Uživatelka:Iannadecz/Pískoviště/Bookpredupravoucitaci

Testing and Assessment of Studens in Higher Education

Contents

1 Introduction

1.1 Foreword

In the world of academia, few topics are more passionately debated than modern methods of assessing students' knowledge. Assessment is a form of communication between teacher and student. It is an opportunity to communicate what the teacher considers is important to learn.

A fair, demonstrable, and objective assessment of students' knowledge and skills is the basis for assessing student performance, the quality of education, and motivates students to continue their studies.

Since tests, as a component of assessment, have a significant impact on student learning, it is important to align tests with educational and learning objectives.

When it comes to high-stake tests, which, for example, can determine whether a student goes on to further study, it is important to ensure the accuracy and reliability of these tests. This book provides a summary of procedures, recommendations, and methods for creating quality didactic tests with subsequent analysis of their results. The book is primarily aimed at members of the academic community responsible for student assessment and is intended for use as a practical tool to help the teacher throughout the process of planning, developing, employing, and evaluating tests.

The authors aim to provide readers with a practically applicable methodology and tools for compiling and evaluating their own tests. They describe with test items, their organization in item banks and the topic of item and test security.

Methodologies and tools for the preparation and evaluation of tests designed to objectively assess learning outcomes are common across disciplines. The text includes selected passages from publications listed in the bibliography. The primary resource was the handbook Testování při výuce medicíny [Assessment in Medical Education], published in Czech language by Karolinum [1].

The text is organized to be used as a guide through the process of test preparation and to enable the reader to become oriented in the issues. Those interested in a deeper study of the topic can consult the articles and books listed in the reference section.

1.2 The Role of Testing in Higher Education

The quality of a college or university is closely tied to the quality of its students, which is why institutions of higher education strive to select the best applicants to admit as students. These institutions then work to prepare their students for what they will face in the working world. At the same time, these institutions must assess the effectiveness of the educational process and examine how well-prepared students are for their future roles. Obviously, the success of graduates in practice is the best measure of the effectiveness of education. However, while this metric may be the most objective, the delay between teaching and its evaluation would make it impossible to maintain effective feedback. To be able to measure educational outcomes in a shorter period, we must choose other ways. One of these is the testing of learning outcomes, which, if possible, objectively, reproducibly, and fairly assess the level of achieved knowledge and skills.

I need to test. How do I go about it?

If you need to begin testing:

Read this entire book,

study the essential resources on which this book draws,

learn to use the statistical language R,

… and you'll never even get around to actual testing.

Overall, the test cycle is a complex process in which you can constantly improve individual steps. You will eventually become a recognized expert in psychometrics. So: if you need to start testing, start now. And you can start just by writing test items. Turn to the chapter in this book that will help you. Your own practical experience will move you forward more than the best theory.

1.3 Types of Tests

For a test to work as we envision it should, we must first clarify what we expect from it. Different types of tests are used in different situations. While all tests share some common traits, each test emphasizes certain qualities while overlooking others.

Depending on which phase of learning the test is a part of and what it is intended to aid in learning, we can divide tests into formative and summative. The purpose of formative assessment is primarily to provide feedback to students and teachers about the progress of the lesson. The test becomes part of the teaching dialogue, supports the active involvement of students in their learning, and contributes to their motivation. The student finds out to what extent his or her knowledge and skills correspond to the course's requirements and tries to use them. The test will help the student identify their strengths as well as areas that still need work. The results of formative testing help the teacher to make teaching more efficient, as they show the teacher which areas need more attention and where, on the contrary, further attention is unnecessary. For both parties, formative testing should primarily be indicative. A formative test is not subject to very demanding procedural requirements. In some cases, the imperfection of a formative test can even help education by stimulating discussion and involvement of all those involved.

In contrast, the objective of summative assessment is to provide an overall picture of the learning outcome. The results of summative tests are often the basis for further steps in studies or career. This type of assessment is most often undertaken following the completion of some integral part of the learning or at the end of a course, or, conversely, it can be used to verify an applicant's ability to enroll in the course or to start performing a certain job.

In practice, purely formative and purely summative tests are the extremes on a continuous scale. We often come across the fact that even the results of a test that is primarily formative, are in some way included in the student's overall assessment, and further progress is conditional on the achievement of certain results. Conversely, even summative tests and exams should, in most cases, provide feedback to both the student and the teacher, helping to improve the quality of the course and develop study skills.

When preparing a test, it is also necessary to consider to what extent and to what depth the acquired knowledge and skills are to be evaluated.

In this regard, the most demanding are proficiency tests – tests and examinations of (professional) competence, which assess the overall ability to perform an activity, for example, to communicate in a foreign language, to perform certain tasks, etc. Professional competence tests usually require workplace-based assessment and separate written testing can only be used in specific cases or as a component of the aptitude test.

Achievement tests evaluate to what extent the student has mastered a part of the course or a certain section of study.

The aim of diagnostic tests is to describe in more detail the test subject’s strengths and weaknesses.

Finally, prognostic tests and aptitude tests are intended to estimate the extent to which the tested person will be able to successfully complete a certain course and acquire the target competencies in it. For example, the Modern Language Aptitude Test (MLAT) measures a student's potential to successfully master foreign languages, the Scholastic Aptitude Test (SAT) assesses academic ability and the potential to graduate from college or university.

Standardized test design and evaluation are intended to ensure provability, reproducibility, and long-term reliability of the results of the most important tests. With a standardized test, it must be guaranteed that the result depends primarily on the skills of the examinee, not on the specific test version, the environment in which the test is taken, the supervisor or the evaluators.

In psychometrics, the term “standardization” has a number of meanings, and these will be discussed in more detail later, in a separate chapter. Standardization of a test is required especially where the result of the test is a recognized certificate, or the outcome is important for the further career of the test taker. Standardized testing includes the collection of testing data and its statistical processing with the aim, among other things, of detecting “non-standard phenomena” (copying, item leaks, hinting...). One of the basic tools used when preparing standardized tests are calibrated test items, with psychometric characteristics combined so that the test as a whole has the desired properties. Comparability between individual test runs is checked using anchor items, which make it possible to compare the difficulty of tests given on different dates. Standardization also includes the objective setting of the threshold for passing the test and the consistent ensuring of comparable conditions for all those tested. Standardized tests must be prepared and performed in such a way that the objectivity of their results can be proven, even in court. The requirement for standardization means increased costs. It is therefore always necessary to consider where these costs are justified and where it would be sufficient to use common, non-standardized examinations and tests.

In a non-standardized test, the specific examiner or evaluator plays a more important role. They often focus more on the individuality of the examinee and can better assess his or her personal talents and achieved competences. However, this approach is not suitable for comparing examinees with each other.

In situations where standardization would be inexpedient or even unfeasible (for example, for a too small number of examinees), steps are often taken that lead to the impartiality of a non-standardized assessment and thus to the reduction of undesirable effects on the assessment, especially the subjectivity of the examiner.

In person tests have always been the primary form of testing, where students are in direct contact with the instructor. In recent years, due to the pandemic, remote tests and exams have become more prevalent, with development in forms of testing that do not require direct contact between teacher and student. Major progress has been made in proctored testing methods, and open book testing is also being developed.

1.4 Limitations of Testing

Although education at universities and higher education institutions in general has been around for centuries, there is still no clear consensus on its actual objective [2]. There are probably multiple such objectives, and it also depends on the focus of the university. In general, we can probably say that a university graduate should leave as a professional prepared for the performance of a certain profession, occupation, or role. The traditional idea is that this requires the acquisition of knowledge and numerous skills. However, this alone is not enough. A university graduate should be able to work more or less independently, i.e., perform certain activities. After all, the authorization to perform a certain profession is often linked with obtaining a university degree.

The performance of a specific activity often requires more than simply acquiring a range of factual knowledge. You need to understand a certain area, have certain skills, but also adopt a professional attitude. If we are to guarantee that a university graduate is capable of professionally performing work in the given field, we should verify that he is sufficiently competent not only in terms of knowledge, but also in the corresponding “higher” levels.

1.4.1 Dimensions of Knowledge, Skills, and Attitudes

When describing the learning objectives, for example in the annotation of a subject or its division into a syllabus, their description is usually based on the substantive content of the subject – a list of topics that we want to teach. But this kind of division is not enough on its own. For high-quality learning, and then also for the assessment of its results, it is useful to add a second dimension to the list of thematic areas – to divide each topic into several levels according to the complexity of the educational objectives.

The most used model that describes the complexity of the objectives of education, upbringing and professional training is Bloom's taxonomy. It is actually not one, but three hierarchical models (also referred to as domains):

The cognitive (knowledge-based) domain,

The affective (emotion-based) domain, and

The psychomotor (action-based) domain.

In education, one works most often with the first domain, i.e. the domain of educational objectives, which relates to knowledge, and the ability to engage and use it. However, many authors repeatedly point out that the other two areas are also an equally important, albeit harder to grasp, part of education. Education or upbringing in the emotional domain leads to the development of professional attitudes. The psychomotor domain then includes the acquisition of practical skills.

Several versions of Bloom's taxonomy are currently in use. In the 1990s, it was extensively revised, and a two-dimensional map was created. Instead of three domains, this revised version works with four knowledge dimensions – factual, conceptual, procedural, and metacognitive knowledge. Each of these knowledge dimensions then contains six dimensions of cognition: remember, understand, apply, analyze, evaluate and create.

Given that the term Bloom's taxonomy refers to several different concepts and versions, we’ve chosen to work in more detail with only one of these, traditionally referred to as Bloom's taxonomy of educational objectives. It more or less corresponds to the cognitive domain of Bloom's original taxonomy from the turn of the 1950s and 1960s and the dimension of factual knowledge from the revision mentioned above.

1.4.2 Bloom's Taxonomy of Educational Objectives

The taxonomy within the scope of learning objectives describes the levels of competences and skills relating to factual knowledge, its comprehension and understanding of context. It has six levels: the lowest is knowledge, then understanding, application, analysis, evaluation, and the highest is creation. Traditional education tends to work precisely with this set of objectives, primarily with its lower levels.

Knowledge (also remembering)

The lowest level of learning objectives is knowledge, i.e. remembering and the ability to recall facts and the most fundamental concepts. The student learns basic terms and definitions. We mainly want the student to state something, repeat something, explain a concept, classify something in a certain classification scheme, etc., which the student can do even without a full understanding of the concepts he or she is working with – for example, a student can correctly classify a plant species in a family solely thanks to the fact that the student has learned which species belong to the respective family. At the same time, the students do not need to have any idea what the plant they are talking about looks like, what the characteristics of the given family are, and why the given plant actually belongs to that family. Knowledge at this level often has the character of isolated data from an encyclopedic dictionary.

Comprehension (also understanding)

Accomplishment of this educational objective can typically be demonstrated as the ability to explain or interpret certain material. A student who has achieved understanding is able to explain the main idea. With a grasp of the material, the student can also describe, compare, sort, or translate something into another language that he or she speaks, etc. Although the student understands the given subject, he or she may not yet be able to put it to use – typically this is reflected in the fact that the student cannot combine it with other knowledge that he or she also has and understands. For example, the student comprehends the movement of the Earth around the Sun and its connection with the changing of the seasons, as well being well familiar with the shape of the Earth, but he or she cannot answer the question of what the position of the sun is above the horizon in different geographical regions at the time of the summer solstice.

Application

A typical indication of having achieved this level is the ability to use acquired knowledge in new situations. The student can solve questions and problems that he or she has not encountered before. To do this, they must recognize connections and relationships and find a way to use them to solve problems. Often, the student must use rules or procedures that they already know, but in a different way than they have up to now. However, the student is only able to work with relatively uncomplicated assignments. This level is not sufficient for solving a complicated question, because, for example, the student does not yet distinguish between data essential for the solution, insignificant data, and data that is possibly missing and must still be acquired. A more complicated problem is therefore unsolvable at this level, because even though the student has the necessary component knowledge, he or she becomes lost in it.

Analysis

In Bloom's taxonomy, analysis means breaking down a complex whole or problem into smaller parts, which will enable a better comprehension of it. Analytical skills are needed, for example, to distinguish cause and effect, or to find evidence that supports some generalizing statement. This level also includes the ability to recognize the structure of some information, break it down into individual components, assess the relationships between them and, thanks to this, estimate the credibility of the information source. The achievement of this educational objective can also be demonstrated, for example, with a thought experiment in which the student can estimate how a certain intervention would change a certain event. To do this, they must analyze the action, recognize its components and the links between them, determine what exactly would change with the intervention under consideration and what consequences it will lead to.

Evaluation

Evaluation is one of the highest objectives in Bloom's taxonomy. It refers to the ability to evaluate information and, based on this evaluation, to make informed decisions, take positions or defend opinions. Evaluation requires the ability to first analyze information and is therefore linked to the preceding objective. Fundamentally, evaluation aims to assess the individual parts of the information and judge their significance and validity. The resulting verdict, meanwhile, should be key to solving a certain problem or creating something new. A typical question that demonstrates the achievement of this objective can be, for example, the preparation of a professional review of a scientific article, including recommendations for its publication, rejection, or modifications. Clearly, accomplishing such a question requires not only the achievement of all previous objectives, but also some creativity. It is not surprising then, that the order of the two highest objectives sometimes differs in different versions of Bloom's taxonomy.

Creation

The ultimate objective that a student can achieve through learning is creation. It is an expressly creative, productive objective. Achieving this objective allows the student to propose a new original solution, design something, invent something, etc.

Bloom's taxonomy forces us to think about how people learn, which is also valuable when considering how to assess learning outcomes. Although this is undoubtedly the most commonly used “compartmentalization” of the educational process, it still has its drawbacks and critics. The previous description makes it clear that while the nature of the lower categories of Bloom's taxonomy is quite unambiguous and easily applicable in practice, the higher levels are prone to increasingly abstract definitions, are more ambiguous, and there are even doubts about how exactly to organize them. Bloom's Taxonomy suffers from multiple limitations [3][4][5], for example:

Bloom's Taxonomy assumes that a person learns in a linear, sequential fashion – that he or she progresses from the most basic objectives to the higher ones. In reality, this is not the case. A learner can jump between individual “levels” and repeatedly return to lower categories. If a person learns to evaluate or create something, then he or she reanalyzes the result of his or her work, adds to his or her knowledge and learns to understand it, etc.

A person can also learn from the top of the pyramid towards its base, by creating something. In this case, other processes become applied, ones that Bloom's taxonomy does not describe very well: research, trying out a solution, creating a prototype, revision or critical assessment. Only these activities then force one to seek out sources and acquire new factual knowledge. [6]

Bloom's Taxonomy assumes that a person learns in isolation from others – it is individualistic. It overlooks the social and connectivistic aspects of learning, which are very important in higher education. [7]

1.4.3 Other Taxonomies

Bloom's Taxonomy is relatively complex – recall that in the text above we only worked with one of the knowledge dimensions, so we completely omitted everything related to procedural skills or attitudes. And this, together with its incompleteness, led to the emergence of other, variously understood models of learning and levels of competences achieved. [8][9]

In the 1990s, specifically for the assessment of knowledge and skills, the so-called Miller's pyramid began to be used in medicine [10], and it then gradually spread to other fields [11] [12]. The original four levels of assessed competencies were later supplemented by a fifth level [13].

Knowledge (the student is on the level of “knows”).

Competence (student knows how). The examinee can integrate the knowledge from the previous level into the current context.

Performance (shows how). The skill is already comprehensive, the examinee “is self-oriented in” and combines a wide range of knowledge and skills, which he or she often acquired in various subjects and areas of study.

Action (in practice, the student performs all the necessary actions correctly; action, does). This level should be reached, for example, by a candidate for the state final exam.

Identity (he or she is a true professional). A person who has reached this level can consistently demonstrate the attitudes, values and behaviors expected of a representative of a certain professional group. It can be said that the given person thinks, behaves and feels like a doctor, teacher, lawyer, designer, etc.

If we accept that a university graduate should be a professional prepared to perform a certain action, we should verify during the course or at the end of the studies that they have actually acquired the relevant competencies. While knowledge and understanding can usually be assessed very well using both standardized tests and non-standardized methods (such as oral examinations), assessing the highest levels of competence is more complicated. We would certainly consider it absurd if, for example, an orchestra hired a violinist based on a written test or an oral exam on violin playing. It would be equally absurd to claim that someone is, for example, a well-prepared teacher, lawyer, historian, or doctor based on a simple written test or oral exam. This would verify with greater or lesser credibility that he or she has acquired the knowledge and understanding that are a necessary prerequisite for the performance of the profession, but we would not ascertain whether he or she can also apply the knowledge in an appropriate manner, whether he or she has acquired the necessary skills and whether they can actually perform the activities that the specific work involves.

Knowledge and understanding can be accurately assessed using a well-prepared and standardized written test. Assessment of skills by written test, on the other hand, is only possible in specific cases. We can use a written test if the skill is, for example, the ability to solve a mathematical problem or to describe some reaction using chemical equations. However, most skills cannot be assessed by written test or oral exam – it would be difficult to test, for example, laboratory tasks, practical work with surveying instruments or blood sampling in this way.

To verify skills and activities, it is possible to use a technique generally referred to as practical examination or workplace (based) examination [14], [15]. As a rule, these are methods that can be standardized only with great difficulty. For items that are tested in practice, it is simply not possible to ensure identical (standardized) conditions for a larger number of candidates. Moreover, these methods often assess skills that cannot be fully classified in a standardized fashion, such as communication with a client or teamwork.

However, even these practical tests can be made objective, i.e., arranged in such a way as to suppress the influence of undesirable factors – for example, the subjectivity of the examiner or the variability of the conditions under which the test takes place. The next step is the validation of the practical tests, i.e., verification that the test result truly reflects the skills acquired, which are needed in real world practice.

Methods that make practical testing more objective tend to have several typical features:

Long-term or repeated performance is monitored rather than one-time performance within a single test session. If the practical exam takes place in a short period of time (e.g., during a single day), it is divided into several separate parts (often referred to as stations). Each of them is assessed by different assessors and each is focused on a different range of skills and activities.

The examinee is evaluated independently by a larger number of evaluators. The evaluators include experts in the given field, but often also other persons – for example, classmates, role players who conduct model communication with the examinee during the exam, and sometimes even technical staff.

The exam and its evaluation are structured, i.e., the evaluators comment on the monitored aspects of the performance (so-called exam rubrics) according to predetermined criteria.

The exam is validated as a whole, or its individual parts are validated.

A major shift in practical examination formats was brought about by the introduction of so-called objective structured clinical examinations (OSCE) in medicine in the mid-1970s. Twenty years later, this approach began to be used in other fields as well, which is why it is sometimes referred to as objective structured practical examination, OSPE. During the OSCE, students go through a series of stations in which they are confronted with common situations arising in everyday practice and are to perform a certain procedure. They are evaluated using a structured questionnaire, which is filled in by the attending evaluators. More weight is given to steps and procedures of a more general nature (in medicine, for example, the prevention of the spread of infection, communication with the patient during the procedure, patient instruction and explanation of the procedure, etc.), while less weight is given to actions that are narrowly specific.

Another way to assess the achievement of higher educational objectives is to compile portfolios. A portfolio is a systematically created set of samples of a student's work that demonstrates their efforts, progress, and achievement of educational objectives throughout the course or curriculum [16][17][18][19]. It is a distinctly constructivist assessment tool. Its advantage is that it faithfully reflects the achievement of the highest, creative educational objectives, as well as the adoption of professional habits and attitudes, value rankings, etc. On the other hand, portfolio is a highly individualized evaluation tool that does not allow for standardization, and it is also difficult to increase its objectivity. Compared to other tools, portfolios are quite time-consuming for both the student and the teacher. Implementing portfolio assessment requires careful preparation and seamless integration with the curriculum. [20]

1.4.4 Quantitative and Qualitative Forms of Evaluation

In the preceding text, we mainly approached the evaluation of education results as a measure of the extent of the student’s achievement of the expected knowledge and skills. Taken this way, the result of the assessment is a certain quantity, grade, or numerical value. The charm of this concept lies in its comprehensibility and in the fact that the validity of the conclusions we draw can be easily examined by scientific methods. We can talk about the accuracy and reliability of such conclusions, where we can quantify the degree of accuracy by statistical methods, we can even measure the degree of uncertainty with which we communicate the result. We can also track the impact of every change we make in teaching and testing. Altogether, it allows us to standardize exams – to ensure that assessment results are substantiated, reproducible, objective and valid.

However, even the hierarchization of learning objectives according to the most common concept of Bloom's taxonomy already shows that standardized testing and examinations cannot capture the entire breadth of higher education. We have shown that it can only capture lower levels of cognition. Higher educational objectives can be evaluated with non-standardized, but still objective methods, and the result can still be of some value. However, for the evaluation of the most complex objectives, skills and attitudes, a simple one-dimensional expression is not enough.

Comprehensive competence, attitudes, behavior, or the achievement of professionalism in a certain area cannot be measured by a number. Although they are not measurable (or at least cannot be expressed by a one-dimensional quantity), they are describable. They can be described by verbal evaluation. This, however, always has a subjective component, and is typically a non-standardized assessment.

Approaches to evaluating educational outcomes are constantly evolving. Over the past few decades, there has been intensive development in the area of standardized assessment, and this method has become an integral part of education in developed countries. Thanks to the verifiability and defensible nature of the results, it has gradually completely displaced non-standardized methods in many areas. In recent years, however, the limits of such an approach are being pointed out [21]. Standardized methods remain the best-known tool in certain stages of learning, but some authors caution against these methods being the only tool used [22] [23]. Non-standardized methods are not inferior, nor are they superior to standardized ones. They are merely different tools, each of which is suitable for something different.

2 Planning a Test

A test should provide a good assessment of the results of the learning and, as such, is often a part of the learning process. For the preparation of learning and its subsequent evaluation, it is therefore necessary to be able to define as precisely as possible what the graduate should be able to do, i.e. what the learning objectives are. In practice, however, actual learning only approaches these objectives. In some areas, students even entirely fail to achieve the learning objectives, but at the same time they often acquire other, unplanned, knowledge and skills. The set of competencies the graduate actually acquires are called learning outcomes.

In terms of content and scope, a test we use to verify whether the graduate has acquired the desired knowledge and skills, should correspond as best as possible to the learning objectives, and at the same time, the outputs of the learning should correspond as much as possible to the test. In practice, this will never be a perfect match. To make it as good as possible, it is necessary to properly plan both the learning and the test. If students are aware of the objectives they need to achieve and how their performance will be assessed, they are more motivated to study [24]. If, on the other hand, the content of the test or exam differs from the actual content (i.e., the outputs) of the learning, students feel cheated and label the testing as unfair, often replacing the motivation to learn with the motivation to simply get "the answers required on the test".

Another reason to carefully plan tests is the fact that we often cannot test all students at once and need to create multiple parallel forms (versions) of the test. If we follow the same, sufficiently detailed plan it becomes relatively easy to achieve equivalence of all the versions.

2.1 Blueprint

One of the best proven methods of preparing a test plan is the construction of a so-called blueprint [25]. The first step is to create a spreadsheet with rows that match the content objectives. This step is relatively simple – it is mostly based on the course syllabus, the order of chapters in the primary textbook, etc. Each line corresponds to one topic. It is advisable that the breakdown of topics is sufficiently detailed – one lecture, lesson or chapter in the textbook usually corresponds to several lines with partial subtopics.

The columns of the blueprint (specification table) correspond to the aspects (more precisely, the learning domains) from which we can look at the topics [26]. Finding views that can describe most of the content objectives (i.e., table rows) at once is key, and often the most difficult step in creating a blueprint. The most general advice is to base it on the objectives of Bloom's Taxonomy, e.g.:

Knowledge – for example, knowledge of terminology, definitions, naming of a certain phenomenon;

Understanding – for example, comparing, interpreting graphs and data;

Application – for example, solving a new problem using an analogy;

Analysis – for example, identification of causes and consequences, ability to explain a certain phenomenon;

Synthesis – for example, predicting the outcome of an event, estimating the consequences.

Since the domains created according to Bloom's taxonomy are relatively detailed, and because there are just too many of them for the creation of shorter tests, sometimes simpler diagrams are used, e.g. acquiring knowledge – application – problem solving. In some fields, common aspects naturally follow from the material taught and are relatively easy to find, e.g. in clinical fields of medicine, the columns can often be labeled etiology – symptoms – diagnosis – treatment – prognosis, etc.

In any case, the blueprint should be constructed so that as many combinations of rows and columns as possible make sense. Individual fields then contain the planned number of items, the type of items or the method of testing. It is not necessary to fill in all the fields of the blueprint, but the test plan should be balanced – there should not be any empty or almost empty row or column, and no larger blank areas should remain.

