. Most of the raters are consistent on most of the items, but some of the raters are inconsistent on most items. An explanation is that the inconsistent raters do not understand the rating instructions. If extra training does not lead to sufficient consistency, the inconsistent raters should be removed from the pool of raters.

3. The raters are consistent on most of the items, but inconsistent on some of the items. An explanation is that the rating instructions are not clear for the inconsistent items. If improved instruction does not lead to sufficient consistency, the inconsistent items should be removed from the set of concept items.
4. Raters disagree on many items. An explanation is that the rating procedure is inadequate. If improvement of the procedure does not lead to sufficient agreement, the items are probably not suited to measure the construct.
5. Most of the raters agree on most of the items, but some of the raters disagree on most of the items. An explanation is that the raters who disagree do not understand the rating instructions. If improved training does not lead to sufficient agreement, the raters who disagree should be removed from the pool of raters.
6. The raters agree on most of the items, but disagree on some of the items. An explanation is that the rating instruction is not clear for these items. If improved instruction does not lead to sufficient agreement, the items should be removed from the set of concept items.

### 2.11 Compiling the first draft of the test

The concept items that survived the pilot studies are used to compile a concept version of the test that includes instructions for the test takers. Usually, the instruction contains some example items that test takers have to answer to ensure that they understand the test instructions. The concept test may consist of a number of subtests that measure different aspects of the ability or achievement.

The conventional way of assembling a maximum performance test is to start with easy items and to end with difficult items. Test takers may be anxious or nervous, and easy items at the start of test administration are used to set their minds at rest. Some test takers take a long time to answer the items. Difficult items at the beginning or in the middle of the test may mean that they spend too much time on these items. Therefore, the difficult items are put at the end of the test.

The concept test is submitted to a group of experts. The group can be the same as the group that was used in the experts' pilot study on item quality (Section 2.10.1). The group has expertise in (1) the content of the ability or achievement, and (2) test construction. The experts evaluate two different

properties of the concept test. First, they judge whether the test instruction is sufficiently clear for the population of test takers. Second, they study whether the test yields adequate coverage of all aspects of the ability or achievement being measured by the test. This part of their work is called *content validation* of the test. Finally, if necessary, the sensitivity panel judges whether the test is balanced with respect to multicultural material and references to gender (Ramsey, 1993).

The comments of the experts are used to compile the first draft of the test. This section described the conventional way to compile tests. Recently, new methods for compiling tests have been developed that are appropriate for advanced applications. Modern methods for compiling tests are introduced in Section 14.2 of this book.

The first draft of the test is administered in a try-out to a sample of at least 200 test takers from the population of interest. The try-out data are analyzed using methods of classical and modern test theory, which are discussed in Part II of this book.

## 2.12 Software

Cohen's kappa, weighted kappa, and their confidence intervals can be computed using the program AGREE that is distributed by SciencePlus (http://www.scienceplus.nl).

*Exercises*

2.1 Write three short-answer items (one recall, one understanding, and one application) for an achievement test on Chapters 1 and 2 of this book.

2.2 Write three three-choice items (one recall, one understanding, and one application) for an achievement test on Chapters 1 and 2 of this book.

2.3 Discuss these six items with a participant of the course, and revise the items using his (her) comments.

2.4 Two raters (I and II) independently rated students' essays in five categories (A, B, C, D, and E). The rater x rater frequency table is:

| | | Rater II | | | | |
|---|---|---|---|---|---|---|
| | | A | B | C | D | E |
| Rater I | A | 10 | 5 | 2 | 1 | 0 |
| | B | 4 | 20 | 5 | 2 | 0 |
| | C | 1 | 5 | 18 | 4 | 1 |
| | D | 0 | 1 | 4 | 8 | 2 |
| | E | 0 | 0 | 1 | 2 | 7 |

Compute Cohen's coefficient kappa for interrater agreement.