When filling out the number of items, we already take into account the total scope of the test. It may happen that some fields of the blueprint, corresponding to less essential knowledge and skills, will be used only in some versions of the test, and other versions will contain other items instead. In any case, a test plan constructed in this way will help ensure that the number of items devoted to a certain topic reflects its importance and that all aspects of a certain problem are adequately tested [27].

The two-dimensional blueprint is often very detailed and extensive, so it resembles a technical drawing in its dimensions – that's why the term blueprinting was used for this method of test planning. The detailed blueprint is usually not public, it is used only by the test creators. Its publication could lead to the speculative behavior of test takers, who would concentrate their preparation more on the test itself than on acquiring the knowledge and skills needed for practical work [28]. On the other hand, the content of the test (i.e. the lines of the blueprint) should always be published, as well as the share of each area in the total scope of the test.

3 Test Items (Questions)

Items, or questions, are the basic building block of every test. We will deliberately avoid the frequently used term question, which, as we will see below, can also have a different, narrower meaning.

In general, skills can be assessed directly or indirectly. During direct examination, the student is given the task of directly performing a certain activity. Direct testing often uses practical testing techniques (workplace-based assessment), but written tests can also include some direct questions.

Examples:

Create your structured resume in English.

In the R programming language, create a function that...

Direct items make it possible to assess the extent to which the candidate has achieved the target competencies and how prepared they are, for example, for the performance of a certain work activity. Their disadvantage is difficult grading. The candidate's performance must be evaluated by several evaluators, and usually several areas or aspects of the performance are evaluated. Evaluators must be trained in advance, and they use a structured evaluation form for assessment.

More common are indirect testing methods. The student's skills are not tested directly but are assessed on the basis of knowledge and skills that are a prerequisite for a particular ability.

A typical written test tries to assess the achievement of target competencies indirectly. The authors of the test create a certain construct of which knowledge and skills are essential for the achievement of competence. By testing for these, they try to estimate whether the examinee could achieve the target competence. From this perspective, we are not examining whether the examinee can actually do a certain thing, but whether he or she has the prerequisites to do it.

It is not possible using indirect testing methods alone to reliably decide whether a candidate is capable of independently performing a certain activity, and it is not possible to completely replace direct methods with these. However, these methods are much faster, easier to evaluate, cheaper, and can more readily achieve reproducibility.

There are basically two types of indirect test items (Chvál 2015):

Open-ended Items

The examinee must create an answer – write a text, perform a calculation, draw a picture, etc. The answer can have a range of formats and lengths – on the one hand, it can be about filling in the missing letter, on the other about writing a multi-page essay. Since the examinee creates their answer, these items can also be characterized as productive.

Closed-ended Items

The examinee chooses a solution from a finite range of options that have been offered to him. He or she does not create the solution – their item is only to select and mark it. Most often, they choose between several possible answers to a certain question. However, this also includes other types of items in which the examinee chooses from several predefined alternatives, e.g. they have to match related concepts to each other, arrange items in a certain order, fill in i or y in the missing places in the text, or decide whether in a quantity will decrease, increase or remain unchanged in a certain situation.

The line between open-ended and closed-ended items can sometimes be blurred. In some cases, the examinee must choose an answer from a practically unlimited range of possibilities, without it being classified as a productive item. An example could be items in which the student has to mark (i.e. select) a certain place on a photograph of a microscopic preparation.

Both open-ended and closed-ended items have their irreplaceable place in testing and in university learning, while each is suitable for something different:

Open-ended items make it possible to evaluate more complex skills, especially skills of a productive, creative nature (Schindler 2006). To formulate an answer, students are forced to actively use professional terminology. It is often possible to follow the thought process that led the test taker to the solution. During the evaluation, it is possible to recognize how well the test takers understood the assignment and whether the item is not poorly formulated. Open-ended items are an invaluable tool for continuous formative testing during learning, which is primarily intended to provide feedback to students and teachers. They provide feedback more effectively than multiple choice items and are well suited for use as a starting point for discussion of the topics covered. Open-ended items can also be part of the final summative tests, in which they are used to verify the achievement of not only component knowledge and understanding of the learned facts, but also the use and incorporation of this knowledge in more complex items. However, the preparation of open-ended items for summative tests is demanding, as is their evaluation.

Open-end items cannot be graded automatically – the answers must be assessed by qualified evaluators. It may thus happen that the assessment is plagued by subjective error. Each open-ended item is successively evaluated by several mutually independent evaluators, who receive them anonymized if possible. Very detailed instructions are prepared in advance for the evaluators. Nevertheless, examinees may challenge the objectivity of the test and it may be more difficult to justify the evaluators' decisions beyond doubt. Therefore, if open-ended items are to be used in a test of fundamental importance, these items and the rules according to which they are scored need to be prepared very carefully. Open-ended items are often evaluated on a wider scale than just “correct/incorrect” and all evaluators must equally assign partial points for partially correct solutions. In general, the more open-ended the item, the more difficult it is to ensure its objective assessment. It is also necessary to have procedures in place in the event that the evaluators' opinions on a specific solution differ. Therefore, the preparation of open-ended items tends to be time consuming and its difficulty increases with the importance of the test.

Open-ended items can place students with communication impairments at a disadvantage, since the formulation of an answer often affects the evaluation.

Closed-ended items include multiple choice, pairing and arranging items (Schindler 2006). They are easier for processing, tests can often be evaluated automatically by computer, or possibly they can even be graded by a less qualified worker. In most situations, closed-ended items are the fastest and most effective tool to find out how much knowledge the student has acquired and how well they have understood the subject matter. To a limited degree, these items can be used to evaluate even the mastering some simple skills.

The big advantage of closed-ended items is that it is easy to decide whether the examinee responded correctly. As a result, the test evaluation is reproducible. Scoring the test is also very fast. The answers do not depend on the formulation skills of the examinee, their graphomotor skills, their typing speed, etc. On the other hand, closed-ended items do not allow testing of many types of skills. Closed-ended items place students who are less attentive or function less accurately under stress at a disadvantage.

3.1 Multiple Choice Items

Multiple-choice items predominate in written tests today. Their main advantage is that they are easy to evaluate. As we shall see below, they can take a number of forms. All have in common the fact that the examinee chooses one or more answers from the options offered. How the options are offered—checking a “radiobutton”, "checkbox", or selecting from a drop-down menu—is not decisive.

When it comes to properties and use in tests, it is important to divide multiple choice items—regardless of their formal appearance—into two groups:

Dichotomous items (TRUE/FALSE items)

One offered option (or more of them) is completely correct, the others are completely wrong.

Example:

Indicate whether the statement is true:

A fin whale is a mammal that lives in the sea TRUE – FALSE

Scoring dichotomous items is simple. Most often, one point is awarded for a correct answer, nothing for an incorrect one. Less common are scoring schemes in which other numbers of points are assigned, e.g. the point gain is weighted according to the difficulty of the item, or points are deducted for an incorrect answer.

The disadvantage of individual dichotomous items is that by simple guessing you can get an average of 50% of the maximum possible score. At first glance, this may not be a problem if the threshold that the student must reach in order to pass the test is correctly set. However, this reduces the discriminative power of the test. Therefore, some authors recommend various modifications of dichotomous items, for example requiring that, along with each “FALSE” answer, the student states how the question would need to be changed to get a "TRUE" answer [29]. This actually creates a combination of a multiple-choice question and an open-ended question.

Bundles of dichotomous items

(Also referred to as: multiple true/false, MTF; multiple response question, MRQ.)

Sometimes, several dichotomous items are combined into a bundle with a common core.

Example:

A dog is a popular pet. There are many breeds that vary in size, color and temperament. Which statement about dogs is true?

a) Some dog breeds have no hair at all. TRUE FALSE

b) Regardless of size and color, all dog breeds belong to a single biological species. TRUE FALSE

An important feature of dichotomous item bundles (MTF) is that the examinee has to decide on each statement independently, without regard of the other statements in the bundle. In other words, we can break down the above set of dichotomous items into two separate dichotomous items:

Item 1:

A dog is a popular pet. There are many breeds that vary in size, color and temperament.

Indicate whether the statement is true:

Some dog breeds have no hair at all. TRUE FALSE

Item 2:

A dog is a popular pet. There are many breeds that vary in size, color and temperament.

Indicate whether the following statement is true:

Regardless of size and color, all dog breeds belong to a single biological species. TRUE FALSE

The formal appearance of dichotomous items and their sets can vary. Most often, a TRUE/FALSE or TRUE/FALSE answer is chosen for each statement. It is less appropriate to ask the test taker to mark the statements that are true and leave the false statements unmarked. In this case, the bundle of dichotomous items (MTF) is similar to items with a single best answer (SBA), which, however, have different characteristics and do not require an answer for each of the offered options separately.

There are, however, other possibilities, in which mutually exclusive alternatives are indicated.

Example:

We have five test tubes available. Each of them contains 1 ml of a solution of one of the carbohydrates listed below with a concentration of 1 g/l.

We add 1 ml of potassium hydroxide solution (2 g/l) to each of the test tubes and boil the mixture briefly. Then we add a solution with complex bound divalent copper to all test tubes. The resulting color of the mixture in some tubes is blue, in others red.

For each carbohydrate, circle what color you expect the mixture to be after the described experiment is over:

a) amylose BLUE – RED

b) fructose BLUE – RED

c) glucose BLUE – RED

From individual dichotomous items, their set often differs in terms of scoring. Different scoring schemes are used:

All or nothing - 100% (most often 1 point) is given if all answers in the set are correct, 0 points in all other cases

Partial score – each partial dichotomous question is scored independently, e.g. 0.25 points

Partial weighted score – each partial dichotomous question is scored independently, each has a different point value (e.g. according to importance or difficulty)

Penalty scoring – negative points are given for some wrong answers

Partial scoring – e.g. PS50: If the test taker answers the entire set correctly, they receive 100%. If they answer more than half of the component dichotomous questions correctly, they will receive 50%. In other cases, they get nothing.

Guessing correction – an estimate is made of what score the examinee could have achieved for the set by random guessing and the result is corrected

Other more complex methods. Most frequently used are the All or Nothing and PS50 methods, while others are being discarded.

Problems with only one correct answer (Single Best Answer, or SBA)

The most often used and at the same time the most effective type of multiple choice problems are problems with a single correct answer. In appearance, they may resemble sets of dichotomous items, but they are constructed differently. Again, the item has a stem followed by an offer of several options. The task is to choose an answer that is significantly better than all the others. Thus, the examinee does not evaluate each option separately and does not try to determine whether it is valid or not, as in a set of dichotomous problems, but compares the offered options with each other. At the same time, none of the options offered may be completely correct in all circumstances, and none may be completely wrong. On the other hand, it must be possible to rank the options offered from best to worst.

Comparison of TRUE/FALSE and SBA type multiple choice item

TRUE/FALSE Item

An item with a single best answer

The shape of the Earth is close to

The shape of the Earth is close to that of a rotating body.

Which of the following is it most similar to?

Sphere TRUE – FALSE

Ellipsoid TRUE – FALSE

Ovoid TRUE – FALSE

Cylinder TRUE – FALSE

Sphere

An ellipsoid

Ovoid

Cylinder

Both problems in the example ask the same question and offer the same solutions. In both cases, the author considers option 2 to be the correct answer. Note, however, that the TRUE/FALSE item is not completely unambiguous: it can be argued that the Earth does not have an exact ellipsoid shape, and on the other hand, its shape can be approximated accurately enough by a sphere for some purposes.

In the case of an SBA-type item, the situation is different: the examinee has to choose the most accurate (not necessarily completely accurate) answer. The solution is clear.

At the general level, it can be said that for higher education, SBA type items are more suitable than sets of dichotomous items. The fact that it is impossible to say with absolute validity whether an individual option in SBA is completely correct or, on the contrary, completely wrong, reflects real life. Testing with SBA better prepares students for real world experience. On the other hand, it tends to be difficult to create a larger number of MTFs for a certain topic in a way that ensures the items are truly unambiguous. The pursuit of clarity often leads to the refinement of the assignment, which is then increasingly longer and more detailed, but often also more instructive, resulting in an MTF item that is clear but at the same time very easy. It can therefore be said that, for a certain topic, more SBA items of high quality can be created than MTF items. The common concern that a problem with a single correct answer will be easier and more predictable than a set of dichotomous problems that may have more than one correct answer is not justified. In practice, however, it turns out that properly constructed SBA-type items tend to be more difficult and usually differentiate better than MTF items.

The reader can find more detailed information on the creation of SBA type items in the chapter entitled Recommendations for the Creation of Test Items.

Note:

We often come across the term Multiple-choice question, MCQ. This is a more general term that includes multiple true-false (MTF), single best answer (SBA), and other types of items. In this publication, we deliberately avoid the term MCQ, as its meaning is not clear-cut. This is because in common communication, the term MCQ is often narrowed down to cover only the most common type of items, and depending on the customs in a specific geographical area, it means something different each time:

In the literature, MCQs are most often synonymous with single-answer items, i.e. SBA.

In some parts of the world, MCQs are most often used as a designation for bundles of dichotomous items, i.e. MTF.

Due to the fundamental differences in the construction and properties of SBA and MTF, the term MCQ can cause unpleasant misunderstandings.

Matching Questions

The matching question consists of a set of premises and answers. The examinee's item is to assign the best answer to each premise. Matching problems can have different ratios between the number of premises and answers, and various subtypes are sometimes distinguished accordingly. In the simplest case, the number of premises and answers is the same, and it is given that each answer belongs to exactly one premise. Another possibility is that there are more premises (and some answers are used more than once).

Example:

Put each animal into a group according to the type of food it eats

1. Domestic pig _____

A. Carnivores

2. Desert lion _____

B. Omnivores

3. Steppe zebra _____

C. Herbivores

4. Przewalski’s horse _____

5. Nile crocodile _____

Conversely, there may be more answers offered than premises. An extreme case are so-called extended matching questions, or EMQ. In many ways, they resemble several SBA items in a row, but the range of options is significantly larger (typically more than ten) and the same set of answers is used for multiple premises. EMQ-type items have become widespread in the medical field, where they have mainly been used for testing clinical disciplines.

Example:

Choose the most likely diagnosis for each case report of back pain from the following menu:

A. Ankylosing spondylitis

B. Dissection of the aorta

C. Intervertebral disc herniation

D. Lumbar spondylosis

E. Vertebral fracture

F. Intervertebral disc infection

G. Pars interarticularis defect

H. Metastasis to the vertebral body

I. Renal colic

J. Herpes zoster

Item 1:

A 23-year-old man has a six-month history of lower back pain. The pain mainly affects the thoracolumbar junction and the right buttock. The pain is usually worst in the morning, and makes it difficult for him to get out of bed. There is a partial improvement during the day. During the examination, we find limited mobility of the lumbar spine, especially lateral flexion.

Item 2:

A 32-year-old woman comes in due to sudden pain in the lower back. The pain is constant, it does not depend on her position. All spinal movements are limited and painful. Three weeks ago, she had a urinary tract infection which was treated with amoxicillin.

Matching items can take a variety of graphical forms. For example, the test taker can write the letter or number of the chosen answer for each premise, or they can connect the premises and answers with a line. When testing on a computer, answers are often selected from a drop-down list, or answers are dragged to the premises using the mouse. In a broader view, matching items also include, for example, placing labels into an image.

Matching items are widely used, for example, in language learning. Their features are very similar to single best answer questions, essentially a sort of bundle of SBA questions. In many fields, matching questions are gradually being abandoned, they are being replaced by questions of the SBA type. A smaller number of types of items that are used in a certain test is usually an advantage, because the test taker does not have to think so much about what form of answer is expected from him, and can better concentrate on answering the questions themselves. This also makes the test “friendlier”, reducing test anxiety.

Matching tests are scored using procedures similar to those for scoring dichotomous item bundles (MTF), most commonly all-or-none, subscore, or PS50 methods.

Ordering Items

The examinee has the task of ordering the presented items (e.g. concepts, events) according to a certain rule. It can be, for example, the ordering of the steps of a certain procedure or the arrangement of some objects according to some quantity or property.

Example:

Rank the liquids from highest to lowest freezing point.

Water

Oil

Alcohol

Glycerine

From a formal point of view, ordering items can resemble matching questions, since the examinee assigns its order to each item. In some cases, arrangement items can have more than one correct solution, e.g. the seasons follow one another in the order spring – summer – autumn - winter, but also autumn – winter – spring – summer, etc.

The weakness of the ordering items is that they are difficult to evaluate. An all-or-nothing method is sometimes used, but this assessment tends to have low sensitivity. Therefore, evaluation is most often performed sequentially in pairs and it is examined whether the items in the pair are arranged correctly or not:

Example:

On the table lie four cubes of the same size, each cast from one metal – iron, aluminum, copper and gold. Sort the cubes from lightest to heaviest.

The correct order: aluminum – iron – copper – gold

Examinee's answer: aluminum – copper – iron – gold

aluminum – copper: correct order

copper – iron: wrong order

iron – gold: the correct order

The examinee receives 2/3 points for the item.

3.2 Open-ended Items

A short form answer item

These are also referred to as items with a short answer, short-answer question, SAQ. The examinee is most often required to answer the item with a short form answer consisting of one word or a phrase. Often the answer is also the result of a calculation, a sketched graph or picture, a chemical formula, a mathematical equation, etc. Depending on the construction, these items are sometimes divided into production and supplementary items:

Production item: What is the name of the capital of Great Britain?

Supplementary item: The capital of Great Britain is _________________.

Short answer questions are an excellent part of formative tests. Their benefit lies in the fact that they provide the teacher with information not only about which parts of the curriculum the students have mastered and to what extent, but at the same time they can learn about possible mistakes, misunderstandings and erroneous concepts from which the students start out. Thanks to this, it is possible to react in a more targeted way when preparing the next lesson and better guide students.

Short-answer questions are also useful even for summative tests. In some areas, they are quite common – for example, in the learning of languages, mathematics and geometry, etc. However, preparing short answer questions for a summative test must be done very carefully in order to avoid ambiguities.

Example:

A properly constructed SAQ:

What is the name of the capital city of Great Britain?

Correct answer: London.

Even in this simple case, the evaluators must agree on how they will grade an examinee’s answer of, for example, “Londra”.

A poorly constructed SAQ:

The capital of Great Britain is _________________ .

In this case, it is not clear what the author of the item is asking: is the capital of Great Britain London, big, in England, historical, on the Thames...

In general, it can be said that there is more than one correct answer to most SAQ items and the answers must usually be evaluated by a qualified evaluator – an expert in the relevant field. Assessors must either agree on which answers should receive full points, which ones should get partial scores, and which answers to consider incorrect, or they must be given detailed instructions.

What is a unit of mass?

Answers can include, for example, kilogram, gram, ton, metric cent, pound, ounce, carat, quart...

What animal is in the coat of arms of Moravia?

Red and silver checkered eagle with golden crown

How to evaluate the omission of a color, or the description of a color combination as red and white? How to evaluate the answer “single-headed eagle”? Is it a mistake to leave out the information about the crown?

Due to the number of possible answers, it may even happen that it is not possible to reliably assess whether the answer is correct.

Give the name of at least one painter.

How many points will you award for a name you don't know at all? Can it be reliably verified that the answer is correct? How much time and effort will such verification cost?

Outside of summative testing, this feature of SAQ can be used. Different answers to the same question can stimulate discussion during the lesson and activate the students. They can also help us create closed-ended SBA items: if you need to find distractors for this kind of a item, ask the same using a SAQ. The answers obtained can provide valuable inspiration.

Fill in the blanks test (cloze deletion test)

In a fill-in-the-blanks test (cloze test), the examinee receives a text with portions left out, which the examinee must complete. Most often, he or she has to fill in the missing words in several different places. In some cases, the examinee is supposed to complete some part of a word.

Example:

Complete the missing text

A plane body that has three sides and three vertices is called a t_____angle. A special category is made up of right _________ le t________les. Their longest side, the so-called hy__________use, compared with the right __________, which is created by the_________.

Supplementary items are widely used in language learning, for example to test vocabulary or to test the ability to understand the spoken word.

Example:

Based on listening, fill in the missing words and data

Earth's only natural satellite is called ________. Its average distance from the center of the earth is _________ km. It orbits the Earth about once every ________. Man stepped on it for the first time in _________.

In this form, supplementary items are actually a bundle of items with a short-form answer. Sometimes similarly constructed items, in which the examinee chooses their answers from a predetermined list of words, or the number of possible answers is limited (e.g. items of the type Complete -its-/-it’s-) are sometimes referred to as fill-in-the-blanks (cloze). In that case, this actually involves only a differently indicated matching assignment, i.e. closed-ended items.

When scoring completion items, a partial score is most often assigned for each correct completion, or a partial PS50 score is used.

Modified essay

A Modified-Essay Question (MEQ), is different type of short-answer problem set. The introductory text is followed by the first question, then supplementary information and other questions alternate.

Example:

A 78-year-old man, a widower who lives alone, came to the in-patient clinic with a complaint of fatigue and weight loss. He was admitted to the general internal medicine department, where you work, for further evaluation.

Question 1: What are the three most likely diagnoses?

Question 2: Write five questions you would ask the patient that would best help you distinguish between the three diagnoses.

Laboratory tests showed mild anemia with a hemoglobin concentration of 104 g/l. The red blood cell count is lower than the reference range. You therefore conclude that the patient suffers from microcytic anemia.

Question 3: Give two typical clinical signs that you will look for when examining the patient.

Question 4: Briefly write how the given result changes your initial diagnosis.

The modified essay is somewhere between short-answer and wide-open questions. They are especially valuable in formative tests, in which they can to some extent simulate a problem-solving dialogue between the student and the teacher. However, preparing this format for summative tests is challenging. In addition to all the requirements and limitations mentioned for short creative items, there is the fact that a mistake at the beginning of drafting a modified essay can also affect the following sub-questions. This would not happen in an orally conducted dialogue, as the teacher would correct the initial mistake in an appropriate way. Of course, it also affects whether the student can go back while completing the test or not.

Note that in a modified essay, individual questions tend to be approached in a much broader perspective than in typical short-answer questions. This time we no longer expect the test taker to answer with a word or phrase. Rather, their answer will consist of several separate statements. The student must not only master the material he or she is being tested on, but must also be able to formulate their answer concisely and precisely in a short time.

Essay

An essay is a item with a long, formulated answer. The examinee writes a text of extended length, ranging from one paragraph to several pages.

The essay usually forms a separate part of the exam, and its evaluation is not combined with other items. The grading may be subject to the subjectivity of the evaluator. Therefore, if an essay is used for summative assessment, it is usually assessed by multiple assessors and, ideally, the essays are anonymized. In order to make the assessment more objective, it is usually assessed according to a predetermined outline, i.e. the extent to which the examinee fulfilled certain assessed aspects (“rubrics”) in the essay is scored. Apart from the knowledge, i.e. the essay content side, the ability to give a well-organized interpretation, analyze and describe contexts, express oneself in a clear, structured and comprehensible manner and use professional terminology correctly, adhere to conventions customary in the field, etc., typically have a great influence on the grade as well.

Using the essay as an assessment tool varies throughout different parts of the world. To some extent, it can be said that the alternative to the essay is the oral exam, during which the examiner conducts a conversation with the examinee. An advantage of the oral exam can be the dialogue, which makes it possible to more accurately identify the examinee’s strengths and weaknesses. However, the disadvantage is the non-reproducibility of the assessment. While the essay can be re-graded at any time by another assessor, the oral exam cannot be repeated. Even if it is audio-visually recorded, repeated assessment can be difficult, especially if the examiner conducted the interview inappropriately or incorrectly at some point. Examining in front of a committee can contribute to increased objectivity, but in practice a larger number of examiners does not automatically mean that they can be considered mutually independent evaluators who do not influence each other during grading. Despite all these reservations, the oral exam has its pedagogical value, thanks to the possibility of basing the act of examination on professional dialogue and interaction. However, it is conditional upon the high erudition and professionalism of the examiner.

3.3 Other Types of Test Items

Script Concordance Test (SCT)

The item begins with a text similar to the one in an SBA. This is followed by a question that both offers a possible solution (hypothesis), and also brings new information and asks to what extent the new information supports the proposed hypothesis. The test taker usually chooses an answer from five options (from the hypothesis being very unlikely to the hypothesis being very likely).

Example:

You are examining a 14-month-old Holstein heifer who is bloated and anorexic. Her body temperature is 39.5 °C, heart rate 115 beats/min., respiratory rate 64/min. She is not dehydrated, rumen contractions are inaudible, there is very little excrement.

If you are considering dislocation of the spleen to the left as a possible cause, and in the laboratory finding fibrinogen is 10 g/l, this cause becomes

-2 very unlikely

-1 less likely

0 neither less nor more probable

1 more likely

2 very likely

A vignette is often followed by several questions:

You are examining a 48-year-old man with Fournier's gangrene who has repeatedly undergone surgical removal of necrotic tissue. The patient is treated with broad-spectrum antibiotics.

If you were planning to...

...and found out that...

the planned solution is

1. skin transplantation for a scrotal defect

in part of the defect there is granulation tissue, but the disease continues to progress to the groin, where there are new necrotic areas,

-2: absolutely contraindicated

-1: relatively contraindicated

0: equally indicated or contraindicated

1: indicated

2: highly indicated

If you were planning to...

...and found out that...

the planned solution is

2. further debridement

the patient is septic, intubated, cardiopulmonarily unstable,

-2: absolutely contraindicated

-1: relatively contraindicated

0: equally indicated or contraindicated

1: indicated

2: highly indicated

If you were planning to...

...and found out that...

the planned solution is

3. hyperbaric oxygen therapy

the patient is septic, intubated, cardiopulmonary unstable,

-2: absolutely contraindicated

-1: relatively contraindicated

0: equally indicated or contraindicated

1: indicated

2: highly indicated

The scoring of the answers is determined separately for each question, based on the opinion of a group of experts. Each expert marks one option that they consider to be correct. The option marked by the largest number of experts (the so-called modal option) is scored with the full number of points. Scores for other options are assigned by relationship

[Score for possibility] = [number of experts that have marked this option] / [number of experts that have marked the modal option]

For example:

The item was evaluated by 10 experts. They marked the correct options as follows:

Option

-2

-1

0

1

2

The number of experts that have marked the option as correct

0

0

2

5

3

The modal option is the answer "1", which was marked by the largest number of experts – so this option will be awarded full points. The full scoring will look like this:

Option

-2

-1

0

1

2

Score for option

0/5 = 0

0/5 = 0

2/5 = 0.4

5/5 = 1

3/5 = 0.6

3.4 Recommendations for creating test items

Item writing is a item that requires imagination and creativity, but it also requires considerable work discipline and knowledge of learning objectives. The creation of items must be based on a clear idea of the learning objectives. A test should measure a single cognitive area.

Before you start creating items

Before it comes to the actual creation of items, it is necessary to return to the question of what we actually want to test and why. It is not enough just to take the textbook, flip through the chapters that cover the content of the upcoming test, and create items for the text that catches your eye.

Ideally, we have a prepared test plan (blueprint) at our disposal. If not, the test plan should be created before we start writing items. If even that is not possible for any reason, we should at least have the learning objectives written down in as much detail as possible (i.e., not just the thematic areas that the learning covers). It is advisable to add to the learning objectives an idea of how big a part of the test should deal with them.

A test plan or a detailed list of learning objectives gives us a clear indication of what items we need for the test and in what quantity. At the same time, already at this stage we should have a basic idea of the types of items that we will use in the test.

Selection of Item Types

When compiling a summative test, we will not go wrong if most of the items are single best answer items, and, if necessary, we supplement them with open-ended questions with a short form answer. We should use other types of items judiciously, or according to the customs of a specific field. Single best answer items provide the best value for money. The time required to answer them is relatively short, so that a sufficient number of such items can be included in the test, and at the same time they allow sufficiently sensitive recognition of students' skills. Items with a short form answer are more difficult to evaluate, but on the other hand, they are a suitable addition, as they provide better feedback to the teacher.

In formative tests, especially if the group of test takers is small, the ratio can be reversed – the majority of the test can be made up of short-answer questions supplemented by multiple-choice questions with a single best answer. In formative tests, there is no need to be afraid of other item formats, if their use is expedient. But the test should never combine too many different formats (no more than three or four), otherwise it will be confusing and students will spend a lot of time researching what is actually being asked of them and how they are supposed to answer which item.

If the test combines multiple item types, students should be clearly told what is expected of them. The instruction must be very specific.

Example:

Instruction for single best answer problems:

Appropriate: Circle the best answer.

Unclear: Choose the best answer.

Instruction for short formed answer problems:

Appropriate: Answer with a single word or phrase.

3.4.1 Recommendations for Creating Multiple Choice Items

Recommendations for the creation of closed-ended (multiple choice) items will be given here for items with a single best answer (SBA), i.e. the type of items that should be the basis of most tests. For the most part, the same recommendations can also be applied to other multiple choice formats.

Ambiguity is a frequently-occurring problem in multiple choice items. It mostly results from the author of the item having a specific situation in mind when compiling it, but then trying to write the item as succinctly as possible. This results in the loss of details and assumptions from the text, which are important for answering it. The student taking the test must first guess what the author of the test actually had in mind, and only then can they answer. The item usually measures the student's ability to guess what the teacher wanted to ask, rather than the actual knowledge and skills in the tested area.

Therefore, the basis of a good-quality multiple choice item is a well-written stem. Problems with a single correct answer tend to have a relatively long stem, several lines long (sometimes referred to as the statement). The stem should tell a story – describe a simple but real situation, or perhaps an experiment. A clear description of the situation the author had in mind will avoid most ambiguities. Moreover, students will find items compiled in this way to be motivating – stories that remind them of real experience, remind them that they are learning something practical that they will need in their future employment.

For novice test writers, writing statements is sometimes difficult. The basic recommendation is that they should write items as if they were designing research to answer specific (but of course simple) questions in their field.

The statement is followed by the actual question. That question should be short, unambiguous, and it must ask only one thing. In single-best-answer items, it must be clear from the question that the student is really being asked to mark only one option that is a better answer than all the others. This is especially important in the event that other options offered would make sense, at least under certain conditions. Properly constructed questions tend to take the form of, for example, “What is the most likely cause?” or “What is the most appropriate next course of action?” “What will the events described most likely lead to?”

Another simple principle follows from the above: Good quality items usually have a relatively long stem (text) followed by a short but clear question. This is followed by a menu of options, which should also be short.

Regarding the options offered, it is usually easy to create the correct answer (the so-called key). The rule here is that the expert should give the correct answer after reading the stem text and the question, even without an offer of options. Choosing from several options actually just makes it easier to evaluate the test and provides some help to the student who is not yet as proficient as an expert in the given area.

It is more difficult for teachers to suggest incorrect answers – distractors. The authors of the items may no longer fully remember what the problem was when they were learning the tested topic, and then propose distractors that are irrelevant for the students. It may therefore be advantageous to first give the item to students as an open-ended question, preferably in a formative test. Distractors are then created from incorrect responses. At the same time, the teacher gets an idea of how difficult the created item will be. If the teacher combines this kind of procedure with a discussion of the answers, or if the teacher invites the students to “think aloud” during the solution, it will provide the teacher with additional valuable information both for teaching and for further modifications to the test items.

More experienced test writers are often helped by a quintet recommendations of how an item with a single best answer should look. These five recommendations will also help reviewers during opposition proceedings for new test items.

Focus on an important problem.

Higher education should prepare students for work in the real world, and therefore test items should relate to problems encountered in practice. Don't waste time with trivial or overly complex questions. There is no point in testing little essential, marginal knowledge – such items usually do not give any indication of readiness for practice, rather they test how well a student “knows how to take tests”. Do not use “trick questions”, and avoid negatively worded items. Test knowledge and understanding, not concentration.

Test for use of knowledge, not recall of a concept or an isolated fact.

A longer text of the item requires the student to somehow evaluate the described situation and interpret it. In summative tests, avoid testing definitions and classifications – test for the student’s understanding of the content of the concepts and ability handle them, or whether the inclusion of a concept in a certain category is also connected with an understanding of the properties of that category. Whether this recommendation has been fulfilled can often be easily verified: copy the entire item stem into an Internet search engine – the correct answer should not appear in front of you.

It must be possible to answer the item even with the answer options hidden

Have your peers review the question. First, give it to them without the options provided – they should be able to answer it correctly. If not, rewrite the item.

Do not use relativistic nor absolute terms.

The text of the stem or the anwers choices offered should not contain relativizing terms such as often, rarely, exceptionally, mostly, etc. The use of such words will make the item unclear, whereas the relativizing term conceals the author's specific point of view, which the student may not guess correctly. For example, a certain situation may be rare from the point of view of the general population, but at the same time it is relatively common from the point of view of an expert who professionally deals with the solution of such situations.

“Forbidden words”, especially in the multiple-choice answers, also include terms expressing the absolute, such as always, never, etc. Only rarely does something really apply 100%, so the use of such a word in one of the offered choices usually means that it is a distractor – and students can easily recognize these.

All answers offered must be homogeneous.

All the options offered must appear similar – they should be written in a similar style and should be of a similar length. They must fall into a single category – for example, all the options offered are possible causes of an event, parts of speech, biological taxonomy, work activities, chemical substances, etc. It must be possible to rank the answers from best to worst.

OBRÁZEK

Fig. 3.4.1 In the example on the left, the options offered for the question can be ranked from worst (option C) to best (option D). This confirms the homogeneity of the options offered. If you pose the same question to several experts, they will all rank the offered answers in the same order. On the other hand, in the case on the right, it is not possible to sort the offered options – each option falls into a different category, answers the question from a different aspect, so we would be “comparing apples with oranges” when trying to sort them.

An example of a question with non-homogeneous options

Select the best statement about Marfan syndrome:

A. It affects men more often

B. It is a disorder of the collagen ligament

C. It is treated with hyaluronic acid

D. It is often associated with oligophrenia

E. It is manifested by conspicuously short limbs

The options offered should be ordered randomly or alphabetically. If the options contain a numeric value, they should be sorted by that number.

3.4.2 Recommendations for Creating Open-ended Items

Follow these guidelines when creating short-answer questions for written tests: [30] [30]

Formulate questions simply and clearly, avoid linguistic tricks and catches. A good short answer question tests knowledge of specific facts or the ability to analyze and interpret a scenario. It is not appropriate to simultaneously, in the same item, test the student's ability to understand a complexly constructed question – the assessment results would then be practically uninterpretable.

Try to answer the question from different points of view. A question that asks about one specific fact should have only one correct answer. Conversely, a question that asks about possible variants (e.g. differential diagnoses) will have several correct solutions. Keep in mind that even a question that seems obvious to you can be understood differently by different readers. It is always advisable to have the questions checked by a reviewer.

State the length you expect the answer to be, e.g., Answer in one word or phrase., or Sketch the graph of a function.... Also write how the item will be graded.

Use negatively worded questions with caution.

Positively worded questions ("What is the best course of action...", "What is the most likely cause...") have more didactic value than negative questions ("What is the wrong course of action"). If you are already using a negatively worded question, emphasize the negative, for example by using capitals ("Which antibiotic is NOT appropriate in this situation?").

Make sure that the answer is not dictated by, for example, the size of the space for entering it.

Prepare evaluator instructions carefully and have them checked by your colleagues along with the item.

3.5 Answer Guessing Techniques

Test wisdom (test-wise, testwisness) reflects an ability that allows the student to correctly select the answer to the question even without being knowledgeable in the area that the item covers [31] [32]. It takes advantage of the fact that the authors of multiple choice items usually first have a well-thought-out correct answer and supplement it with distractors. They often do so stereotypically, so their considerations can be guessed or they commit typical oversights [33].

The longest answer is correct

Test wisdom tells students to prefer the longest answer. Sometimes the correct option is also at first glance more complete than the other options, or is more specific or detailed [34]. This is because, in an effort to be unambiguous, the authors of the items write the correct option in great detail, and then pay less attention to the distractors. It may also happen that they use an incomplete part of the text of the correct answer to create distractors. A typical example could be earlier items, e.g. in driving school, or in work safety tests, where the longest of the answers offered was usually the correct one.

The answer “in the middle”

If the answers can be arranged logically (e.g. if they are numbers) and the student does not know which one is correct, he or she guesses one of the middle ones. When item authors devise distractors, they often choose some smaller and some larger values. Students who are aware of this rule therefore guess one of the answers within the options offered. By eliminating the lowest and highest values, the probability of a correct guess increases substantially. The answers offered also sometimes “circle around” the correct solution, so it is enough to trace what the mutual similarity is and the correct answer can be guessed.

Example:

What is the area of an A4 format sheet of paper compared to A6 format?

a) Half

b) Triple

c) Fourfold

d) Eightfold

A student who does not know the answer excludes the extremes, i.e. options a) and d). So, the student then chooses between b) and c). Answer c) is similar to options a) and d) (they are powers of 2), while option b) is different. So, the student tries answer c) - which is the correct answer.

Grammar instructions

When creating the item stem and distractors, care must be taken that the grammatical form does not predict the correct answer. When creating the stem of the item, the author usually already has the correct answer in mind, and the formulation of the stem will correspond to it. This may not apply to distractors, which are often invented after the fact or changed at the last minute.

Example:

A resource that provides the fastest way to check if there is anything new in the field is

a) Books

b) Professional journals

c) The Internet

d) Scientific conference

Absolute and relativizing expressions

When designing the answers, the author attempts to exclude misunderstandings by using specification, but providing a guide for guessing. If some option contains any more refined (“extreme”) term, eg, always, never, only, necessarily, must, all, none, impossible, still, it is an incorrect response (distractor). Conversely, relativizing expressions such as often, rarely, perhaps, sometimes, usually, mostly, can be in the correct answers.

Example:

Which statement about mammals is correct?

a) They are exclusively terrestrial animals.

b) No mammal can fly.

c) They can have fins.

d) They have all developed eyesight.

Logic key - opposites or an exhaustive list of possibilities

If there are two or three answers that cover all the possible options, one of them will surely be correct.

LOGICKÝ KLÍČ - PROTIKLADY PRO VYČERPÁVAJÍCÍ VÝČET MOŽNOSTÍ

Pokud jsou mezi nabízenými odpověďmi dvě nebo tři, které pokryjí všechny možnosti připadající v úvahu, jistě bude jedna z nich správná.

Example:

Consider a mathematical pendulum with weight m and thread length l. If we increase the mass m,

a) The period of swing is shortened

b) The period of swing does not change

c) Swing time is extended

d) Maximum deflection will decrease

e) Maximum angular velocity will decrease

Options a), b) and c) together cover all possible cases (the swing time is reduced, unchanged or increased). A student who uses test wisdom will think only about these options and will not waste time at all on options d) and e).

Too simple an answer

Students anticipate that the items will contain catchers and complexities. They therefore tend not to choose an answer that is simple and self-evident. Sometimes it is appropriate to include a correct answer of this kind to break the pattern.

All of the above

If the offered answers include “all of the above” or “none of the above” (and similar), then students will prefer this answer. Moreover, it turns out that formulations of this type in the answers do not differentiate well between better and worse students. They should therefore not be used in tests.

Repetitive is correct

The teacher often prepares the distractors so that they seem as similar as possible to the correct answers. It may happen that the answer can be guessed by comparing the options offered and finding where they agree.

Example:

Express the Ohm unit in base SI units.

a) m ·kg·s-3·A-2

b) m2·kg·s-2·A-2

c) m3·kg·s-3·A-2

d) m2·kg·s-3·A-2

Expressing the ohm unit in basic units is undeniably a difficult item. However, we see that the answers differ practically only in exponents. And the exponents are repeated in the answers. So, we choose the option in which all repeated characters appear. In our case, it is the option d).

Hint among items

Tests may contain items that are clues to the answer of another item. In longer tests in particular, it is very difficult to keep this aspect under control. Two strategies are used to prevent this situation from arising. The items in the item bank have set relationships with each other, and the system does not allow “related” items to be selected for the same test. The second strategy is to offer students questions gradually, and not allow them to go back to already answered questions.

Verbal Similarity

When there are verbal similarties between the stem of the item and one of the offered answers, this answer tends to be correct.

3.6 Automation of Test Item Creation

As the use of computer-assisted testing increases, and especially with the development of adaptive testing, methods that could simplify the creation of test items are attracting attention. In the traditional approach to test construction, each item is created by area specialists. First, the item is written by the author, then other experts oppose it, then the teacher checks it in a pilot test and revises and modifies it according to the results. Only then is the item finally used for testing. The whole process is long and expensive. As a result, it is increasingly difficult to meet the increasing demand for test items [35]. Automatic item generation (AIG) could greatly save time and resources and is therefore being intensively researched. Some concepts for solving this item have already reached the stage of practical testing.

In the first of the concepts, we can divide the process of automatic item cloning into two steps. Test item writers first create item models that serve as a kind of template. They try to distill the essence of the item, which is fundamental for demonstrating knowledge. Various alternating terms are then suggested for appropriate places in these templates (often by machine, e.g. using synonym dictionaries). Using a set of wildcard terms, the algorithm then turns this template into a group of related items by creating all possible permutations. This generates “new” but not independent items. No more than one item from each clone group can be used in a particular test run. In addition, some permutations will lead to nonsensical or improbable combinations, so they must be excluded. [36], [37], [38]

It is then a matter of discussion whether the item banks should only contain the resulting clones or the source templates and variable components of the items. The purpose is to obtain items whose psychometric characteristics would be appraisable from the known results of another item from the same series of clones. Thanks to the demanding nature of the creation process, which leads to the need to clarify the essence of each item, the items created in this way are often of surprisingly high quality. Note that the usability of the machine-generated versions also depends on the specific language. For example, in Czech, with its complicated grammar, it would be extremely difficult.

A similar procedure was also tested in efforts to create item-comparable tests for reliability verification using the test-retest method. Attempting to modify originally functional items by changing the alternating terms was shown to result in the creation of items of greater difficulty. [39]. This somewhat undermines the original notion that the cloned items will have the same psychometric parameters as the original. It is therefore a question whether the whole process makes sense when new items are created, but they still need to be calibrated anyway.

The first papers dealing with using artificial intelligence for the automation of test item creation are beginning to open the second concept. Its model procedure was demonstrated at a workshop at the meeting of the European Council of Medical Assessors in Braza, Portugal. A group of test writers was given the task of creating a cognitive map for the given topic (abdominal pain). A cognitive map helps to describe the problem successively by elements (e.g., age, gender, context, vital signs, cause, diagnosis). Each of these elements can have a set of different values. Experienced test developers need several hours to create a cognitive map. The computer then generated a set of items that represent different combinations of elements of the cognitive map. During the workshop, this mixing of elements was done with the help of the Excel application. Deploying a similar system could make life easier for test writers in the future. [40] The problem with this approach lies in the time-consuming and expensive nature of creating a cognitive map. A paper on automated item generation in first grade mathematics shows that automated generation is cost-effective (compared to traditional generation) if a set of more than 200 items can be generated from a single cognitive model. [41]

In the same year, another system was presented, which can use artificial intelligence to mine data from a bibliographic branch database and use it to create item stems and distractor suggestions. These draft items can serve as a semi-finished test for human writers to create new items more easily. [42]

3.7 Test Item Reviews

In the test preparation process, especially for exams of great importance, the checking of items with the help of expert reviews before their use in the test (so-called panel review) plays an irreplaceable role. While in the case of a fourth-grade quiz, which the teacher uses to determine the students' knowledge of science, it may not be necessary for other teachers to assess the content of the test, for tests that are part of an entrance exam or a professional certification exam, it already is necessary. Items go through several levels of independent review before being seen by the first test taker.

The opposition, or item review, is divided into several phases, which always focus on a specific area. Its objective is to reveal the shortcomings that items and tests usually contain in their initial form. The motivation is to ensure correctness, optimize the test and eliminate subjective influences. Even if the review is initially somewhat time and organizationally demanding, its benefit is undeniable and increases with the importance of the test. After successfully managing all the revisions listed below (content revision, fairness revision, editorial revision), the author team should go through the final form of the individual items again and approve all the changes made.

Why is it necessary to check the items and the whole test?

Test items are part of a tool that we use to measure a certain skill of the test takers. Checking the correctness, wording accuracy and non-contradiction of the items makes the test a better measurement tool and reduces the likelihood that the test will be unfair and that any of the participants will complain about it or its individual items.

Who should check the items?

This can vary widely depending on the importance of the test. For tests of minor importance, one additional reviewer is more than sufficient. You simply ask a colleague to take the test for you and check it. For exams of high importance, such as entrance exams, graduation tests, etc., the item must be reviewed by several reviewers with clearly assigned roles. The reviewing experts must be both experts in the given area and at the same time they should be familiar with the population being tested.

What do the controllers check?

It depends on the type of reviewer and their role in the review process. Testing institutions often create checklists for reviewers to follow. The reviewer can check that the item stem is well worded. Whether or not it is grammatically instructive and does not facilitate choosing the right answer. Whether the key is correct and the distractors incorrect and whether all options are of comparable length. The controller can check the correctness of punctuation, the correct use of superscripts and subscripts, compliance with writing conventions for variables and units.

How is the review work organized?

Although the review form (checklist) can be in paper form, it is more common for it to be in electronic form. It is often directly integrated into the item bank, so items do not leave the bank's secure environment even during review. The test administrator can check the status of reviews in the item bank and motivate reviewers to perform better.

The process of opposition hinges on cooperation, similar to the preparation of a complete test program. Several participating experts independently assess the suitability of individual items and work together to eliminate all shortcomings that could hinder practical implementation. Teamwork plays a crucial role in opposing tests and test items.

The process of opposing items and the test itself can be divided into three phases, through which the opponent is guided by the item reviewer form (discussed in more detail below).

3.7.1 Content Review

Are the answers correctly and accurately worded? Aren't the distractors debatable?

As part of the content review, it is highly recommended that both the co-authors of the entire test and independent experts who were not involved in their creation check the questions and the answer options. A writer’s subjective attitude may have resulted in an ambiguous, i.e. incorrectly worded test item, the use of which would reduce the value of the test.

For most educators, creating alternative answers (distractors) tends to be a particularly difficult activity. In general, distractors should not be meaningless statements or absurd possibilities that the examinee will automatically exclude, but on the contrary, they should force him to think and then eliminate them after logical reasoning. MTF-type multiple-choice items are particularly susceptible to ambiguous distractor wording.

Other types of content deficiencies may arise for other types of questions. Single Best Answer (SBA) questions must be reviewed so that there is expert consensus on the clear best answer.

If the teacher is faced with the item of creating more test items and, in addition to correct answers, also designing a number of suitable distractors, he can help himself by assigning his new items to students as short-answer questions, as part of formative testing. Students will often design highly functional and attractive distractors when generating responses.

In general, when it comes to content it is recommended especially:

to check the accuracy of the wording of the assignment/item stem,

whether the options offered in each item are formulated in such a way that under no circumstances, in any interpretation or in any considered case, the distractor cannot be the correct answer and vice versa (applies especially to MTF),

whether the items in the test correspond to the test plan (blueprint).

3.7.2 Editorial Review

Are the questions sufficiently comprehensible, typographically uniform and without typographical errors?

Editorial review may at first glance appear to be not very time-consuming, but in practice it can be more complicated. It is necessary to go through all the test items and verify whether they are sufficiently readable, comprehensible and formally and typographically uniform. It is advisable to rework complex sentences, double negatives and difficult questions into a simpler form so that the student cannot get lost in the wording. The assignment of the item and the options offered should be constructed as clearly as possible. The uniformity and style of test item creation varies among test writers. In this phase of opposition, both terminological and typographic aspects are homogenized. Grammatical correctness is an integral part of checking any texts. This also applies to creating test items. Eliminating all grammatically incorrect or questionable expressions according to spelling rules should be the final stage of editorial review.

In practice, it turns out that a single review is absolutely insufficient. The ideal of 5-7 reviews is hard to achieve with limited funds, but 3 reviews seems like a workable minimum. Often, only one of the reviewers draws attention to a problem. Therefore, the review processor must be very attentive to the reviewers' suggestions in order not to overlook a possible problem.

Example:

During the editorial review, we can also reveal grammatically or graphically instructive wording of questions (so-called suggestive assignment): Jan Amos Komenský's birthplace was:

Uherský Brod

Nivnice

Komňa

Brno

3.7.3 Item Reviewer Form

From a practical point of view, it is beneficial to provide the reviewers with a form that will guide them through the opposition of the test items. Answering the individual questions on the form forces the reviewer to engage with the test item from all the points of view covered by the form. It is not absolutely necessary that each test item completely passes in all monitored parameters; however, the opponent should register and comment on any deviations. Below is an example of such a form for item reviewers.

Table X.X Review of a single best answer question

Item assignment

Reviewer

Yes ✓

or

No ✗

Comments

Tests essential knowledge.

Corresponds to the topic according to the test plan.

It tests the application of knowledge, not just recollection of isolated data.

It corresponds to the required level of knowledge.

The assignment is clearly formulated.

The entry does not contain trick question elements (e.g. double negative).

An expert will think of the correct answer, even if he or she does not know the options offered.

Distractors are homogeneous.

The wording of the options does not indicate the correct answer.

None of the options are disproportionally difficult.

It does not take the form of "which statement is correct" or "all statements are correct except".

It does not contain the words "always", "usually", "rarely", "never", etc.

One of the offered options is the best.

The offered options are sorted alphabetically or in some other logical order.

The options are of similar length and content.

The options are compatible with the question.

3.7.4 Fairness Review

Do the items only measure the specific knowledge or skill required and nothing else?

Every item, every test, should test the required knowledge, know-how or skill and nothing else. By definition, the fairness of a test is the extent to which conclusions drawn from test results are valid for different groups of test takers.

If knowledge and skills are required to answer the question, which for whatever reason were not comparably available to all tested persons, i.e. if all test subjects did not have the same opportunity to acquire the required knowledge or skills, the item is not fair. Such a question is easier for a group of students who have been advantaged in some way, and conversely more difficult for another group who have been disadvantaged through no fault of their own. An example can be the excessive use of technical terms or complex sentence constructions that may not be understandable to everyone. Although the author of the question wanted to verify certain knowledge, in this case he or she is inadvertently testing language proficiency and proficiency in professional terminology. In this context, another complication can be testing the students' attention through "tricks in the question", or the use of double negatives and the like.

The item should not favor any group based on age, sex, origin, social and economic status, religion, race, native language, etc. Since the breakdown into groups is not restricted in any way, it is not realistic to examine fairness for all possible groups in the testing participant population. It is therefore recommended to examine fairness towards those groups that experience or research has shown might be adversely affected. These are often groups that have been discriminated against based on factors such as ethnicity, disability, gender, or native language. Students from different groups with the same level of knowledge should be equally likely to answer the question correctly.

Basic recommendations and rules for the creation of test items and tests with respect to the fairness of the items are given, for example, in the ETS Standards for Quality and Fairness. [43]. These standards recommend verifying that test items:

are not offensive or controversial,

do not reinforce stereotypes of any groups,

are free of racial, ethnic, gender, socio-economic and other forms of bias,

do not have content that would be considered inappropriate or offensive to any group.

The unfairness of items can often be revealed by a thorough review of the fairness of the assignment itself. However, sometimes even an experienced opponent fails to detect it. This is why, when analyzing the test results, we also examine the differential behavior of the items, as we will show in the chapter dedicated to item analysis.

4 Performance of Tests

4.1 Pilot Testing

Credible testing of learning outcomes, especially if it affects students' further progress, presupposes that we know the properties of the test being used before it is actually used in a live setting. Pilot testing and pretesting are used to determine test properties. Technická poznámka

A note on the terminology: The two terms partially overlap. The term pilot testing is mostly used in this book as a broader designation of both steps. If differentiation of the two steps is needed, the term pilot testing refers to a more general "proof of concept" – a kind of feasibility study that reveals possible errors in the concept and design of the test on a small group of students and can also provide useful subjective feedback. The term pretest then refers to a more formal and detailed pre-screening of the test, which makes it possible to estimate the psychometric properties of the questions, their difficulty, the ability of the test to distinguish between stronger and weaker test takers, and which makes it possible to obtain subjective and objective feedback from the tested group.

Pretesting uses comparable procedures for test evaluation to those used to draw conclusions from "live" testing. While a smaller group of students (for example, 20 [44]) with the same level of knowledge and motivation as the target group is sufficient for the pilot run of the test itself, a larger group of at least 100 respondents is needed for the pretest, which is used to calculate the statistical parameters of the items.

In view of the demanding and time consuming nature of building a relevant group, the first "live" run of the testing itself is often used as a pre-test. The inputs obtained from the evaluation of the preliminary tests need to be incorporated into the design of the final version of the test. It is usually necessary to modify at least some items. If the pretest shows significant deficiencies, however, it may also be a case of reworking the entire test concept [45].

4.1.1 Subjective Feedback

Subjective feedback provides very important information from a selected sample of the target group of respondents – typically from selected students. They can help us identify ambiguities or errors in the questions with their subjective opinions. The opinions of each member of the selected group must be taken into account and their comments and suggestions must be considered. The composition of the pilot group should be balanced, i.e. it should not be composed, for example, of only pupils with above-average results, or, on the contrary, expressly underperforming students. Multiple resources are available for the implementation itself. With respect to the efficiency of further processing, the most widespread is the questionnaire format, in electronic form, where the answers can be easily processed and passed on to the working group in a clear format. Below is a list of suitable options for how subjective feedback can be obtained:

questionnaire,

discussion group,

frontal teaching discussion (in the case of a smaller number of students, in greater numbers, this option becomes ineffective),

notes in the test or so-called thinking aloud (see [46]), when students are asked to comment or record their thought processes while solving the test.

4.1.2 Objective feedback

Objective feedback is important for its irrefutability, which is based on the mathematical processing of pilot test results. The conclusions of the objective feedback give indications for the possible modification of unsatisfactory test items. Among the most well-known and widely used test evaluation outputs are:

evaluation of test item difficulty (identification of easy and difficult questions, unsatisfactory items, the possibility of arranging items according to difficulty),

determination of sensistivity of individual items (analysis and adjustment or exclusion of items with undesirable sensitivity)

evaluation of test quality as a whole, primarily its reliability and validity.

When evaluating the results of the pilot group test, we must bear in mind the possible differences between the pilot group and the target group, caused, for example, the different motivations of the two groups. It is a good idea to minimize these differences in advance, e.g. with a suitable “legend” accompanying the pilot test.

4.2 Practice Tests

If students have the opportunity to take practice tests in the subject they are studying, this often has a positive effect on the results of the education – this is the so-called testing effect. For important tests, therefore, a practice test, or “mock test” (UK English), is often organized. This gives students the opportunity to make sure that there will be no problem with the technical side of things (for computer or remote tests), to try out the test format (items with one or more correct answers, etc.), and to check their knowledge and time requirements in advance. These mock tests significantly reduce test anxiety, increase motivation and, through the “testing effect”, the readiness of students. From the organization's point of view, these tests make it possible to calibrate new items, to test the organization of the exams before live implementation, and especially to support students’ preparation by providing feedback on their current performance.

A practice test usually contains the same components as a live test. All parts of the test are evaluated and participants are offered feedback pointing out their strengths and weaknesses. This allows students to learn from their mistakes and gain practice and confidence before the final test.

For the student, practice tests have a number of benefits:

Adopting the right strategy. If the test is time-stressing, the student can become aware of this when taking the practice test and adapt his or her strategy for working with time accordingly (e.g. put off time-consuming items to the end).

Getting practice. High-stakes exams can stress participants and reduce their natural performance. This can be reduced by preparing under conditions similar to the actual exam.

Analysis of own performance. After each test, students should spend time analyzing their mistakes. They should go through each part of the test carefully to find out where they make the most mistakes and focus their preparation on those areas. Using this kind of preparation, students can better understand the questions that might be used in the final test.

A meta-analysis conducted in 2017, as well as other papers, have shown that practice tests and their feedback have a large effect on learning outcomes and can be used as an effective tool to support learning. It turns out that students who participated in practice tests often achieve better results than students who prepared in other ways, such as by reviewing the material, practicing, etc. According to some studies, practice tests are more beneficial for learning than reviewing the material and all other compared methods. Practice tests can therefore be recommended to effectively support learning and as part of feedback for both students and teachers. [47], [48]

4.3 Administering Tests

Administering tests – in-person or remotely?

Thanks to the development of computer technology, it is possible to choose how the test will be administered. It can be a written (paper-based test, PBT) or an online (computer-based test, CBT) test. Each of these approaches has its advantages and limitations.

4.4 Paper Testing

Paper tests consisting of multiple-choice questions saw massive expansion as early as World War I in response to the personnel needs of the US military. It was then necessary to quickly and efficiently classify a large number of recruits, and this could not be achieved in the short time available by the usual individual-basis work of psychologists at that time. [49]

Thanks to its efficiency, preprinted testing quickly spread to other fields that previously relied on individually administered tests – education, intelligence testing, and other areas.

In the English-language professional literature, two terms are differentiated: purely computer-based testing and computer-supported testing. In the second case, the collection of answers can also take place using paper questionnaires (this is paper-based testing), but the tests are then evaluated and further analyzed using computer technology.

Computer-based testing is certainly the direction in which the entire field is moving. Nevertheless, paper testing is important, not only when there is a lack of computer equipment, but also as an easy entry into the world of testing and the use of related methodologies. Appropriately chosen programs and technologies can make our work significantly easier.

In the simplest form of a paper test, freely printed questions with proposed answers are enough. The traditional evaluation of answer sheets using transparencies with a template of correct answers showed a large error rate due to the human factor, often comparable to the number of errors made by the respondent in the answers. With the advent of optical scanners and optical mark recognition (OMR) technology, reading the answer forms is no longer a problem. Corrections and changes to the answers made by the examinee can also be analyzed relatively easily. For automated evaluation, however, the forms must be designed so that they are easily machine-readable, i.e. they meet the requirements for optical mark recognition. Examples of machine-processable questionnaire sheets can be found on the Internet under the terms “bubble answer sheet”, “OMR answer sheet”, or “scantron test sheets”.

Tests can also be generated and printed directly from testing support programs such as the specialized Rogō test program. It supports the printing of machine-readable forms, including the creation of several versions of the test with differently ordered items. Printing test forms is also possible with LMS Moodle, which has the Quiz OMR extension for creating machine-readable forms.

While the printing of test forms is often included in testing programs, the reading and recognition of completed forms is not addressed in the test systems mentioned. It is necessary to use an external solution, such as the proven commercial software Remark Office.

Advantages

Paper testing usually uses pre-printed forms on which the test taker marks their answers. The advantage is that a large number of tests can be administered simultaneously.

Answering on paper is more intuitive and comfortable for some test takers, since it doesn’t raise concerns whether they can manage the technology.

Disadvantages

One disadvantage is the inflexibility of the entire process due to the technologies used.

For example, is not possible to obtain certain information, such as regarding the speed with which the testee responded.

4.5 Computer Testing

Electronic evaluation has largely evolved from conventional forms of evaluation. The original paper tests and answer sheets were converted into digital form and delivered to the test taker either by an application running on a local computer, or, with the development of technology, now more frequently online, via the Internet. A massive increase in electronic testing can be seen especially in the last ten years. [50] Added to this now is testing using mobile platforms. [51] A number of software tools are available for online testing. On the one hand, there are specialized programs that deal only with testing (e.g. Rogō) or test modules that are part of various comprehensive tools (e.g. LMS Moodle).

Advantages

Computer testing is incomparably more flexible than paper testing. It allows you to use multimedia in items, and there is no loss of quality in images. Computer testing also brings a number of advantages for the administering of the test and controlling its progress (e.g. it is possible to set one-way passage through the test, finished itemks can be locked, etc.). A huge advantage for test security is the ability to track how the test taker answered over time and how long each item took. In addition, computer testing opens up entirely new possibilities for adaptive testing. Direct filling and processing of the test on a computer significantly speeds up its evaluation, which is greatly appreciated by students expecting feedback. Electronic testing is generally less error-prone and leads to higher quality assessments. [51] Thanks to computerized testing, new formats of test questions are entering the assessment, using, for example, the possibility to mark the answer in a picture. And finally, computer-based testing is more cost-effective and environmentally friendly than its paper counterpart. [52]

Disadvantages

Compared with paper testing, computer testing is limited by the computing technology available to the operator. A computer system can be infected by a virus or attacked by hackers, or fail due to a loss of power or connectivity. Prior to electronic testing, test takers and staff must be trained in the use of the electronic testing system. Potential dispute resolution with disgruntled test takers can be more complicated because there is no “paper proof” of what the item was and how the student answered. When testing many clients at once, there may be increased demands on transmission capacity, especially in the case of mobile devices with wireless connections. High initial costs (HW + SW) may discourage the deployment of electronic forms of assessment, but this disadvantage is offset by low subsequent operating costs.

4.6 Distance Testing

With the development of distance education, there was a need to also shift testing to a new distance form. In contrast to oral distance testing, which takes most of the methodology from face-to-face testing and only adds a video conference tool for communication, distance testing differs significantly from face-to-face testing. Emphasis on credibility and minimizing the temptation to improve test results using unauthorized aids and sources of information in the teacher's absence comes to the fore.

In order to maintain the credibility of the assessment, the testing needs to be adapted to the distance conditions. There are basically two ways to achieve this: proctored testing and “open book” testing.

4.7 Proctored Testing

There are situations when it is necessary to test candidates who are not all in the same place. In most cases, everyone is required to come to the exam venue. But this may not always be effective, or even feasible.

The classic solution to such a situation is for the examinees to gather in several centers, where the examination board will come to them. The exam can take place simultaneously or at different times in all places. Therefore, parallel test sessions are created, or parallel versions of the test are also used.

A fully qualified examiner may be represented by an invigilator, a proctor. The proctor is not an expert in the area being tested, so he or she cannot be a test evaluator, they cannot provide feedback on the test, and they cannot replace an expert in other roles either. However, they can ensure the environment and conditions for the proper execution of the test.

Proctored testing first gained popularity during the verification of the competence of army officers, when it was necessary to test people scattered around the world, or with international language tests. In a typical arrangement, the examinee arrives at an equipped testing center with trained personnel who verify the identity of the examinee, administer the test, and ensure that the examination is conducted according to the declared rules.

Even the invigilator does not have to be physically present at the testing site. He or she can supervise test takers remotely (remote proctoring). This route makes it possible to use the existing type of tests and complements these with consistent online supervision. The described methodology of online supervision of distance testing developed into a separate discipline about ten years ago – “online proctoring”. Proctoring strives to use technical means to eliminate the risks of unwanted behavior of test participants. It covers the whole process from verifying the identities of the participants, checking the space in which the test is administered, positioning the camera(s), checking the programs running on the computer, to sending and evaluating the results.

Online proctored testing has two main modalities. On the one hand, it concerns testing on a large scale, where economies of scale appear to be an advantage. The second modality is distance exams of great importance, which have higher security standards and are used, for example, for entrance and graduation exams, for certifications, etc. Both types are also offered commercially as a service.

Large-scale proctored testing is usually organized for hundreds to thousands of participants. To deliver the test, it most often uses slightly modified tools for common electronic testing (Moodle, BlackBoard, ...), possibly extended by modules limiting, for example, the opening of other applications. Artificial intelligence is often used for surveillance in such large-capacity systems, which detects non-standard behavior of participants. When choosing software for online testing, you need to take into account how it behaves when the connection is lost. If, for example, the connection goes down during testing in Moodle, all data filled in by the testee is lost. If the connection goes down while testing in Rogō, only the last (currently being answered) item is lost.

4.7.1 Prevention and detection of cheating during distance testing

Proctored test cheating prevention follows the same rules as test cheating prevention, which we discuss in a separate chapter on test security.

A specific feature of proctored testing is the absence of a teacher on site, which may tempt some students to find unethical methods of influencing the result. Therefore, in remote testing, great attention is paid to technical solutions that replace personal supervision and limit the possibilities of cheating. The implementation of surveillance can take several forms, differing in the level of security, the number of people tested and the price.

Live proctoring in real time

Real-time streaming video from a web camera and mobile phone is used for control. This kind of remote surveillance creates a difficult confluence of image transfer requirements. Unlike video conferences, where a high-resolution transmission from the presenter is sufficient, surveillance systems require a high-resolution image transmission from all tested persons. This would lead to the rapid exhaustion of the transmission bandwidth when sharing the full video and thus limit the maximum number of people in one test run. Therefore, powerful video compressions are used, the image is switched between tested ones, or only individual photos taken at random times are transferred.

Proctoring Using Recording and Subsequent Review

The course of testing is recorded and the recording is then evaluated or kept for later evaluation. The advantage is the ability to test larger groups with fewer supervisors. Webcam proctoring has been shown to have beneficial effects in reducing academic dishonesty in online tests. [53] However, the bandwidth issues are the same as with live proctoring.

Proctoring Using Artificial Intelligence

Supervision in this case is two-tiered. In the first level, the tested students are monitored by artificial intelligence, which evaluates the student's behavior in real time. In the event of an incident, it will alert live supervision, which will resolve the situation. The advantage is again the ability to serve a large number of test takers with a small number of supervisors. If the artificial intelligence application runs on the test taker's computer, bandwidth requirements are reduced, opening the possibility of serving really large numbers of test takers.

In addition to supervision, technical solutions to limit the possibilities of illegally obtaining information from the Internet are usually deployed during testing. This involves the use of specially developed internet browsers, such as Safe Exam Browser and others. Programs of this type usually work well on standardized classroom computers, but when used in students' home environments, problems may arise when working with different operating systems and incompatible combinations of other programs.

A special browser is not the only solution to the abovementioned problem. One alternative tested was the JavaScript program PageFocus, which very sensitively and selectively monitors the attempt to open another window and warns the examinee of his or her illegal actions. [54]

4.7.2 Recommended Preventive Measures

If you test all test participants at once, by testing in a “time window”, you reduce the amount of shared information.

Measure the duration of the test in advance and set the time allowance very tightly.

Allow test takers to take the test only once.

Arrange the items in the test assignment of each participant randomly so that the order of the items is not the same.

Do not allow test takers to change answers or go back to previously answered items.

During the test, monitor or limit the activities of test takers using a specific browser.

4.7.3 Proctored Examination Controversy

The much-debated pedagogical problem of proctored testing is that it a priori views the test taker with distrust and with its technical measures it anticipates unfair competition – “head to head…”. There are several disadvantages of proctored online testing and objections to it [55]:

There is an undesirable intrusion into the student's privacy, for example by keeping video footage from the student’s home. Problems arise with the storage of personal and biometric data.

An environment of mutual mistrust and suspicion is created.

Test anxiety increases.

A number of technical tools are put between the examinee and the examiner, each of which can fail (loss of connection, system freeze…). Some testing steps become more complex (e.g. identity verification, workplace verification). Therefore, extensive instructions and scenarios are created for both proctors and examinees. Training everyone involved can undesirably divert attention from the tested field itself, which is what students should focus on prior to the test.

Online testing is influenced by the extent to which the examinee is familiar with the technical means used. It depends on the speed at which the student is able to send the answers, but also the confidence with which they answer and their test anxiety. The transition to remote testing can therefore introduce a significant bias into the assessment.

Currently, the debate whether the benefits of online proctored testing and proctored testing using artificial intelligence can even balance their disadvantages and risks is far from settled [55]. However, sometimes it is unavoidable and proctored testing should be used, even with awareness of these issues.

4.7.4 Alternative Approaches

A large part of the problems associated with proctored testing in higher education is related to the knowledge-based concept of assessment. Distance testing is threatened to a greater extent than face-to-face testing by a handful of threats: identity confusion of the examinee (i.e. someone other than the designated examinee answers the questions), illegal cooperation (i.e. someone helps or provides hints to the examinee) and, in particular, the use of unauthorized sources and tools.

Although current technical means make it possible to verify the identity of the examinee relatively reliably under ideal conditions (e.g. also according to biometric features), when time stress, a large number of test takers and poor connection quality come into play, identity can be easily forged. Especially if the cheater is aware of the technical means used and the limitations faced by the examiner. In such a case, it is appropriate to consider uploading and saving the identification procedure, which gives an additional possibility to clear up any potential doubts.

A major threat to the validity of the results is unauthorized collaboration and the use of unauthorized resources and aids during the exam. If we exhaust the technical and organizational means to exclude them, these risks can be further reduced by appropriate design of the test and test items. Hints (mutual cooperation) are easy for closed-ended itemks. The effect of illegal aids is significant especially when answering knowledge questions.

Therefore, the appropriate approach to take seems to be to construct remote exams in such a way that they above all verify the achievement of higher educational objectives, not just memorization and recall of factual knowledge. When taking the test, specific factual information can most easily be found using a search engine, in notes or in the literature, and quickly be applied. On the other hand, when solving tasks aimed at a deeper understanding or even certain skills, neither searching on the Internet nor copying from books or textbooks will help a less prepared candidate – isolated facts alone are not enough to answer the question. Hinting and cooperation with another person is also more difficult.

A frequent recommendation is that remote exams should be designed as “open book exams” to the greatest extent possible. Currently, there are still not enough studies to support this recommendation unequivocally, but it is well-founded theoretically.

The relationship between the test taker and those administering the test must also be considered. An atmosphere of trust and fairness largely suppresses attempts at cheating. Less strict supervision therefore appears to be a suitable solution especially if the teacher/examiner has been working with the students over the long-term, solid relationships have been established between teacher and students and among the students themselves, and if higher levels of educational objectives are being tested. In contrast, a typical situation where a “hard” approach and strict supervision cannot be dispensed with are highly competitive exams where candidates and examiners do not know each other, such as entrance or certification exams.

4.8 Open Book Testing

A notable contribution of the COVID-19 pandemic has been the rapid development of open-book testing. Open-book tests and exams allow educators to ask questions that cannot be answered based on access to information sources alone. They require higher cognitive skills, information retrieval and processing, and critical thinking instead of memorization. In many ways, the open book exam is closer to normal work experience. Open-book testing prepares students for work in the digital world. [56]

It seems that supervision and restrictions employed to ensure a level playing field in remote testing are not the only possible path to fair online testing. This becomes especially true when stepped up checking by educators causes a reaction on the part of students who, feeling pressured, seek ever more sophisticated ways to bypass the restrictions. This changes cooperation between student and teacher into an unwanted “head-to-head” competition in cheating. This situation is especially exacerbated when testing in an online environment, where, in proctored testing, the supervision is very noticeable, and every flaw in the test security can easily lead to its invalidation.

Social and technological progress also plays a role. New students are “digital natives” and are used to working natively with new technologies. New communication and computing devices are becoming more compact and sophisticated. In the future, it will be almost impossible to prevent students from using them in distance exams, and it will be increasingly difficult to prevent them even in face-to-face exams. The use of online resources during an exam will thus become uncontrollable and their ban will be practically unenforceable. Radical restrictions, such as turning off data connections and mobile services throughout the country on the day of entrance exams, do not seem to be the right (or desirable) answer in our prevailing local conditions. [57]

To maintain a level playing field and preserve academic integrity, we can, in these new conditions, shift the focus of assessment from traditional (closed-book) tests to tests with open access to information. To no longer so much test knowledge itself, but shift attention to testing skills. We use classic tests out of inertia and often also just because we were not able to assess skills before – it's time to start changing that.

The transition to open-book testing means adapting the entire test agenda to the new situation. Classic testing allows you to ask questions focused on the recall of individual facts. If you are considering open-book testing, this means moving to the higher levels of Bloom's Taxonomy. Questions should be asked that require students to apply their knowledge to new situations and use analytical and critical thinking. For this approach to be fair to students, it is recommended to have students practice these more advanced cognitive skills sooner than when they will need them on a test.

It is extremely difficult to rework existing tests into the “open book” form. In practice, it is necessary to abandon all knowledge test items and develop new ones, at higher levels of Bloom's taxonomy, that would test understanding and skills.

Benefits

One of the most challenging, but also most rewarding things you must deal with when switching to open-book testing is changing your perspective on academic education. If you are honest, you will admit that in your own work, you are constantly searching for various information, patterns, and details. These you then apply to a specific situation, synthesize them and gradually create your own work. Our students will do the same in the future and we should prepare them for it. We need to structure the open-book assessment to measure their ability to perform this application and synthesis, rather than testing their memorization of individual pieces of information that they will forget in a month.

So, can open book tests/exams help solve the problem of cheating in distance testing? It appears that they can, as recent systematic reviews have shown that closed- and open-book tests produce comparable results. [56],[58]

But not only that. Open-book tests can help engage students in the processes of reflective and critical thinking. They also foster their digital literacy, critical thinking and lifelong learning processes, all important ingredients for graduates' future employability.

Zagury-Orly and Durning are not alone in thinking it likely that in the future we will see a hybrid model in which students are assessed using a combination of open-book and closed-book tests. The first part of the exam could be closed book and assess students on what they should know without looking at textbooks. The second parts of the exam (open-book) would focus on higher cognitive levels, on skills that are relevant to evidence-based practice. [59], [60]

Risks

One of the risks of using open-book tests is that teachers may not initially know how to design effective test items that require critical thinking. Students may be lulled into the false notion that they will be able to look everything up during the exam and fail to properly prepare for it. There may be a false assumption that the exam will be easy and all the answers can be found in the textbook or from other authorized sources.

Even in an open-book exam, the teacher must define the scope of permitted sources of information, to maintain equality of opportunity.

4.8.1 Recommendations for creating questions for open-book tests

For open-book testing, open-ended item types that give students more space, such as constructed-response questions, will be particularly useful. Use story-based items that require students to apply critical thinking in response to a trigger scenario. Present the data to the students and ask what it might mean in the given scenario. What else could have affected it, how can it be verified, etc.

Here are some examples of open-ended items suitable for open-book testing (sorted according to Bloom's taxonomy levels):

Application

Arrange ... to demonstrate ...

Analysis

Identify the error in the proof or calculation

Explain this situation in terms of theory...

What are the counterarguments...

Why is result A different from result B

What is the relationship between X and Y?

Synthesis

Description of the experiment. What do you expect the result to be?

Describe the next step in this process...

Which method is best for this

Which argument is the strongest?

Evaluation

Assess the situation under this state of criteria

Evaluate, assess, recommend what would be better, ...

What changes would you make?

What would happen if...

In questions, use wording such as: “what is most appropriate” or “what is most important”, which guides students in formulating judgments and stances.

Open Book Testing and Digital Age Competencies

In times of high availability of information, the student's ability to correctly formulate a question becomes more important. The information itself is easily available, but without the ability to assess the situation and formulate the question that arises from it, it is impossible to use it effectively.

The need to navigate in a world where there is a lot of information and its relevance is uncertain comes to the fore. Information often contradicts itself, including scientific studies following an “evidence-based” approach. Orienting yourself in this jungle will be one of the key competencies of the digital age.

5. Evaluating and Grading Students

5.1 Standardization of Testing

Standardized testing means that testing is demonstrably objective, fair, reproducible and valid. At the same time, these attributes do not come from the good will of an individual teacher, but are achieved through the systematic use of demonstrable procedures and methods.

Using explicit and known standards in advance allows teachers to provide students with objective feedback on learning outcomes and thus bolster their motivation. Students perceive standardized assessment to be fairer than other assessments that do not address the comparability of questions and conditions.

Standardization also ensures that the threshold for passing the test will be set according to objective (defensible) criteria, that equal conditions will be ensured for those tested, and that the results will be comparable to each other, regardless of the date and specific examiners.

In the field of assessment and psychometrics, the term “standardization” is used in several senses, which can be somewhat confusing:

Standardization as setting an objective threshold (cut score) for passing the exam. This uses certification methods such as Angoff, Ebel, the bookmark method and others. The point is to set the threshold for passing the exam according to objective and demonstrable procedures, so that the threshold that separates the successful from the unsuccessful cannot later be brought into question.

Standardization as a guarantee of equality of conditions during the test.

Correct selection of candidates and correct assessment of learning outcomes require that the process be objective and equal for all involved. We must therefore ensure that all students receive an equivalent test with the same time limit and all other conditions, and unfairly favoring some examinees is avoided.

Standardization as ensuring compliance with standards.

In order for the evaluation procedures of individual schools and institutions to be comparable with each other, or for individual institutions to issue valid testing certifications, they themselves must adhere to the standards that are key to testing. One such example would be the Standards for Pedagogical and Psychological Testing.

Since the assurance of equality and reproducibility of conditions, procedures and evaluations is not a given, various methodological aids and tools are used for this. For example, to ensure the reproducibility of tests carried out by multiple teachers, at multiple schools, or over a longer period of time, the test team creates methodological material for evaluators, which can be referred to as, for example, instructions for evaluators, test manual, examination committee instructions, methodological instructions for evaluators , instructions for exam organization, etc. [61], [62], [63] The teacher thus receives precise instructions for the preparation, execution and evaluation of the test in order to ensure the reproducibility of the results.

Benefits of Standardization

One of the main advantages of standardized testing is that the results are sufficiently valid and reliable and can be objectively documented and reproduced. This distinguishes them from regular in-school evaluations, which are dependent on a particular teacher. Thanks to standardized testing, it is possible not only to compare the results of examinees across individual schools, but also to compare their performance in different years.

Standardized testing not only provides information regarding an individual’s knowledge, but when aggregating the results of entire tested groups, it can provide other useful information, for example, the possibility of comparing the results of different classes, schools or other groups on a timeline, with relative accuracy.

Risks of Standardization

Through gradually increasing adoration, standardized testing has become an icon in some countries and also used for some assessments for which this format is clearly not suitable. According to some authors, “standardized tests cannot measure initiative, creativity, imagination, conceptual thinking, curiosity, effort, irony, judgment, engagement, goodwill, ethical reflection, and a whole host of other valuable dispositions and attributes. What they can measure are specific skills and knowledge, that is, the least interesting and least significant aspects of education"[64]. Critics of standardized testing point to the uniformity of such an educational model and the production of “assembly line-like” graduates [65]. However, this uniformity is not the result of standardized testing, but of its uncritical use. Another objection is that the overuse and abuse of standardized tests harms instruction by narrowing the curriculum. The use of standardized testing regardless of the objectives of learning leads to the fact that what is not tested is not learned. The method of testing then becomes a model of how to teach the subject. Proponents of standardized testing respond that this is not a criticism of standardized testing, but its inappropriate use.

5.2 Determining the cutoff test score

Finding the threshold to pass a test is often referred to as “standardization”. The objective is to find a line between those whose performance is considered adequate for the purpose for which the exam is intended, and should therefore pass the test, and those whose performance is considered insufficient from this point of view. Determining a cutoff score, like any human activity, will contain some margin of error, which can lead to false positive and false negative decisions. The objective of standardization is to minimize these errors.

The method for determining this threshold should simultaneously be: [66]

defensible,

credible

supported by evidence in the literature,

easy to do,

acceptable to stakeholders.

It is usually not sufficient to merely give points on a written test. For most tests, it is also necessary to say which students passed the test and which did not, or grades must be assigned to individual point gains. Determining the cut score (passing grade) is sometimes an underestimated, yet extremely important step. When compiling a test, it is rather difficult for its writer or writers to estimate how difficult individual items will be for students, and it is even more difficult to “hit” a certain value for the overall difficulty of the created test. Nevertheless, cut-off scores are often determined tentatively, based on the estimation of one or a few teachers. If testing is of greater importance, such an approach is questionable – the test results can be contested by claiming that the evaluation was unreasonably strict, or, on the contrary, that the test was too benevolent and allowed even students who have no business there to further study or practice. Therefore, in the case of standardized testing, the cut-off scores should also be set in a standardized way. As a result, the established cut-off score is substantiated, justified and much more reliable. There are several ways to find the cut-off score using a standardized procedure. The individual approaches differ depending on the purpose of the test, and there are also significant differences in their complexity and demands on qualified experts and their time.

Relative, absolute and compromise methods

Student evaluation can be based on comparing students' performance with each other. We call such an assessment relative. Or the evaluation can be based on the fulfillment of some absolute (independent of the performance of others) criteria. We call such an assessment absolute. Alternatively, it can combine elements of both, and in this case we are talking about the compromise method.

The method of relative assessment is based on the assumption that in large groups there is always a (approximately the same) part of the test takers that are prepared to pass the test. There is a certain optimism in this, because if all the test takers were poorly prepared, the method will still select some part of them as satisfactory. That is why this method is especially suitable where we are not focused on the specific competence of the applicants, but about selecting the best from the given group. A typical use of this method is, for example, acceptance tests.

Absolute assessment, on the other hand, requires test takers to demonstrate specific knowledge and skills that entitle them to pass the test or to perform some activity. An example of this kind of test evaluation is the exit test from a driver's training course, state exams, certifications, etc.

In theoretical considerations about the assessment and classification of students, we can view these two different concepts of assessment as a manifestation of two different views on the purpose of higher education.

In the first case, we can view education as a perennial intelligence test that sorts individuals according to their intellectual skills and work habits. This approach reflects the interest of potential employers to select the most suitable candidates for a limited number of prestigious positions and helps to ensure that the most capable are selected for key positions. In the course of studies, this approach pits students against each other, letting them compete with each other. The evaluation method in this case will be relative evaluation.

The second view is different. It assumes that the purpose of education is to enlighten, strengthen and socialize citizens. According to this view, the educator should not focus so much on sorting students according to skill, but on helping them to find the right view of the world and themselves, with the objective of equipping them with the knowledge, tools and habits that will make them useful and culturally literate members of society. Assessment of students within this concept is based on the fulfillment of absolute criteria and is therefore an absolute assessment.

5.3 Relative Determination of Threshold for Passing the Test

Relative standardization is the method of test evaluation, in which the performance of the tested individual is compared with the performance of the relevant population. This means that it is ascertained whether the tested individual achieves better or worse results than others who are tested. Tests in which the performance of the test taker is assessed in relation to others are called norm-referenced tests, (NRT). For example, SAT tests, which are used as a decisive criterion for admission to many universities in the US, use this approach to evaluate the individual's performance in the context of the performance of others. In our setting, relative standardization comparing the performance of students with each other, is a common part of entrance examinations and various classification tests.

Relative assessment is based on the assumption that the performance of mutually comparable study groups (across space and time) is basically the same.

Advantages of Relative Assessment

Relative assessment is not linked to the content of the test, but evaluates individual participants against each other. So, the advantage is that it prevents inflation of the highest grades, clearly differentiates the best students and it is not necessary to individually standardize each test separately.

Disadvantages of Relative Assessment

Grading students according to relative standardization discourages cooperation and teamwork because students realize they are competing with each other for a limited number of top grades. It also reduces students' motivation to study by weakening the relationship between their effort and their final grade, as it depends not only on their own performance but also on the performance of others. The disadvantages of relative grading include fluctuations in the quality of successful students according to the quality of the group. Especially in smaller groups, it may happen that even students with a level of knowledge that does not meet our requirements succeed. And conversely, some students may not succeed in the test, no matter how well they know the material. Relative standardization can exaggerate insignificant differences, especially in smaller and homogeneous groups. Considering these limitations, we should use relative evaluation mainly in large, heterogeneous groups in which cooperation is not expected. Relative standardization, on the other hand, should not be used in groups consisting of fewer than 40 students.

From the student's point of view, this method is inherently “unfair”, because the grading depends not only on the student's own performance, but also on the performance of others with whom they are compared. It is therefore possible that with the same level of knowledge, a student would be graded better in one year than in another. To minimize this risk and ensure year-to-year comparability, leveling of test difficulty is used, which will be discussed in a separate chapter.

Practical Application of Relative Assessment

With relative standardization, the group is divided up according to the number of points achieved and is graded. A z-score or percentile ranking is used, for example, to determine specific grades. When using a four-level classification scale, the boundaries between the individual classification levels correspond, for example to the z-scores of -2, 0, 2, as indicated in Figure X.X. The setting of the cut-off score in the case of relative grading can be arbitrary, for example on entrance exams it can be based on the capacity of the school for which the entrance exam is being conducted.

OBRÁZEK

Fig. 5.3.1 Relative standardization compares the performance of an individual with other examinees. In doing so, the total score is converted to derived values. To express a student's result in the group, one of the methods of relative test standardization can be used:

The percentile scale roughly indicates what percentage of the tested population performs worse than the student in question.

The z-scale describes how far (as measured by the standard deviation of the data) a given student's score is from the mean.

The T-scale uses the same metric but expresses it on a scale of hundreds.

5.3.1 Percentile scale

The most well-known method of comparing the performance of examinees is to show their performance using a percentile scale. A percentile is determined for the student's result, which roughly tells how many percent of students in the reference group had a worse result than the given student. The percentile thus approximately determines the student's ranking converted to an interval of 0 to 1 (or 0-100%).

When calculating a student's percentile, the number of students who scored worse than the student is counted and half of the students who scored the same as the student are added. Then it is determined how large a part of the total number of students this group makes up. The percentile rank for the person with the i-th worst total score can be derived through the relationship:

```
VZOREC + písmenka v textu
```

where Ni is the cumulative frequency for the given outcome, ni is the frequency of the given outcome, and n is the number of students tested. Cumulative frequency expresses the number of students who achieved a given result or worse.

5.3.2 Z-score

Another method of standardizing a student's result is to calculate his z-score. For a given student, his z-score shows how much his result is above or below the mean (measured in standard deviation units). So, we can simply calculate the z-score as the difference between the student's raw score and the average of the whole group, divided by the standard deviation:

VZOREC + písmenka v textu

Using the z-score, the teacher can easily identify excellent students ($z > 2$) and, conversely, very weak students ($z < -2$). The teacher can also easily compare a student's performance on different parts of the test.

A more detailed analysis of other methods of standardization (e.g. C-scale and others) is available, for example, in Jeřábek and Bílek’s Teorie a praxe tvorby didaktických testů. [67]

5.4 Absolute Determination of Thresholds for Passing Scores on Tests

Absolute assessment (standardization) is a way of evaluating a test, in which the student's performance is compared with absolute criteria – with the requirement to acquire knowledge or skills that they must have in order to be able to consider their knowledge sufficient for successfully passing the test (and the course). The criterion is the achievement of specific knowledge and skill, not the achievement of a certain number of points on the test. For example, we stipulate that after completing a first aid course, the trainee should know the recommendations regarding cardiopulmonary resuscitation, otherwise he or she must re-take the course. Another example of an absolutely standardized test is the driver's training test: it is important not to turn loose drivers on the streets who do not know the basic rules, even if they are among the relatively better ones within a particular group of applicants. Absolutely standardized tests are called criterion-referenced tests (CRT) and are used, for example, in the National Council Licensure Examination (NCLEX) for nurses in the USA.

With absolute test assessment, the line between a successful and an unsuccessful student must be correctly chosen, i.e. the boundary between students who have mastered the given area sufficiently and those who have not mastered it sufficiently. Unfortunately, intuitive or "traditional" procedures and arbitrarily set limits (50%, 60%, 75%, etc.) are sometimes used to determine this limit without further justification.

There are a variety of methods for setting reasonable cutoffs for different types of student assessments. The reader can find an overview of them, for example, in the comprehensive Handbook of Test Development [68].

Among the most well-known methods for determining the cut score based on absolute criteria are the Angoff method, the Ebel method, the bookmark method, the contrast group method, and the cut off group method. Another group of "mixed" methods uses elements of both absolute and relative approaches – this group includes, for example, Cohen's method.

The gold standard methods (Angoff method and Ebel method) are based on the expert opinion of relevant experts who assess the individual items of the test one by one and seek consensus on the probability with which students should be able to answer them correctly. These methods are considered the most reliable, but at the same time they are very laborious and expensive. Therefore, simpler and faster methods are often used and, in case of doubt, compliance with gold standard methods is verified.

Let's now look in more detail at the two most important methods of setting the test pass boundary – the Angoff and Ebel methods.

5.4.1 The Angoff Method

The Angoff method, or its modification pursuant to Hambleton and Plake [69], is based on the concept of a minimally competent candidate. This means a model candidate whose knowledge and skills are just on the lower edge of the permissible minimum. In other words, he or she is the weakest student that should still pass the test.

The test items are assessed by a group of 4-20 experts who have an idea of both the topic and the actual competences of the students, i.e. most often teachers of the given field. As part of the preparatory meeting, the group should be trained in the methodology and familiarized with the required performance standards corresponding to the curriculum, in order to unify the idea of the required competencies. In the next step, the panelists go through one item after another of the test, separately, and write down their estimate on the prepared forms with what probability the least competent candidate should answer it correctly. It is recommended that the first few items be assessed together for training purposes. The experts' results are then entered into a common table. If the estimates for an item differ by more than a pre-agreed maximum allowable deviation between estimates (usually 15%), such a question is discussed with the whole group and consensus, i.e. agreement on the assessment, is sought. Items for which consensus cannot be reached are dropped from the test, as differing expert opinions usually indicate a problem with the item itself.

The required percentage score for passing the test is then determined as the average of the probabilities of successfully answering all the questions on the test. The advantage of this procedure is its objectivity and independence from personal preferences. The disadvantage is the time-consuming and professionally demanding nature of the procedure. [70]

tabulka a obrázek 5.4.1.

Table 5.4.1 Table of expert estimates of the probability of a minimally competent student answering the question correctly.

Item number	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	Average
1	0.70	0.70	0.65	0.65	0.80	0.60	0.70	0.69
2	0.50	0.50	0.60	0.60	0.55	0.50	0.60	0.55

3	0.80	0.75	0.70	0.70	0.70	0.80	0.70	0.74
4	0.70	0.60	0.70	0.70	0.75	0.60	0.60	0.66
...	...	...	...	...	...	...	...	...

Average 0.66, i.e., 66%

Note: If there were to be TRUE/FALSE questions on the test, the lowest possible estimate of success would of course be 0.5, because even a student who does not know the answer has a 50% chance of answering correctly. Analogously, for a multiple-choice question with five options and only one correct answer, the minimum will be 0.2.

In order to set the correct cutoff for passing the test, it is essential in the Angoff method that experts must know the target group well and be able to estimate how difficult specific items will be for that group. Despite the fact that the Angoff method appears to work in practice, some authors debate whether the idea of a minimally competent student is a sufficient anchor for setting the standard. Some argue that it might be more appropriate to consider what is important to achieve (rather than how difficult it is to achieve) and what all candidates should achieve (rather than just what a group of successful candidates would achieve) in determining the required performance standards. The suspicion arises, therefore, that experts will be more likely to imagine an average examinee and speculate how such an examinee would pass the test, rather than looking for a minimally acceptable level of competence in relation to the desired learning objectives. [71]

Using the Angoff method

Test items are judged by a group of experts, and for each item each expert estimates what percentage of minimally competent students would answer the given question correctly. Experts work independently so as not to influence each other. The results are entered in a table in which the rows represent the items from the test and the columns contain the estimates of individual experts.

After filling in the table, it is usually assessed whether the experts agreed in their estimates. Items where the estimate variance is greater than a pre-agreed percentage (typically 15%) should be discussed; often, this reveals ambiguous wording or another problem.

Finally, the average of all estimates in the table is calculated. This average tells what percentage of the total possible number of points a competent student should achieve. In other words, this average indicates the threshold of success for the given test - i.e. the line between “passed” and “failed”. The success threshold divides the set of test-takers into successful and unsuccessful.

Fig. 5.4.1 The success limit determined using the Angoff method divides the set of test takers into successful and unsuccessful (pass/fail). This method is also supported by some testing programs, for example Rogō (which will be discussed in a separate chapter), from which you can receive this very chart.

Support for Angoff's method is included not only in the Rogō test program, but is also offered free (for registration) including instructions on YouTube by Assessment Systems, a manufacturer of programs for testing and test analysis.[72], [73]

Conditions for using the Angoff method

For the successful use of the Angoff method, it is necessary that the participating experts have sufficient experience in the given field and agree fairly precisely on the idea of what the students must be able to do in the given course. Experts must therefore be able to imagine what at least a competent student can do, or should be able to do.

5.4.2 Ebel Method

To set the limit for passing the test, in addition to the gold standard—the Angoff method—the Ebel method is also used. It also uses a panel of experts (most often teachers) who are assumed to be intimately familiar with both the subject being tested and the students’ level.

The method has three stages:

categorization of questions,

the evaluators’ estimate of what share of examinees should correctly answer the items included in the individual categories,

calculation of threshold.

Stage 1

Categorization of questions

In the first step, each item needs to be classified into two orthogonal dimensions, i.e. its difficulty and importance are estimated.

The “difficulty” dimension distinguishes three levels of item difficulty: “easy”, “moderate” and “hard”. Assessors individually estimate the difficulty of each item and assign it to the appropriate category. [74]

The "importance" dimension has degrees of “essential”, “important” and “useful”.

Based on this classification, each item is placed in the so-called Ebel grid:

Fig. 5.4.2 The Ebel grid for classifying items into categories of difficulty and importance. You can find this grid, for example, in the Rogō test program, which offers standardization using the Ebel method.

Phase 2

Success Rate Estimate

In this step, each expert has to estimate what portion of the items from each category should be answered correctly by a minimally competent student. Even in this case, the experts should reach a predetermined agreement, for example the extreme estimates for each of the nine fields of the Ebel grid should not differ from each other by more than 20%. Otherwise, the experts must meet, discuss the differences and repeat the estimate.

Phase 3

Threshold Calculation

The products of these percentage estimates of the success rate of a minimally competent student and the number of questions in each category are then added together. The sum is divided by the total number of items, resulting in the desired threshold of success.

The Ebel method also has a modified version, in which experts evaluate only the relevance of items and group them from essential to useful. The judges then determine the percentage of items in each of the three categories that the borderline student should be able to answer correctly. The correct value for the student's passing of the exam is then the average across the categories. While the traditional Ebel method serves to determine an expert estimate of the minimum result that a student should achieve in order to still pass the test, [75], [76] the modified Ebel method serves rather to prepare a content-valid test. [77], [78], [79].

Example:

Table 5.4.2 Ebel Method – Phase 1

Experts divide the questions according to criteria – importance and difficulty (see Ebel Grid above). The table shows the number of items in each category.

	Essential	Important	Useful
Hard	3	2	1
Moderate	4	2	1
Easy	3	2	2

Table 5.4.3 Ebel Method – Phase 2

Experts estimate the success rate for answering the questions in each category by a minimally competent student.

	Essential	Important	Useful
Hard	50%	50%	30%
Moderate	70%	70%	50%
Easy	90%	80%	60%

Table 5.4.4 Ebel Method – Phase 3

Parameters are calculated for the individual categories:

Number of questions · Estimated success rate for minimally competent student

Essential

Important

Useful

Hard

3 · 0.5 = 1.5

2 · 0.5 = 1.0

1 · 0.3 = 0.3

Moderate

4 · 0.7 = 2.8

2 · 0.7 = 1.4

1 · 0.5 = 0.5

Easy

3 · 0.9 = 0.3

2 · 0.8 = 1.6

2 · 0.6 = 1.2

Finally, we calculate the threshold of success: we add up the estimates for each category from the previous table and divide the result by the total number of questions. In our case, therefore, 13 ÷ 20 = 0.65, i.e., we consider obtaining at least 65% of the total number of points to be a successful completion of the test.

The same calculation, which was broken down into three steps above, can also be performed in one step, as shown in the following table:

Table 5.4.5 Passing Threshold Calculation Table

Difficutly

Importance

(Relevance)

Number of questions (n)

Proportions

(P)

Product

(n · P)

Hard

Essential

3

0.50

1.5

Moderate

Essential

4

0.70

2.8

Easy

Essential

3

0.90

2.7

Hard

Important

2

0.50

1.0

Moderate

Important

2

0.70

1.4

Easy

Important

2

0.80

1.6

Hard

Useful

1

0.30

0.3

Moderate

Useful

1

0.50

0.5

Easy

Useful

2

0.60

1.2

Total

20

13.0

Passing Threshold

13: 20 = 0.65, which is 65%

Note: The example give is for illustration only, in practice we would work with greater numbers of items.

The importance of the Ebel method is also evidenced by the fact that the Ebel grid is part of some testing programs. Specifically, it is directly included in the Rogō testing program, which will be discussed below.

Controversy

Rating the difficulty of the items is necessarily subjective and the result depends on the experience and training of the evaluators. Some papers point to the problems this causes. It seems that it is not easy to achieve acceptable agreement among item evaluators. There also appears to be a systematic tendency for raters to underestimate the difficulty of difficult items and overestimate the difficulty of easy ones. [80] In an experimental study, a group of evaluators was unable to reach a consistent estimate of item difficulty, and when given data on how the item performed on a test, they attempted to revise their judgments to match that data, even when the data were falsified. [81]

Even gold standard methods may not be as robust as they seem. They may depend on the experience and training of the evaluators. [82]

It is therefore advisable to occasionally back-check even proven standardization methods, for example using anchor items and IRT analysis, if only to make sure that they really work as they should. [83]

5.4.3 Bookmark Method

The bookmark method works by delivering the test to a representative sample of participants and calculating difficulty values for each item based on that group. We then rank the items according to difficulty and invite the experts to place a bookmark where they think the cut-off score should be. The average for all assessors is calculated and discussed in the group. All evaluators are then asked to place a second bookmark, the same or different from the first, depending on whether their opinion has changed during the discussion. The level required to pass the test is then based on the average or median of the second score.

Today, we use computers for this, but in the past, it was really often done by printing items into test notebooks and experts literally bookmarking them. In contrast to the Angoff method, with the bookmark method, the difficulty of the items must be mapped in advance based on the administration of the test to a representative sample of test takers.

5.4.4 Contrasting Groups Method

Another procedure for determining a defensible threshold score (cutscore) is the method of contrasting groups [84]. Unlike the previous methods, it does not assess individual items of the test, but only works with total point gains.

The assumption of this method is that we have two contrasting groups available. For example, we present the test to a group of beginners (i.e., students) and to an advanced group (for example, people from the field). A curve is created for each group that shows the distribution of the scores obtained. The cut-off point for passing the test is determined as the intersection of the curves of both contrasting groups (Fig. 5.4.3).

Fig. 5.4.3 Finding the cut-off score using the contrasting groups method. (The figure illustrates the use of the contrasting groups method. The dashed normal curve represents the performance of the novice group. The dotted normal curve represents the advanced group. The vertical black line through the intersection of the two curves represents the pass/fail cutoff scores.)

In a real case, especially with small groups of individuals, the curves will not be so smooth and “normal”. Commonly, therefore, the real collected points are interpolated with a smooth curve and the intersection of these interpolated curves is sought.

Implicit assumptions must be met in every cutoff scoring method. While in the Angoff method we implicitly assume that expert judgments are correlated with item difficulty, in the contrast group method we assume that test performance is correlated with another available assessment method. The objective of the contrasting groups method is actually to evaluate how the test results predict some “gold standard” evaluation of the test takers, i.e., the division of students into contrasting groups. This gold standard might be a teacher evaluation or some recognized performance metric. If a suitable "gold standard" cannot be found, other methods must be used - e.g. the bookmark method or the modified Angoff method.

5.5 Compromise Methods of Determining the Pass Thresholds

5.5.1 Hofstee Method

In principle, both absolute and relative standardization have their limitations. Hofstee, for example, drew attention to this some time ago, when prof. Wijnen from Maastricht University proposed an interesting method of relative standardization (Wijnen's method) [85]: Wijnen assumed that the average student would try his best and should pass, so we can take the average test score as a starting parameter. As a cutoff score, we can then use an arbitrarily chosen value between this average score and its value reduced by two standard deviations. The advantage of this solution is that it corrects the effect of the unreliability of the test because it uses the standard deviation as a measure. Prof. Hofstee was not satisfied with this method, arguing that the procedure does not take into account the absolute results of the test and, like any relative standardization, condemns a more or less fixed percentage of students to failure in advance, regardless of their performance.

Hofstee Method

Hofstee therefore proposed mixed standardization offering a compromise between absolute and relative standardization. In this method, the cut-off score for passing the test is set using expert estimates.

It is assumed that each of the experts is thoroughly familiar with:

the test

the nature of the tested group

the expected level of knowledge of the examinees

Experts must answer two questions:

In what range should the number of students who fail a given test be (e.g.: 10-30% of students from a given group should not pass this test)?

In what range should the minimum score for passing the given test be (eg: The minimum for passing this test should be somewhere between 50 and 60%)?

Analogously to absolute standardization, experts are therefore asked about the threshold of success and at the same time, similar to relative standardization, they are asked about the desired percentage of success. After discussing the suggested values, where the experts can still modify their suggestions, we get 4 values:

the minimum and maximum permissible proportion of failures, fmin and fmax,

the minimum and maximum allowable limits of success, kmin and kmax.

All four values are determined as medians of individual experts' suggestions.

Fig. 5.5.1 The Hofstee method of determining the cut-off score for passing the test

The pass threshold is determined after scoring the test as follows: Based on the test performed, a distribution curve of the test scores is created. Kmin and kmax are plotted on the horizontal axis, fmin and fmax are plotted on the vertical axis. A straight line connecting the intersection of fmax with kmin and the intersection of fmin with kmax is made. The intersection of this straight line with the distribution curve is used as the pass threshold for the test. [86]

The Hofstee method is often classified as a compromise method that tries to resolve the differences between absolute standardization (assessing the percentage of correctly answered items) and relative standardization (assessing the percentage of examinees who passed the test).

The Hofstee method is in many aspects similar to the Beuk method. Both require evaluators to establish cut-off scores directly, without examining individual items, and involve judges' estimates of the performance of the entire group being tested. Both methods need to know the actual distribution of test scores, so they cannot be performed until the test is completed and scored. For both, the necessary expert recommendations can be gathered before taking the exam. Beuk's method, in contrast to Hofstee's method, can propose a cut-off score even if the experts' estimates are higher or lower than the points on the distribution curve of the achieved scores. [87]

5.5.2 The Cohen method

The Cohen method is a very elegant compromise method for determining the cut-off score, combining criterion and relative judgment. It assumes that there are always a few relatively good students in the group of test subjects, who have essentially comparable performance across the groups. They are not the very best, whose variability can be considerable, but the “second best” – excellent students whose results are in the 95th percentile of the given group. It turns out that the scores of these excellent students are a very good and reliable measure of the difficulty of the test. Thus, the first step of the method according to Cohen is to determine the difficulty of the test more or less by methods of relative standardization – by ranking students according to their results and determining how many points correspond to the 95th percentile (i.e. above what point gain the top 5% of students placed).

In the second step, we determine the cutoff score that must be achieved to pass the test. Cohen and Van der Vleuten tracked students' performance on various tests for nine years and proposed that this cutoff score be 60% of the score achieved by students in the 95th percentile. A correction for guessing is then added to the calculation, so its final form is:

$$PM = 0.6 \cdot (P - C) + C$$

PM threshold score (passing mark)

P is the score achieved in the 95th percentile (i.e., how many points above the top 5% of students scored)

C is the score that can be achieved by guessing

Example: The test contained 30 “single best answer” items. Each question could be scored 0 or 1, for a total of 0 to 30 points. Each question offered 5 options. It was therefore possible to get a fifth of the points by guessing, i.e. $30 \div 5 = 6$ points. 80 students took the test. The top 5% of them, i.e. the top four students, scored 27, 26, 26 and 24 points. The score achieved at the 95th percentile is therefore 24 points. Using Cohen's method, we will propose a PM cut-off score:

$$PM = 0.6 \cdot (24 - 6) + 6 = 16.8$$

A student who has scored 16 points still does not pass. A student who scored 17 or more points passed the test successfully.

Fig. 5.5.2 Finding the cut-off score according to the simplified Cohen method (i.e. without correction for guessing).

On the histogram showing the distribution of test scores, the first dashed line (right) shows the score values of outstanding students who placed in the 95th percentile. The middle dashed line shows the cut-off for passing the test, which is at 60% of the performance of excellent students. The last dashed line (left) shows the value of the score that can be achieved by simple guessing.

The Cohen method was validated on a large set of tests and it turned out to have very good results and excellent agreement, especially with the Angoff method. It overcomes the disadvantages of widely used criterion and relative cutoff scoring methods. Its advantages are simplicity, speed and low cost. [88] A disadvantage may be that the point limit for passing the test cannot be announced in advance, but only after the test has been given and scored. This can cause doubts both among students and especially among people responsible for teaching, even though the algorithm by which the cut-off score is determined is clearly and unambiguously announced in advance

5.6 Equalization of Test Difficulty

Equalization of test difficulty (also test levelling) is part of standardization. Its objective is to ensure the mutual comparability of different runs or parallel forms of the test (for example, in individual years, at individual schools, etc.).

Equalizing is a technical procedure to recalculate students' grading from individual runs (parallel forms) of the test so that student results achieved in one run can be compared with student results in other test runs [89].

Leveling of difficulty is an important aspect of testing quality and directly affects its validity. It is an essential tool in educational evaluation, as it plays a vital role in determining the validity of the test in all forms and years.

Two procedures are used when comparing tests with each other: linking tests and equalizing tests. The linking of two tests means that we create a relationship (prolinking) between the results of these tests. E.g. we can create a table of corresponding scores from both tests, always achieved by students of the same level in both tests. Based on this table, we can say that students who scored X on the first test will most likely score Y on the second test.

The claim that the difficulty has been equalized is much stronger. If the two tests under consideration were successfully equalized, then we can state that students who scored X on the first test and students who scored Y on the second test have very similar levels of knowledge and skills measured by these tests.

In other words, to say that two forms of a test are balanced (equivalent) means that they measure the same content and support the same conclusions about what students know and can do. If, on the other hand, we say that there is a link between the two tests, this is a much weaker claim, which only means that there is a statistically measurable relationship between the scores on the two tests. This is because the fact that students scored X on the first test and scored Y on the second does not mean that the two tests are really measuring the same thing (the same construct). The linking of tests is therefore not a sufficient argument for us to replace one test with another. To do this, it would be necessary to verify that the tests are equivalent, i.e. to obtain confirmation from experts that both tests cover the same domain with the same means.

Equalizing the difficulty of tests can either precede the test (pre-equating) or follow it (post-equating). Pre-equalization of test levels refers to the preparation of a new test so that its format, content and characteristics correspond to the initial test. In the case of additional leveling of test difficulty, the test can also be compiled according to the rules for preliminary equalizing, but the final equalizing is carried out only with the help of data obtained from the analysis of the completed test.

To balance the difficulty of the two tests, we need some comparable data. One possibility is to give both tests to a sufficiently large group of people and compare the results. To limit the effect of test order, the group can be divided and each half will receive the tests in reverse order. The disadvantage of this approach is the impracticality and time-consuming administration of two tests. The security risk also increases, as the exposure of two tests increases the risk of their items being divulged.

To limit these negative aspects, we can use the so-called anchoring of the test, which is where a certain number of items are included in the test that are the same in all versions. These so-called anchor items are then used to compare different versions of the test. Anchor items should be representative, should cover the range of test difficulty and should comprise at least 20% of the test length [90]. The selection of anchor item topics should replicate the content of the entire test. A set of anchor items can be considered a “mini version” of the entire test. [89].

Anchor items can be either “intrinsic” or “extrinsic,” depending on whether or not they count toward the test score. They can be “embedded“ if they are scattered throughout the test, or “attached” as a separate block of items at the end of the test.

There are many methods for equalizing tests.

Linear equalization is a tool for establishing equivalent scores between two parallel forms of a test within classical test theory. Linear equalization assumes that the tests differ only in the value of their average raw scores and the variability of the results (i.e. the size of the standard deviation). Under these assumptions, we can convert scores from one test to another using a linear transformation. So, we can first transform the mean score of the second test to the mean score of the first test and then transform the value of the score of the second test one standard deviation above and below the mean. The result is a linear transformation of the scores from the second test to the point scale of the first test. This method has several limitations:

Linear equalization will not work in cases where the relationship between test results is not linear (e.g. with an asymmetric distribution of scores).

The transformation applies only to the test set for which it was calculated.

The transformation works best for scores that are less than one standard deviation away from the mean.

The advantage of the linear transformation is that it is easy to understand and computationally simple.

If we would like to use a more robust calculation that also works for students at the edges of the investigated skill range, we can use, for example, equipercentile equalization.

The equipercentile method provides greater accuracy in aligning results across the entire range of possible outcomes. In this equalization of results, we first determine the percentile order of the scores achieved on both tests. The percentile ranks between the two tests are then matched using a table. The second option is to first convert the raw scores to percentiles and then score them (for both tests together). A number of computer programs offer the ability to calculate equivalent scores or establish a percentile ranking for all scores achieved. Percentile ranking is also often used to communicate results to students. Disadvantages include that, like linear test alignment, equipercentile is dependent on a specific selection of students, and is not readily applicable to other groups. Both methods mentioned so far are similar in many ways. Sometimes the linear adjustment is referred to as the equipercentile approximation. [91]

IRT-based equalization methods. In practice, methods based on item response theory are more widely used, which have proven to be more accurate and reliable than methods derived from classical test theory and do not include dependence on a specific group of test takers.

IRT-based test equalization methods can be divided into two groups:

methods of equalizing observed scores,

methods of equalizing actual scores.

In the first case, the actual scores in the two test forms are compared. Based on the knowledge of the behavior of the anchor items present in both test forms, we transform the scores of the second of the tests so that the difficulty of the anchor items in both tests converges. In the second case, the estimated distributions of total scores in two forms are derived from the IRT model, where we plot the characteristic curves of two or more compared tests in one graph and equalize them using the equipotential alignment methodology. [92]

One of the limitations of IRT-based test equalization methods is the required number of test takers, which should not fall below 500. Estimating parameters in small sample conditions is not satisfactory and worsens with the complexity of the IRT model.

For equalizing the difficulty of tests based on IRT, the IRTEQ free software is available [93], or the R equate package can be used. [94]

5.7 Student Grading

Grading generally means sorting, classifying examinees into classes according to some criteria. In education, grading is understood as an evaluation of student results. If we are working with a specific test that measures some knowledge or skill, we are looking for an objective, reproducible, and fair procedure to grade student performance on the test.

In this sense, grading is a continuation or rather an extension of standardization in the sense of setting the threshold for passing the test. The establishment of the grading scale, i.e. the setting of the relationship between the performance on the test and the grade level, is the only subjective element that enters into the entire testing process. Therefore, it should be given due attention [95].

Note: In the text devoted to the grading of student results on the test, we will limit ourselves to the discussion of closed-item tests with a choice of answers. The classification of other types of test items is discussed in the literature (e.g., [96] or [97]).

In order to establish a relationship between test performance and grades, it is necessary to clarify what the task of the test should be. The reasoning is similar to choosing between relative or absolute cutoff scoring methods. So, we are considering whether to let the students compete with each other and classify them into “performance groups” by comparing them with each other, or if we grade according to the extent to which the student has achieved the target competencies.

5.7.1 Grading based on Comparison with Group Performance

Grading by comparison with the performance on the group (relative grading – norm-referenced) is based on the performance of the student in the context of the group. It derives the grade from the student's ranking within a certain group. We rank the students according to their performance on the test and then assign them grades according to pre-agreed limits. Relative grading assumes that the performance of different study groups (across space and time) is basically the same. From the student's point of view, this method of classification contains an obvious injustice, because the evaluation depends not only on their own performance, but also on the performance of others. It is therefore possible that if the student had been in a different group (for example, if he or she had taken the test in a different school year), they would have received a better grade with the same level of knowledge.

If we want to use an evaluation based on relative standardization, two decisions need to be made. First of all, it is necessary to determine what grade we will assign to the average performance. In the frequently used four-level grading system A, B, C, D, we can intuitively choose the border between B and C as the grade corresponding to average performance; however, that is not the only option.

Furthermore, it is necessary to decide in advance on the boundaries separating the individual grade levels. For example, z-scores or percentile rankings are used to determine specific grades based on performance, in a similar way to that described in the chapter on relative test-passing cut-offs.

For a four-level grading scale then, the boundaries between the individual grade levels correspond to, for example, z-scores –2, 0, 2, as indicated in Figure 5.7.1:

Fig. 5.7.1 Example of grading a student's result on a test using z-scores (relative grading). The group is divided according to the test scores in such a way that the resulting subgroups are separated from the average by an agreed number of standard deviations. We grade the subgroups created in this way according to the relevant classification scale – in this case four-level. Note that the highest grade “excellent” is given in this case to only 2.2% of the test takers, the average grades “very good” and “good” are given to a large majority of students (each of these groups comprises 47.7%) and again only a completely negligible number of students (2.2%) gets the lowest grade of “fail”.

Splitting into subgroups based on standard deviations from the mean (z-scores) is not the only option. An alternative is to divide the group according to the achievement scores into subgroups of equal size, and give these subgroups the same grade. For example, the distribution could be such that the top 25% get an “excellent” grade, another 25% get a “very good” grade, etc.

Advantages and Disadvantages

Grading systems based on student comparisons are simple and easy to use.

They work well in situations where students need to be lined up, for example as part of entrance and admission tests to a department of study in which there is a limited number of places.

They are appropriate for large courses that do not encourage collaboration among students, but generally emphasize individual achievement.

The obvious disadvantage is that individual are graded not only on results, but the results of other students also determine the student’s results.

The assessment threshold can only be established after the test has been given. Therefore, it is not possible to comment in advance on how difficult the test will be (although it is known in advance how the threshold will be determined).

Relative assessment will be more applicable in large non-selective groups that will be representative of the entire student population. In small classes (under 40), this group may not be a representative sample. One student may get an excellent grade because he is in a low-achieving group, while their classmate with the same result in a better group gets a lower grade.

A second objection to grading in relation to others is that it encourages competition rather than cooperation. This method of assessment sets up a relationship of direct competition between students. When students are pitted against each other for a few top grades to be handed out, they are less likely to cooperate with each other in their studies.

A compromise solution for small groups is to use so-called “anchoring” in the relative evaluation. This means grading will be adjusted according to the overall (average) level of the students in the group. [98] If a teacher has used a similar test repeatedly in different years, he or she can use the accumulated test results as an anchor. The teacher then compares the current group with this collected large group. Similarly, a well-constructed pretest can be used as an anchor, in which we estimate the ability of the entire group using absolute criteria. Modifying the relative grading system with anchoring helps reduce feelings of competition between students because then they are no longer competing only with each other.

5.7.2 Criteria-Based Grading

Absolute grading (criterion-referenced) measures the student's success in relation to the criteria required to achieve that grade level. Usually, the criterion is the number of points or a percentage of the total number of points that the student must achieve in order to receive the appropriate grade. The simplest way to grade is to determine what percentage of the total number of points is needed to achieve a certain grade. For example, an A grade requires 90% and up, a B grade 80-90%, and so on. The problem with this approach lies in the arbitrary setting of the boundaries between individual grades. If we set similar point boundaries in advance, the author of the test must “hit” within them. If the test or one of its versions is more difficult or, conversely, easier than the creator of the grading scale assumed, the resulting assessment will also be perceived as disproportionately strict or, on the contrary, benevolent. Therefore, for more important tests, it is advisable to set the borderlines using the estimates of a larger number of experts, for example according to the Angoff or Ebel method.

In absolute grading, unlike relative grading, a student's grading is not influenced by the performance of others and is not based on mutual comparison with students in a group. If we were to test a group of significantly above-average students, they might all get good grades, and conversely, if a group of weak students happened to come together, no one might get good grades. Students are not competing with each other and are therefore more likely to work together. This can also be beneficial for their active involvement in learning, which is often based on cooperation. The grading of an individual student is not affected by the overall result of the class.

Absolute and relative grading are actually somewhat intertwined. Most teachers set criteria based on their experience with typical student performance. This brings relative elements into the absolute grading. Similarly, teachers sometimes retain some flexibility in absolute grading by telling students in advance that the criteria on the first run of the test may be relaxed if too few students score well. For example, the 90% threshold for obtaining an A grade may be reduced to 85%. If the test would be more difficult for the students than the teacher imagined, he can reduce the assessment criteria in this way. The opposite procedure, where the teacher would tighten the criteria because too many students achieved a good grade, is not recommended.

Another way to grade students according to criteria is to set course objectives and assign grades based on how well the student has achieved them (e.g. A = student has achieved all major and minor course objectives, B = student has achieved all major and several minor objectives, etc. .).

A more sophisticated form of absolute classification distinguishes between different types or levels of knowledge and skills that a student demonstrates on different tasks. Greater emphasis is placed on those that reflect higher levels of mastery of the material. This approach reflects both the amount of material and its level of cognitive complexity. For example, we can divide the learning objectives of our course into two groups: basic and advanced. The basic objectives refer to the minimum necessary knowledge and skills that students must acquire. Advanced objectives, on the other hand, represent higher levels of skills, such as using critical thinking, solving complex problems, and the like.

It may be easier, at least initially, to use two completely separate tests to determine how well the basic and advanced learning objectives have been achieved. This will make it easier to evaluate the exam and keep records of it. By separating the tests, it is also easier to focus on the individual learning objectives and prepare test questions for them. It tends to be relatively easy to assess the basic learning objectives. Assessing the extent to which the advanced learning objectives have been achieved is usually more difficult, as it is more difficult to devise test items covering the ability to apply the acquired knowledge.

Different requirements for student performance can be set for passing both types of test, as indicated in Table X.X.

Table X.X Example of a possible setting of absolute standardization for grading a basic and an advanced test on a five-point grading scale.

Grade level

Basic test

		Advanced test
A	90% or above	85% or above
B	90% or above	75–84%
C	80% or above	60–74%
D	80% or above	50–59%
F	less than 80%	less than 50%

In the given example, we require students to demonstrate mastery of at least 80% of the basic learning objectives and 50% of the advanced objectives. If we require the setting of success thresholds to be more objective, we can use one of the expert estimation methods described above.

From a higher education perspective, criterion assessment is the most desirable. Although it is more demanding for teachers, it requires thinking about the expected learning outcomes, but it is transparent for students and the derived grades should be defensible from a reasonably objective point of view – students should be able to trace their grades according to specific performances in solving set items. Criterion evaluation, with its transparency, creates an important framework for the involvement of students in the learning process.

With absolute evaluation, it is also appropriate at the same time to monitor the distribution of grades in the study group – in other words, to monitor the results of the criterion grading model from the perspective of the relative evaluation model. If we find that too many students are getting poor grade, or, on the contrary, good grades, or the distribution is skewed in some way, then this may indicate that something is wrong and that the grading process needs to be reviewed. For example, this may be a problem with the overall difficulty of the assessment items (for example, unchallenging exam questions or too few questions, or items that do not differentiate between students of different knowledge and skill levels). Best practices for grading in higher education are based on predominantly criterion-referenced assessment, slightly modified by relative correction and feedback. [99]

5.7.3 Grading Systems and Converting Scores

It is often useful for the results of the student's assessment to be expressed in a way that allows comparison between subjects, possibly also between fields of study or universities themselves. Therefore, standard grading systems (marking scheme, academic grading, ...) are created [100]. These systems allow comparisons within individual universities, but sometimes also within entire countries. The results of specific tests, or of entire sets of written and other works, are converted to a standard scale, according to which grades are then awarded.

As an example, we can take the grading systems of the University of Edinburgh. For example, for undergraduate medical courses, the CMS3 system (CMS3:Bachelor of Medicine and Bachelor of Surgery) is relevant:

Table 7.9 CMS3 System

Points	Grade	Description

90–100

A

Excellent

80–89

B

Very Good

70–79

C

Good

60–69

D

Pass

50–59

E

Conditional failure (may be re-evaluated)*

0–49

F

Fail

- Conditional failure is a form of challenge to the student, to correct the grade within a set deadline. If the student fails to do so, the worse grade is recorded.

The assignment of a grade and the recalculation of a score can be seen on the following example: Consider a test in which students could get from 0 to 50 points. Using standardization methods, the creators of the test determined that to succeed on the test, you need to get at least 24 points out of a possible 50 (the so-called pass mark). The grading system marks the worst grade that corresponds to passing the test with the letter D and assigns a numerical value of 60% as the threshold of success. In this case, a raw score of 24 points out of a possible 50 corresponds to a recalculated score of 60%.

Fig. X.XX Conversion of test scores to grades

Through standardization methods, the threshold for success on a specific test was determined at 24 points out of 50. In the grading system used by the given institution, this threshold gross score corresponds to a converted score of 60% (threshold for grade D). Results better than 24 points out of 50 are then distributed equally to the individual grades.

After we have determined the conversion for cutoff scores, we determine how higher point gains will be converted. A simple linear conversion of the raw score to the converted score is usually used. In this case, a converted score of 70% (minimum for grade C) will correspond to a raw score of 30.5 points out of 50, a converted score of 80% will be achieved by a student with 35 points out of 50, etc. In other words, we first determined which students will pass the test, and then we mechanically divided them into individual classification grades. The conversion can be expressed mathematically as follows:

```
VZOREC
```

where is the converted score (from which we determine the grade according to the CMS3 system), is the minimum gross score required to pass the given test (pass mark) and is the gross score achieved by the given student.

Similarly, we can also convert a raw score lower than the pass mark. Conversion in the entire range of possible point gains can then be doubly linear.

Fig. X.XX Doubly linear (linear dogleg) conversion of raw scores to converted scores.

In this test, standardization methods determined that the cut-off score should be 36 points out of 50. This should correspond to a converted score of 60%. Higher point gains are linearly converted so that a raw score of 50 points out of 50 corresponds to a converted score of 100%. Likewise, lower point gains are linearly recalculated so that a raw score of 0 points out of 50 corresponds to a converted score of 0%.

This transformation takes the form of two connected line segments, but there is a slight bend in the line. It is therefore referred to as "dogleg" in the literature. [101]

5.7.4 Scaling

A more complex way of converting scores to another scale, more suitable for reporting, is scaling. Unlike the previous approach where the recalculation was based on three points (0%, success threshold, and 100%), scaling is more detailed. It primarily enables us to communicate the results of different parallel forms (versions) of the test, which may differ slightly in difficulty (see also Equalization of test difficulty) in a comparable way.

A number of important tests such as the ACT, SAT, GRE, and MCAT are reported on scales that are deliberately chosen to convey certain information. The SAT and GRE historically have a nominal mean of 500 and a standard deviation of 100, while the ACT has a nominal mean of 18 and a standard deviation of 6. These are actually the same scales because they are nothing more than converted z-scores.

The "mean values" were chosen arbitrarily, and then the boundaries of the score range were set using multiples of the standard deviations. As a result, the SAT and GRE scores range from 200 to 800, and the ACT scores from 0 to 36. To make the test taker feel better, the scales are set so that he or she receives 200 points for "submitting the form" on the SAT. A result of 300 points may seem like an encouraging number, but it is only 100 points above the minimum, which corresponds to only the 3rd percentile.

Often the raw score achieved in the test is not reported at all, but only some converted score. If there are multiple versions of the tests being compared, the scaling compensates for the fact that the versions differ in difficulty. The chosen scoring scale should be at least as wide as the number of items in the test, otherwise some of the distinction that test results bring is lost.

The starting point, when scaling, is usually to define the range in which the communicated results should lie. We usually begin by finding the mean and standard deviation of the raw test scores, which are then converted to another, recalculated mean and standard deviation. The already mentioned linear and bi-linear conversions may not be enough, therefore more complex transformations are also used. Equipercentile transformation is suitable, for example, for equalizing parallel forms of the test (see chapter - Equalizing the Difficulty of Tests).

6 Analysis of The Test and Its Items

A summative didactic test can be understood as a tool for measuring the level of knowledge and skills that a student has acquired during the learning. The results of decisive testing can have fundamental consequences for the test participants – for example, acceptance or rejection for further studies, certification for a certain profession or the award of a degree. If the tests were inadequate for their purpose and did not measure the qualities we expect them to measure, substantial errors in decision-making could occur, thereby reducing the effectiveness and jeopardizing the credibility of the entire system. It is therefore important to measure and continuously monitor the quality of tests and test items.

Part of the properties of tests (and items) can be described using intuitively the comprehensible categories of difficulty and sensitivity. Difficulty can be understood as the probability with which the test taker will not answer the given test or item correctly. Sensitivity refers to the degree to which a test or item differentiates between better and less prepared students.

In addition to these intuitive metrics, we also use the terms reliability and validity to describe test properties. Reliability expresses the accuracy and repeatability of the test. Using reliability, we actually find out whether retesting the student with a different version of the same test will lead to confirmation of the previous result. Validity tells whether a test or item measures the knowledge we want to measure.

In addition to these traditional metrics, considerable attention has been paid in recent years to the fairness of tests. It is verified whether the test does not somehow disadvantage certain groups of test takers.

Item analysis makes it possible to evaluate the characteristics of individual test items, especially their difficulty and sensitivity, based on the analysis of the completed test. Item analysis can also include the analysis of distractors, which examines in more detail the quality of the options offered in closed-ended (multiple choice) items. It deals, for example, with how the test takers chose individual suggested answers depending on the test taker's overall performance.

Item analysis provides a range of psychometric data for each item, which makes it possible to construct independent tests with similar properties.

Test analysis should include its descriptive statistics and graphical display of the results, most often in the form of histograms. Comparing graphs from individual test runs will help us to assess whether, for example, some items used in the test were leaked, etc.

Let's first look at the properties of the test as a whole, especially its reliability and its validity.

6.1 Reliability

Ideally, the test result should depend solely on what we want to test for, that is, the score obtained on the test should depend only on the skill of the examinee in the area tested by the test (the so-called real score). But in real life, the test result (raw score) differs from the actual score due to more or less random errors. Every test therefore has a certain reliability and accuracy, which we express as reliability (reliability, precision, reproducibility) [102].

Reliability tells us to what extent we get similar results for repeated independent evaluations of the same individuals. It is influenced, for example, by how well the testee understands the assignment of items, especially if they are complicatedly formulated and the examinee is from a different cultural or linguistic background. The test result also depends on the attention of the test taker, which will be affected by the environment in the room and distractions during the test, or whether the test taker is working under stress. Reliability is also reduced by possible guessing of answers, etc.

Reliability takes on values between 0 and 1 (0% and 100%). To simplify, reliability can be imagined as a measure of suppression of random errors expressed in percentages. A reliability of 50% means that approximately half of the variability of the observed score (raw score) is the variability of the actual score (i.e., the measured skills of the test taker) and the other half is due to chance errors. A reliability of 0.8 means that 80% of the variability in observed scores is due to variability in true skills and 20% is due to errors.

The minimum test reliability rating that can be considered satisfactory depends on the context, e.g. the number of items on the test and the number of test takers. As for the number tested, several recommendations have been published that agree that for a reasonable estimate of reliability, the number should not fall below a few hundred. [103] If the number of participants were significantly lower, it is possible to work with overall utility instead of reliability, as already introduced by van der Vleuten. [104], [105]

For the purposes of pedagogical distinction of individuals, e.g. when deciding on admission to further study, a reliability coefficient of at least 0.8 (and higher) is usually required. For other school practices, a reliability coefficient in the vicinity of 0.6–0.7 is sufficient. [102]. For tests with a small number of items (10 or less), reliability usually does not exceed the value of 0.6-0.8. A lower reliability value does not necessarily mean that the test is downright bad, but it must be treated with caution and should not serve as a stand-alone basis for decision-making. On the contrary, a very high reliability coefficient (close to 1) may mean that the items on the test are so internally consistent (so similar to each other) that they are interchangeable and there could be fewer of them in the test without significantly impairing its properties.

Reliability describes the technical quality and internal consistency of a test, but not its correctness. A test can be reliable – have high reliability, while it may not measure what it is supposed to, meaning it simultaneously has low validity. However, the reliability of the test is a necessary prerequisite for its validity.

The following example provides a good illustration of the concepts of reliability and validity and the relationship between them:

Fig. X.XX Diagram approximating the relationship between reliability and validity

6.1.1 Reliability Estimates

In principle, reliability cannot be calculated directly, but we can try to estimate it. When estimating reliability, we try to determine how much variability in test scores is due to variability in actual scores and how much is due to measurement error. (Recall that measurement errors can have random and systematic components.) The objective is to design tests so that sources of error are minimized.

Four main approaches are used to estimate test reliability depending on the situation: [106],[107]

Reliability as agreement between evaluators (inter-rater reliability): This so-called classification consistency is used to assess the extent to which different raters provide mutually consistent estimates of the same phenomenon. It is used especially where subjective factors enter into the scoring of the test. A condition for objectivity is comparable training of evaluators, which unifies the required criteria. It should be kept in mind that high inter-rater consensus does not mean that the test taker will perform the same on a repeat test. Consensus between the assessors and the consistency of their assessment is therefore a condition, but not yet sufficient to guarantee the high reliability of the scores of the tested persons. [107]

Test-retest reliability: This is something used to assess how consistent the results of the same test are when repeated on the same group. Their consistency can be assessed by calculating their correlation. While this method gives excellent results for phenomena where repeated measurements of the same quantity are independent of each other (measurement of length, weight, ...), it is difficult to use for didactic testing. This is because individual test runs cannot be considered independent. With a short break between tests, participants may remember how they answered the first time the test was run, and the resulting reliability will be overestimated. A gap of at least 3 months is therefore recommended, although even there there is a risk of distortion, as students can learn the material in the meantime. When repeating the test with a longer time interval, students may already forget the material and the result achieved will necessarily differ. This “optically” reduces the true reliability.

Reliability of parallel test versions: This approach is used to assess the consistency of the results of two tests created according to the same rules, in the same way, on the same topic. Assessing the reliability of parallel versions of the test (by calculating their correlation) removes the problems with independent repetition of the test that we saw with the test-retest method, but brings new difficulties with the creation of equivalent forms of the test. Parallel forms should be created according to exactly the same test design and their items should have the same psychometric characteristics. Sometimes there is an attempt to create “parallel” items by changing the numerical values in the examples, changing the names and titles in the item text, etc. In practice, however, it turns out that the newly derived items are usually more difficult, so it is necessary to already create pairs of items when writing them, and then randomly use them in the tests.

Reliability as internal consistency: This approach assess the consistency of results across items within a test. In the previous paragraph, we discussed the assessment of the reliability of parallel forms of the test, i.e. the correlation between the test and the parallel (repeated, but independent) test. Since it is difficult to create a parallel independent test, a random split of one test into two halves is used as an approximation (a substitute for a parallel test). We then consider the resulting halves as two independent parallel tests. The correlation between the two halves (corrected for test length) is a good estimate of the “true” test–retest correlation. The problem with this approximation is that we do not know the effect of randomly dividing the test into halves. Perhaps a different split into two halves would yield a different correlation and thus a different estimate of test-retest reliability. While we could alternate between all possible splits and then take the mean correlation as a measure of reliability, this could be very laborious in a multi-item test. It is simpler to divide the test into the smallest possible parts (individual items) and calculate the correlations between them. This approach is a good measure of internal consistency and the basis for the widely-used Cronbach's alpha. [108] Cronbach's alpha can be taken as the average of reliability estimates for tests divided into all possible halves. [109]

6.1.2 Cronbach's alpha

Cronbach's alpha was developed by Lee Cronbach in 1951 to provide a measure of the internal consistency of a test, that is, the extent to which all items in a test measure the same construct and the spread of the test's measurements. If the items in the test correlate with each other, the alpha value increases. The value of the alpha coefficient is also affected by the length of the test. If the test is short, the alpha value decreases. The alpha value is a property of a specific test performance – it depends on the composition of the specific group tested.

When interpreting Cronbach's alpha, it should be kept in mind that the concept of reliability assumes that the test is homogeneous in the sense that the test items examine the same latent trait on the same scale. If this assumption is violated, the reliability estimate may underestimate the test’s true reliability. For multidimensional tests, alpha should be calculated for each measured construct separately. If we are not convinced of the unidimensionality of the test, we must look at Cronbach's alpha as the lower boundary of the reliability estimate. Estimates of acceptable numerical values of Cronbach's alpha range within broad limits (from 0.70 to 0.95) [110].

A low number of questions, heterogeneity of the measured construct, or low correlation between items can all result in a low alfa value. The easiest way to determine the cause of a low alpha is to calculate the correlations of individual items with the total test score. Items with a low correlation (approaching zero) are unrelated to the rest of the test and can be removed.

If Cronbach's alpha is too high, this may indicate that some items are already redundant in the test and do not provide any additional information. The maximum recommended alpha value is 0.90. [111]

The following example can demonstrate how Cronbach's alpha is used:

Let's imagine that we want to test on adding numbers from one to ten. We can easily set up a test in which there will be a large number (say fifty) of supplementary items of the type “3 + 4 = ...”. A person who knows how to add will answer all the questions correctly, or at most will make only a few random mistakes. On the other hand, those who cannot add at all only rarely manage to find the correct solution. The test constructed in this way can be described as internally consistent – it tests a single concept (addition in the given range of numbers). Cronbach's alpha will be close to one.

If we now replace half of the items on the test with examples of the type “12 ÷ 3 = ...”, the situation changes. If we give this modified test to students in the first or second grades of elementary school, we will test two concepts: addition and division. It is possible to imagine that some students will be good at addition, but they will be very poor at division. The test will no longer be as consistent as in the previous case; we can no longer say that any two tasks test the same thing. Cronbach's alpha will decrease.

If we talk about the internal consistency of the test, we should realize that it depends not only on the items themselves, but also on the target group. If we were to give that modified test with simple arithmetic tasks to high school students, it would probably appear internally consistent again and Cronbach's alpha would be close to one: from the point of view of such a more advanced group of test takers, we are once again testing a single concept – basic arithmetic tasks. Whether the particular item is dedicated to addition or division will not make a difference in this case.

The above examples show why the Cronbach's alpha of a particular test should be neither too low nor too high. If the test is inconsistent, we will misinterpret its point results. Let's imagine that we give our test with addition and division problems to second graders. According to the number of points achieved, it is probably quite easy to recognize a group of those who can add and divide well, and a group of students who cannot add or divide at all. Among them there will be pupils who add and divide, but with many mistakes, but also those who add excellently, but cannot divide at all. The results of this kind of test do not tell us whether a particular student passed comparably in both activities, or was excellent in one and failed in the other; it would probably be better to use two separate tests, each focusing on a different skill, instead of one.

Conversely, if Cronbach's alpha is close to one, it means that many students in that group either answered all questions correctly or answered all questions incorrectly. In other words, if the student answered the first few questions correctly, he also answered all the others correctly and vice versa. In the said test, comprising only examples of addition, it would probably be pointless to give the students fifty questions – if we shortened the test, we would probably get completely comparable results. In addition, a test with a very high Cronbach's alpha may not sufficiently distinguish between different levels of knowledge.

Although Cronbach’s alfa is widely used it is necessary to remember all its limitations.

6.2 Validity

Validity (correctness, truthfulness, faithfulness, soundness) describes the extent to which a test measures what we want it to measure. The validity of a test refers to the extent to which conclusions based on its results are meaningful and useful. That is, if the test is designed correctly and if its result is not too affected by systematic errors.

By definition, the validity of a test is the extent to which both theory and gathered evidence support the proposed interpretation of test scores when the test is used as recommended. [107] It is clear from the definition that validity (as opposed to reliability) is a construct that cannot be measured directly. It can only be inferred from the context of other observations.

In practice, we must ask whether our test is really measuring what it is supposed to measure. The resulting validity is affected by a whole chain of assumptions that must be kept in mind. For example, if we use a test on standard high school subjects to select applicants to study medicine, then we should consider:

Whether the test measures the knowledge and skills the student could have acquired in high school.

Whether the ability to master subjects taught in high school predicts the ability to graduate from college.

Whether graduating from college predicts a graduate's ability to be a good doctor.

Whether the test result is affected by any secondary factors (e.g. availability of preparatory materials).

It is clear that making a precise statement of validity runs into some fundamental problems. For example, it is difficult to describe what a good doctor is. Abroad, this is sometimes circumvented by examining the degree of academic and professional success of graduates. But this is a simplification, because even a completely unambitious graduate who leaves to work as a district doctor in the borderlands can be a good doctor. When estimating the validity of entrance exams, we are therefore often satisfied with the success rate expressed as the ability to successfully graduate school in the allotted time span. The compromises continue, however, since in practice it is often not possible to wait for the passing of an entire period of regular study to verify validity and we must satisfy ourselves with academic success after, for example, the first year of study. This adds another link to our chain of assumptions, where we assume that successful completion of the first years of study predicts success over the entire course of study to an acceptable extent. In reality, such an assumption may have only limited soundness, because, for example, the first years of study at medical schools are devoted to theoretical fields and the upper years to clinical study.

6.2.1 Test validation

Since the validity of a test cannot be measured directly, in practice we focus on validating the test by collecting evidence that the test is sound. Test validation involves the collection of empirical data and logical arguments that demonstrate that the conclusions are indeed appropriate. The evidence we seek to demonstrate validity can be of a varied nature. Individual types of evidence are not interchangeable, but rather they intertwine and complement each other.

6.2.1.1 Content Validation

Content validation deals with the relationship between the content of the test and the target competencies that the test taker is meant to have achieved. During test preparation (especially when planning and reviewing the test), several experienced educators are address the question of whether and to what extent the questions included in the test cover the knowledge and skills tested for, and conversely, whether all the questions fall within the area being tested and are not testing something else. It is also examined whether the representation of items devoted to individual topics is proportionally balanced. Assessment of content validity is in a way a check whether the test plan (i.e. its blueprint) has been followed.

It always depends on the purpose of the test. For example, if the purpose of the test is to evaluate the educational program, it may also include topics that have not yet been covered, and the test then actually determines how students are able cope with new topics. On the other hand, if the test is intended to assess whether the test taker can advance to the next year, the content of the test must be strictly based on the content of the material already taught. [107]

During content validation, it is also necessary to monitor whether the interpretation of the achieved test scores does not favor any of the tested subgroups.

6.2.1.2 Criterion Validation

The content validation mentioned above is used to verify whether the test being prepared corresponds to the objectives of the tested field. However, it does not demonstrate how such a test corresponds to the objective criteria (e.g. academic success) with which we would like to compare our test. This is where criterion validation comes in, which examines the relationship between the test result and an objective independent criterion or criteria (grades, progression to further study, successful completion of school, ...).

In general, we distinguish two types of studies that examine the connection between the test and the criterion: concurrent and predictive studies.

When investigating concurrent validity, we compare a validated test and a criterion at the same time and compare whether they are really alternative ways of measuring the same construct [107]. In principle, another, already validated test, can be a concurrent criterion. We then find out to what extent the results of the new test being examined agree with this verified test. The degree of agreement can be expressed, for example, using the correlation coefficient.

Predictive validity describes the extent to which our test predicts future values of some criterion. Predictive validity is a key parameter of all admissions tests. The purpose of entrance exams is to select students with the best dispositions for future studies. It is therefore appropriate to determine whether the tests used really predict success in studies. In practice, this means that the correlation of entrance exam results with study success is determined, or that a regression model is estimated from the data, which can be used to predict study success.

Additionally, we may be interested in whether the given test brings new information beyond that which we obtain in other ways, i.e. what is its incremental or growth validity. In the case of the aforementioned entrance tests, we may be interested, for example, in whether the entrance tests add new information regarding the applicant's future studies beyond that provided by his or her high school grades. E.g. study [112], based on data from students admitted to the 1st Faculty of Economics of the Charles University, showed that secondary school performance explains roughly 15% of the variability in academic success. The result from the entrance exam increases the percentage of explained variability of success to 22%, the addition of information on successfully completed profile subjects in high school to 25% and information on the year of graduation even raises the percentage to 30%. All the mentioned effects were significant (i.e., statistically proven) in the model, thus proving their growth validity. [1].

Those interested in test validation can find more detailed information in a number of sources. [113], [114], [115].

6.2.1.3 Construct Validation

The construct validity of a test refers to whether the test measures the desired psychological construct. It is one of the most important proofs of validity. The test attempts to assess the skills of the student that cannot be measured directly in any way – they are latent. We are therefore trying to create an abstract conceptual construct (model) that helps us understand and describe this latent ability.

As an example, let's imagine a math test. A latent ability may be the ability to solve a certain type of mathematical word problem. If the test is supposed to assess this latent ability, but the test questions are written in long, rambling sentences, it may be that we are actually measuring the ability to navigate a long and complicated text – a completely different concept. Performance is then influenced by a factor that has no connection with the measured construct, so from the test perspective it is therefore a construct-irrelevant variance.

Demonstrating construct validity requires gathering multiple sources of evidence. Evidence is needed that the test measures what it is supposed to measure (in this case, knowledge of basic mathematics) and also evidence that the test does not measure what it is not supposed to measure (reading skills). This is referred to as convergent and discriminant validity evidence.

Convergent validity evidence consists of providing evidence that two tests that are supposed to measure closely related skills or types of knowledge are highly correlated. This means that two different tests end up scoring students similarly. Discriminant validity evidence, by the same logic, consists of providing evidence that two tests that do not measure closely related skills or types of knowledge are not highly correlated (i.e., will produce different student rankings).

Both convergent and discriminant validity provide important evidence for construct validity. As stated earlier, a basic mathematics test should primarily measure math-related constructs and not reading-related constructs. In order to determine the construct validity of a particular mathematics test, it would be necessary to show that the correlations of the results of that test with the results of other mathematics tests are higher than the correlations with reading tests.

6.2.1.4 Generalization of Proof of Validity

For the practical use of the test-criterion relationship in new settings (e.g., the same course in the next academic year), evidence needs to be provided that validity checks obtained in previous settings can be used to predict the degree of validity in a new but similar setting. This step, which is the opposite of the situational specificity hypothesis, is called generalizability of validity and is usually verified through meta-analyses. We try to assess whether the parameters of previous studies assessing criterion validity are reasonably comparable. The results generally support arguments for generalization of validity, suggesting that it is not necessary to conduct a new proof of validity in each new case unless the conditions and parameters of the study are significantly different. [116]

6.2.1.5 Summary of Validity Evidence

Overall validation integrates individual evidence of the validity of the intended interpretation of test scores, including the inclusion of the technical quality, fairness, and score reliability of the test.

6.3 Descriptive Statistics and Graphs

The first step in test analysis is usually the collection of descriptive statistics and their graphical presentation. After the test, you will certainly be interested in how it turned out. Descriptive statistics provide a numerical description of the test and clearly summarize its results. They provide information on the total number of tests, how many were the maximum and minimum achievable points, what the best and worst result achieved was, or what the average number of points was.

From the point of view of formal classification, we find descriptive characteristics of position (e.g. mean, median and mode) and characteristics of variability (e.g. standard deviation, minimum and maximum values of variables, characteristics of kurtosis and skewness).

You will also see similar summary descriptive statistics in the outputs provided by commercial test analysis software tools (Iteman and others). Table of descriptive statistics showing the results of a particular test:

Table 6.3.1 Table of descriptive statistics

number of test participants

1354

number of women

964

number of men

390

minimum possible number of points

0

maximum possible number of points

70

achieved minimum

25

achieved maximum

70

average

28.6

median

37.5

standard deviation

12.4

Fig. no. 6.3.1 Histogram of point gains on real test

Fig. no. 6.3.2 Histogram of point gains on real test with a small number of test participants and inappropriately detailed classification

.

Because the interpreting of some test characteristics from numerical data may not be completely intuitive, an illustrative graphic representation is used. For example, a histogram is preferentially used to show the distribution of students' total points on the test (raw scores).

A histogram is a graphical representation of the distribution of data using a bar graph, in which the height of the columns expresses the frequency of the observed variable in a given range of values, and the width of the column reflects the range of this interval. Ideally, for a large set of values and a softening classification of intervals, the distribution should approximate a normal distribution. In practice, however, the distribution tends to be more complex and reflects specific test conditions.

For example, the asymmetric histogram in Figure XX indicates that the test was quite difficult for the given group of test takers, as most of the observations are concentrated in columns with a low number of points. In this particular case, however, it was intentional and desirable, because only a small group of the best candidates needed to be selected, and the test was set for them.

The second histogram shows how its informative value decreases when a small number of people are tested (simultaneously, the division of the intervals along the horizontal axis was too detailed, whereby fewer cases fell into one column and the graph is thus burdened with a larger random error). The information is noisy and we can only speculate whether the two-peaked distribution is real and indicates that there was a subgroup with unusually good results among those tested. If we take this phenomenon as serious, it would be possible to further investigate the cause of this phenomenon using forensic test analysis methods (see the chapter on security).

6.4 Item Analysis

After completing a live test run, the first thing we'll probably want to do is evaluate the test to see how the students did on it. However, students' answers contain more than just information regarding their knowledge and skills, as the test results also reflect the characteristics of the test questions. Whereas the evaluation of a test can yield information on how individual test participants performed, item analysis can give us the (psychometric) properties. Item analysis is also important for assignment authors and reviewers, as it provides them with objective feedback on how the items they create or review perform in practice. While reviewers are good at assessing, for example, content validity, their estimates of item difficulty are often very subjective. That is why we are interested in item analysis as a source of objective reflection of our items, a tool for their continuous improvement and for educating authors and item reviewers.[117]

The basic assumption of item analysis is that the analyzed test is consistent, i.e. that it was written by qualified teachers, and that it therefore consists of items measuring one area of knowledge or ability. The quality of each item is assessed by comparing students' responses to the item with their overall test score.

The main item characteristics are their difficulty and sensitivity.

6.4.1 Item Difficulty

One of the basic characteristics of a test item is whether at least some test participants can answer it correctly — whether it’s not too difficult for the test takers.

We can estimate the difficulty of the item based on the proportion of test participants who were able to answer it correctly. This proportion is called the difficulty index and is denoted by: P

VZOREC

Where [] is the number of examinees that answered the given item correctly and [] is the total number of examinees.

The difficulty index takes on values between 0 and 100% (respectively 0 and 1). The more students that answered the item correctly, the closer the value of the index is to 100% (or 1). It's a bit confusing since we're talking about difficulty and this index is highest when the item is the easiest.

Therefore, an additional quantity, the difficulty value, is introduced. The difficulty value indicates the proportion of test takers who answered the given item incorrectly, so it is a complement to the difficulty index:

```
VZOREC
```

For more complex scoring items, indexes are calculated using the arithmetic average of the point evaluations of all test takers on a given item and the highest attainable number of points for it.

In summative testing, items whose difficulty value is neither too high nor too low (typically 20-80%) bring the greatest benefit, the best discrimination. This is logical because items that are too difficult will not differentiate between weaker and better test takers, as no one will solve a task that is too difficult. Similarly, at the opposite end of the difficulty scale, an item that is too easy will yield almost no information, because even very weak test takers will solve a task that is too easy. In the case of items with borderline difficulty values, their discriminative ability naturally decreases.

Note that this estimate of item difficulty (introduced within classical test theory, CTT) is dependent on the test takers. The value will be different for each group, and if the groups differ significantly from each other, the difficulty of the same item can be completely different for each group. Overcoming this connection between difficulty and test subjects is made possible by item response theory, in which the ability of the test takers is one of the parameters.

6.4.2 Item Sensitivity

The sensitivity of the item, or its discrimination, describes its ability to distinguish between differently performing students. Let's imagine that we divide a group of students into better and worse, e.g. according to their overall result on a test. The difference between the average success rate of both groups when solving a specific item expresses the ability of this item to distinguish between better and worse students and is referred to as the upper-lower index (ULI).

We calculate the ULI as the difference in success between a group of better (U - upper) and worse (L - lower) students when solving a specific item.

VZOREC

Fig. 6.4.1 IULI Index — difference in the probability of correct answering of an item between better and worse students.

For tests that are supposed to distinguish between the best and the second best, e.g. in admissions tests with a large excess of applicants, we may be interested in how the item differentiates just around the dividing score between accepted and not accepted. In such a case, the ULI index can be used, focusing on the divide between certain percentiles between which the dividing score falls.

Fig. 6.4.2 ULI54 Inex - the difference in the probability of the correct answer to the item between one fifth of the best and one fifth of the other students.

The ULI index can theoretically take on values between −1 and 1, but negative values are indicative of a very gross error in the item (or error in the key) and are rare in practice. A ULI equal to one means that all the better students master the item, while all the worse ones do not. Byčkovský [113] states that:

for items with a difficulty between 0.2 and 0.3, or a difficulty between 0.7 and 0.8, the ULI sensitivity should be at least 0.15,

for items with a difficulty between 0.3 and 0.7, the ULI resolution should be at least 0.25. If the ULI value is lower, the item should be considered suspect.

In practice, items hovering around the stated limits are considered not ideal, but tolerable. However, if the ULI value is too low (ULI < 0.1), the item should be checked to see if it is well constructed and does not contain any serious errors. If we are working with a finer division of the skills interval (as in the case of ULI54), an index value of around 0.1 can be perfectly fine. However, once the value of an arbitrary ULI is close to zero, or even negative, it means that the item is not working. A negative ULI value means that worse students did better than better students. So, there may be something in the item that leads the better students down the wrong track. For example, they may be looking for a catch in the question. A negative ULI can also indicate an error in the key by which the item is scored. Such an item must either be corrected or removed from testing. An interesting problem is the methodology of dividing the interval of skills into smaller parts. It may happen that the interval cannot be "automatically" divided completely ideally, e.g. because there is a large group with the same results on the boundary between the groups. In practice, it turns out that the method of division on the edge of the interval is of little importance in getting an idea of the item's sensitivity. Even if you split the disputed group arbitrarily on the borderline of the intervals, the resulting ULI usually gives a very good idea of the item's behavior.

Some papers use a different division of the skill interval. For example, an examiner divides students into three groups based on their test scores. The division of test takers into “upper third” and “lower third” is often used, but studies have shown that when students are divided into groups that have 27% of students in the “upper” and “lower” groups, the discrimination value increases. [118] It is evident that the 46% percent of students with an average test score do not show up in the discrimination index calculation. This practice is followed, for example, by the Rogō testing system, which calculates the ULI based on the bottom and top 27.5% of students.

6.4.3 Visualization of Item Analysis Results

6.4.4 Examples of Items and Graphical Representation of Their Properties

Let's look at some examples of the behavior of the items used in acceptance tests and their graphical representation.

Example one:

A person hears sound in the range from

A) 16 to 20,000 Hz

b) up to 100,000 Hz

c) less than 16 Hz

d) more than 20,000 Hz

When reviewing this item, we could discuss a number of errors that the item exhibits. For example, the proposed distractors c) and d) do not have the “range” nature referred to in the question. However, let's see how the use of this item turned out in a real test.

Fig. 6.4.3 Visualization of item behavior. The item “A person hears sound in the range of ...”, is so easy that it practically does not distinguish between differently skilled students.

Students were divided into fifths according to their overall result on the test. Correct answer probabilities were calculated for these fifths. We see that even the weakest students achieved over 90% success on this item. Students in the better groups approached 100% success. The item is so easy that it practically does not differentiate between better and worse students.

Fig. 6.4.4 Visualization of item behavior. The item “The energy of a photon is...” is rather difficult. It is very difficult for the weakest students and does not distinguish between them, but it distinguishes very well between excellent and best students. We also see a significant difference between the weakest and best students. The item can be very useful in the test.

Example two:

The energy of a photon is

a) inversely proportional to frequency.

b) directly proportional to the wavelength.

C) directly proportional to frequency.

d) independent of wavelength.

The methodology is the same as in the previous example, again we divided the students into five equal groups according to their overall performance on the test. Note that the last fifth covers the range of 40 points on a 100-point test. It can be seen here that the test as a whole was quite difficult. This particular item behaves similarly. Its maximum resolution is between the fourth and fifth fifths. Items that differentiate on the “difficult” end of the spectrum tend to be quite valuable and not easy to write. This behavior was a surprise for this item, as the reviewers predicted it would be easy.

6.4.5 Analysis of Distractors

An analysis of how the offered options contribute to the quality of the multiple-choice question – i.e. the correct answer (key) and above all the incorrect options (distractors) is referred to as distractor analysis. We are trying to find out whether the distractors are sufficiently attractive for the students and what proportion of the total number of students chose the distractors.

Let's look at the visualization of distractor analysis on a concrete example. On a 70-problem test, students were asked how methanol is formed:

Fig. 6.4.5 Distractor analysis.

What reaction can form methanol?

a) Oxidation of carbon monoxide.

b) Oxidation of methanol.

C) Reduction of formaldehyde.

d) Oxidation of methyl aldehyde.

The authors of the test marked option C as the correct answer.

The students were divided into five groups according to how many points they scored on the entire test. The gray bars indicate what proportion of correct responses corresponds to each of these five groups. So, we see that the weakest group of students—the leftmost gray column—answered correctly much less often than the best group according to the total score achieved (the rightmost column). The height difference of the last and first gray column ULI51 = 0.7 shows that the item discriminates well between the best and worst students, although whether or not it is truly well-constructed is debatable. Even the height difference of the fifth and fourth column ULI54 = 0.14 is satisfactory and indicates a good discrimination between the best and second best students. So, the item as a whole works very well.

Now let's look at the functioning of the offered options. Their behavior is described by colored dashed lines, which for each group of (similarly successful) students show how likely it is that these students would choose the offered answer. The red line (distractor A) is practically an unacceptable choice for all groups of students. Only in the weakest group, about 12% of students, choose this option, but otherwise practically no one else. The blue-green line of the correct answer (key C) rises continuously throughout the skill interval. This indicates that this response is properly constructed. In the weakest group, students choose answer C with the same probability as the other two distractors, so—except for the unattractive distractor A—students in the weakest group are actually guessing. This is again a sign of a well-differentiated item. While distractor D (dark blue line) decreases monotonically over the entire ability interval, indicating that it is working properly, distractor B (yellow line) first increases a little as students' skill increases and only then begins to decrease. Of the best students, practically no one chooses it. However, the fact that the decline is not monotonic means that students in the second weakest group are thinking about it in a way that the author did not anticipate. In this item, the authors used three different names for the same substance – formaldehyde, methanol and methyl aldehyde. The first two are fairly common. In the second weakest group, there were probably many students who, although they knew that methanol can be created by a simple reaction from formaldehyde or methanol, they only guessed whether the reaction was oxidation or reduction.

Let us now consider the distractor analysis in the case of a nonfunctional item. Students were asked about rare gases on a 100-item test:

Fig. 6.4.6 Distractor analysis.

Noble gases

A) They are rarely present in nature and form almost no compounds

B) At least one is used in medicine

c) They are inert, but otherwise normal gases with a diatomic molecule such as hydrogen, for example

d) They are always heavier than air

If we look at the height of the gray bars, we can see that the best students perform the worst on this item. The correct answer should be the simultaneous choice of answers A) and B). While the probability that a student will choose option B) increases with their skill, this is not the case for option A). Students in the two worst groups choose this answer, but after that the probability of choosing it drops steeply. Answer A) contains a fundamental problem that completely devalues this item. If we examine it, we see that it contains several errors. It is not one, but a combination of two statements: “Noble gases are rarely represented in nature.” and “Noble gases form almost no compounds.” The relativizing terms “few” and “almost none” are problematic, as they make the decision whether the option is correct depend on a purely subjective point of view. An even bigger problem is the definition of "in nature", since the author probably meant the biosphere, while for the gifted students, “nature” is probably imagined as the universe. And in this view, this answer is not true. The remaining two distractors (c, d) work correctly, but this can no longer save the item. If the author happens to write such an item, it should not pass review. The analysis of distractors is then the last chance to correct the author's and reviewers' omissions, due to its objective perspective, and remove the item from the test before it is scored.

In order to interpret distractor analyses well, one needs to have test data over a sufficiently large set of students. While the success rate of the individual groups on the item (gray columns) is relatively stable, because it reflects the data of all the students in the group, the individual distractors are no longer chosen by the entire group and are therefore significantly more sensitive to being influenced by random “noise”. If the representation of the behavior of the distractors is to retain a reasonable explanatory power, there must be more than a few hundred people in the entire test group (when divided into five subgroups). If the numbers are smaller, division into a smaller number of subgroups can be used, in extreme cases only two (two gray columns). This means we will lose the nuance of the detailed view, but the result will be less affected by random phenomena.

A distractor is considered functional (plausible) if it is chosen by at least 5% of the tested group. Designing sufficiently attractive distractors can be quite difficult, partly because the teacher may no longer be able to imagine what is difficult for students and what is not. When creating new items, the teacher can use previous, preferably formative testing to design distractors, in which students are presented with a similar item as a short-form response item. They then create distractors for the multiple-choice question based on incorrect answers.

6.4.6 Graphic Preview of the Overall Test Results

Two-color graph

For a quick orientation in how well the test was compiled, we can advantageously use a two-color graph in the item analysis. In the literature, it is also called a difficulty-discrimination plot, or “DD-plot” for short. On the horizontal axis, the items are ordered by difficulty, from easiest to hardest. For each item, the red bar shows its difficulty and the blue bar shows its sensitivity. On this graph, at first glance, we can discern “oddly” behaving items, whose sensitivity is small or even negative, and we can deal with their more detailed analysis to determine the causes of anomalies.

Fig. 6.4.7 A two-color graph (difficulty-discrimination plot, DD-plot) shows the test items sorted by difficulty (height of the red bar). For each item, its discrimination is plotted (blue bars). The horizontal dashed line shows the limit (20%) below which discrimination of a functioning item should not fall. Item #12 is very easy and its discriminative power is very low. Item No. 20 is very difficult and its discriminating power is very small and, moreover, negative, i.e. better students answer worse than weaker ones. This item probably contains some other problem that the author was not aware of. This item must be excluded from the test.

We get a different, perhaps even more illustrative form of the graph if we plot the discrimination of items (on the vertical axis) depending on their difficulty (on the horizontal axis).

Fig. 6.4.8 For the same 20-item test, item discrimination (on the vertical axis) is plotted against item difficulty (on the horizontal axis). The test is rather difficult and the discrimination of the items in it is rather below average. Both suspect items (#12 and #20) stand out clearly in this representation.

6.4.7 Rit and Rir Indices

To assess the sensitivity of the item, it is also possible to use the correlation coefficient between the point gain for the item and the point gain for the entire test, which is called Rit (correlation item-test), or the correlation coefficient between the item and the rest of the test, Rir (correlation item-rest).

The Rit coefficient is calculated as the biserial point correlation coefficient between the item score and the total test score. It tells us to what extent a given item contributes to the selection of correctly responding students from all test takers. In other words, it reflects the distinctiveness of the item and shows the item’s performance against the test as a whole. Positive values close to 1 mean that students successful in solving the given item were also successful on the test overall. Negative values show that students who correctly solved a given test item achieved a rather low overall score on the rest of the test. Correlation indicates whether an item measures the same construct as the rest of the test. If the test is focused on several topics, this should be taken into account when interpreting this coefficient. The Rir value is similar to Rit, but more accurate because the contribution to the correlation from the item itself is not taken into account. Rir is always slightly lower than Rit.

Recommendations exist for numerical values of the Rit correlation, similar to those for the ULI index:

Avoid questions with a Rit value below 0.20.

Always look at Rit combined with P difficulty.

Although discrimination assessment using ULI is more common, CERMAT, for example, uses Rir [119] when analyzing items on tests of high importance.

6.5 Classical Test Theory

The most used theory in psychometrics is Classical test theory (CTT). It is the oldest and probably the easiest to understand. It is based on the central assumption that the observed score (O) is a combination of the so-called true score (T) and the error score (e):

$O=T+e$

Of course, this is not the only assumption that is necessary for the use of CTT. Another important assumption is the local independence of individual observations, i.e. that all measurements are independent of each other. The actual score is the hypothetical score that the student would receive based on their competency alone. But because every test has measurement error, the observed score is not necessarily the same as the true score.

The true score model and its assumptions led to the investigation of the statistical properties of test items that could improve test reliability. Three important item characteristics were identified:

Item difficulty - the proportion of correctly matched test subjects,

Item sensitivity - the difference in item difficulty for (in the test) good and bad students,

Analysis of distractors - analysis of the proportion of wrongly chosen answers for choice items.

It turned out that the most reliable tests were composed of items that had a difficulty around 0.5, a sensitivity greater than 0.3, and with distractors chosen so that a reasonable percentage of students chose them. [120]

Classical Test Theory (CTT) has a number of advantages. It is simple, understandable and widely accepted. For a long time, CTT was the main tool of test research and is still widely used today, especially in item analysis. It enables the analysis of tests and items even with a small number of test takers, which is its main advantage over item response theory (IRT).

On the contrary, the main disadvantage of CTT is the dependence of the characteristics of the item on the tested group and, conversely, the measured ability of the tested students depends on the specific test. A test taker may appear well prepared if the test is easy, and vice versa. But how do we distinguish between the influence of the difficulty of a particular test and the student's readiness? The very definition of item difficulty shows the connection between this characteristic and the investigated group. Whether the item is difficult or easy depends on the skills of the participants being examined and, at the same time, on the result of the measurement of the skills of the test participants depends on whether the items are difficult or easy. [121]

Notice that it would not be possible to create high-quality item banks within CTT itself, because the item parameters depend on the specific test group. In practice, it would also not be possible to create parallel forms of tests. [121]

6.6 Item Response Theory

Classical test theory gives good results if the test takers have a comparable level of knowledge and skill. Let's imagine that this is not the case in some particular instance. For example, that the group of test subjects consists of those who have already completed a driving course and those who are just starting it. If you ask them the same question about the right of way of vehicles at an intersection, the question may be easy for some and difficult for others. We see therefore, that the concept of item difficulty based on classical test theory is not enough in this case. The solution would be to divide the group and measure the difficulty of the item on homogenous subgroups. Thus, we would get two different difficulty values, corresponding to two levels of knowledge.

If we were to divide the group in more detail, e.g. according to the length of training, we could in the end obtain (almost) continuous information about the difficulty of the examined item. This continuous curve describes the behavior of the item for different levels of students' knowledge and skills and is called the item characteristic function (ICF).

Fig. 6.6.1 Probability of the correct answer depending on the student's knowledge level, (IRT derivation)

We will therefore look for weaker students in the left part of the curve (light circles in our graph) and better students in the right part (dark circles). The whole of Item Response Theory (IRT) is based on this concept.

6.6.1 Properties of IRT models

Let's assume that we know the probabilities of obtaining correct answers from different levels of students from the tests that have taken place. If we have enough such measurements, we could try to distribute them along a curve and estimate the probability of success for other possible test takers. The interpolated characteristic function of the item usually has a typical S-shape (sigmate shape), which can be mathematically described as a logistic function. The S shape is also common to other characteristic functions (outside the field of psychometrics), e.g. the blackening function of photographic emulsion depending on illumination, etc. The s-ness of the characteristic curve expresses the fact that the transfer between stimulus and response is effective only in a limited range of stimuli. Let's imagine that we present a group of individuals of different ages with a shape recognition test. It will probably be too difficult for preschoolers and too easy for high school graduates. The flatness of the characteristic curve for the marginal values of the skill level means that in these groups the test will not well distinguish between the more and less skilled.

The characteristic function describing the behavior of an item is the basis of a number of mathematical models that try to describe how test takers respond to items. That is why this approach is called item response theory (IRT).

IRT, also referred to as latent trait theory, is a psychometric theory that was developed to better understand how individuals respond to individual items on psychological and educational tests. The term latent trait is used in IRT because characteristics of individuals cannot be directly observed; must be derived using certain assumptions about the reaction process that help estimate these parameters. The parameter θ on the horizontal axis of the IRT graph represents the level of an individual's latent trait, which may be a human skill or trait measured on a test. This can be cognitive ability, physical ability, skill, knowledge, attitude, etc.

Item response theory surpasses classical test theory (CTT) in a number of respects. It provides a more efficient description of how the items actually work, eliminates the problem of the dependence of the properties of the items on the sample of students tested, allows creation of tests with comparable properties and balancing of different versions (runs) of the test, allows estimating the effect of guessing the answer and allows the use of detailed knowledge of the properties of the items for adaptive testing.

The simplest IRT model counts on one variable – difficulty. Items of different difficulty are represented by characteristic curves of the same shape, only shifted to the left (for easier items) or to the right (for harder items). [122]

Fig. 6.6.2 Characteristic curves of items of varying difficulty in the one-parameter IRT model

A one-parameter IRT model is sometimes also referred to as a Rasch model. It is a bit of a simplification because, although the two models are very similar in appearance, they are based on different assumptions and approaches. IRT is more descriptive in nature as it aims to adapt the model to the data. In comparison, the Rasch model emphasizes the fit of the data into the model. What does that mean? One of the assumptions of the Rasch model is the “unidimensionality” of the test, i.e. that the test measures only one basic construct. If an item measures another construct, it must be excluded from the test. Part of working with the Rasch model is therefore the identification of redundant dimensions of the test and the elimination of the items that cause them to arise. Another assumption is the independence of items. That is, the probability of a correct answer to one item should be independent of the answer to other items. The assumption of independence is not fulfilled if the items have a high positive correlation. To maintain item independence, one of the interdependent items should always be omitted from the test. In this sense, the data is “adjusted” to fit the model. The data is further worked with in the same way as in IRT analysis. Those interested in this topic are referred to the extensive literature. [122],[123],[124],[125] In practice, it is important to always declare which model you are working with, to avoid any misunderstandings.

More complex IRT models, which in addition to difficulty also work with item sensitivity, describe reality more faithfully. An example can be a two-parameter logistic model. While difficulty is, as in the one-parameter model, represented by the position of the curve, sensitivity is represented by its slope. It makes good sense that the steeper the characteristic curve, the more sharply the test will differentiate between similarly gifted individuals. Sensitivity is certainly a desirable property of an item, but it is easy to see that a very sensitive item will only work within a limited range of ability levels of the test takers.

Fig. 6.6.3 Schematic representation of parameters in the three-parameter IRT model. The difficulty of the item is related to the position of the characteristic curve (or the distance from the vertical axis), the sensitivity is related to the slope of the characteristic curve at a given point (imagine that for the ULI coefficient we refine the division on the horizontal axis above all limits, then as a measure of sensitivity we also get the slope in given point.) The third parameter is the guessability of the item, which is represented in the graph by a dashed horizontal line (asymptote). (If we have a six-choice item, then even a completely ignorant student has a 0.17 chance of guessing the correct answer.)

Fig. 6.6.4 The connection between the item's information function and the item's characteristic function

6.6.2 Information Function of the Item

If we look at the typical characteristic function of an item of a standard acetabular shape, then we see that the item discriminates well only around a certain area of its inflection point, where the slope of the characteristic function ensures that a shift on the latent trait axis (level of knowledge) is reflected in a change in the probability of correct answers. As the distance increases, the characteristic curve of the item becomes flatter and the item for these values of the testee's ability ceases to distinguish between better and worse test takers.

The amount of information we can extract from the use of an item is highest around the inflection point of the characteristic function and then decreases rapidly. The function describing the information contribution of the item is bell-shaped, it is called the information function of the item, and we can obtain it by deriving the characteristic function of the item.

6.6.3 Information Function of the Test

The information function of the entire test is obtained as the sum of the information functions of the individual items (we assume that the answers to the items are independent of each other for a specific value of the latent ability).

The shape of the item's information function depends on the shape of the item's characteristic function. Highly discriminating items (with a steep characteristic curve) have a tall and narrow information curve. Such an item has high informational value, but only within a narrow range of difficulty. Items with a flatter characteristic curve, and thus a lower value of the information function, may have lower discriminative power for a given level of the latent parameter, but in turn may benefit over a wider range of difficulties. If we know the information functions of the items, we can monitor the coverage of the latent ability interval by the information functions of the test when planning the test, so that there is no excess redundancy of similarly functioning items and, on the other hand, that the entire ability interval of interest is covered.

Fig. 6.6.5 Diagram illustrating how the information function of the test consists of the information functions of individual items. The dashed curves represent the information functions of the items. The solid line above them shows the information function of the entire test.

6.6.4 IRT Model Calculation Software

While estimates of difficulty and sensitivity within classical test theory are computationally relatively simple, in the case of IRT, the situation is disproportionately more complex. We estimate the student’s unknown latent ability by finding the maximum of the likelihood function of how the estimated parameters describe the behavior of the items. For one thing, these optimization procedures require a sophisticated software tool, and further, a large number of test subjects is required to make the estimation sufficiently robust. At least hundreds, but better thousands. The more accurate (more multi-parametric) the model, the greater the requirement for the number of tested individuals.

Considering that the mathematical models of IRT can be somewhat confusing for a mere mortal, for further study, it is possible to choose literature that stays within acceptable limits of difficulty. We can recommend, for example, the review of literature given by Hynek Cígler in the journal Testfórum. [126]

A number of programs are available to calculate a model within item response theory. You can choose to rent a commercial program, such as Stata (version 14 and higher), IRTPRO, or Xcalibre. Or, on the contrary, look for the relevant libraries in the open source R environment. Commercial software is usually more user-friendly but more expensive, while the R environment is free, but it assumes that you will learn the basics of the environment and use the program codes from the libraries, that you are already familiar with their use, or will be creating your own code, which can be quite time-consuming.

On the boundary between these two worlds is the free web application ShinyItemAnalysis by Patricia Martínková and her collaborators, which we use and can recommend. It is built on the R environment, but its interface is “clickable”, making it easy to use. [127], [128]

6.7 Adaptive Testing – Use of IRT in Practice

During a normal test, the examinee receives a number of items, some of which may not be completely relevant to him or her. They can be more difficult or easier than the examinee’s level. The information functions of the test items cover the range of difficulty levels within which the skills of most test takers range. An unwanted side effect is that each test participant answers a series of questions that are too easy for them, or too difficult for them. At the same time, both are demotivating and, from the point of view of the testing institution, a waste of time. In electronic testing, one can therefore imagine an algorithm that will select items for the test taker, the difficulty of which will be adapted to their performance in solving the previous items.

This approach is called “computer adaptive testing” (CAT). It allows measuring the student's latent skill with the same accuracy as a classic test, but using a smaller number of items.

Thus, adaptive testing adapts the test to the test taker, item by item, based on their responses. A correct answer leads to a more difficult item, while an incorrect answer leads to an easier item. The difficulty of the items is continuously adjusted to the skills of the test taker. A gifted student will receive increasingly difficult items, while an average student will receive easier items. The number of items used is related to the required measurement accuracy. This means that the test stops when the predetermined required accuracy of the psychometric criteria is reached. With adaptive testing, the test is only as long as it really needs to be.

Fig. 6.7.1 Principle of adaptive testing. If we know the psychometric properties of the items, we can arrive at the same accuracy of the “knowledge” parameter estimate using a smaller number of items. Let's imagine that we first give the subject an item of average difficulty (A). The test-taker answers correctly and the adaptive testing algorithm chooses a more difficult item (B), then the test-taker does not answer correctly and the algorithm offers him an easier item (C). Instead of the respondent having to answer all test items, in this illustrative example, answering three items is sufficient to determine the respondent's level with sufficient accuracy.

The method is based on item response theory (IRT), which was discussed in the previous chapter.

6.7.1 Advantages and Disadvantages of Computer Adaptive Testing (CAT)

CAT is a modern way of testing that uses algorithms to optimally adapt the test to each examinee. Traditionally, items are compiled into a test set and presented to students in that set. The most obvious disadvantage of this approach is its inefficiency. The difficulty of the test items does not reflect the skills of the test taker. Let's imagine an exceptionally skilled student who answers all the most difficult questions correctly. We can confidently give it a high score without wasting time answering all the simple questions. While this saving may seem small for one student, if you apply the same method to the entire tested group, the time savings are significant.

Another problem is the uneven accuracy of measurements for students with different levels of knowledge. In traditional tests, items of medium difficulty usually have the greatest representation. There is a good reason for this: the test takers are likely to include a large number of people of intermediate skills. People of average skill will be evaluated very accurately by the test. However, this will happen at the expense of low measurement accuracy for students with a low or, conversely, a high level of skill. These are evaluated with much less accuracy. For the same reason, students with above-average or below-average skills may have a bad experience with the test. Weak students may feel exhausted and discouraged by the fact that most items are too difficult, while above-average students may be demotivated by the fact that most items are too easy for them.

Advantages of CAT:

shorter tests (by up to 50%),

stable accuracy,

favorable feedback from test takers,

better motivation of test takers,

lower divulgence of tested items,

possibility of use for measuring student progress (the student’s test will be different at the end) .

Disadvantages CAT:

impossibility of returning to previously answered items during the test,

sensitivity to test anxiety,

the need for prior calibration of the items,

for items with beneficial properties, using them too often can result in these items being divulged,

requires a sufficient amount of pilot testers (several hundred),

preparation requires highly qualified professionals,

more demanding to explain to the public – higher public relations costs.

6.7.2 Requirements for Computer Adaptive Testing (CAT)

CATs have many advantages, including cutting testing time in half, but they require experienced psychometricians, large pilot samples, and specialized software. Here's a basic overview of what to consider when deciding on adaptive testing:

Items must be evaluable automatically, because the next item is selected in real time based on the result for the previous item. This excludes some otherwise useful forms of test items (constructed answer questions, essay, etc.)

Resources are needed to develop banks with a large number of items. Banks usually need at least three times as many items as the intended length of the test (although this is often no more than is needed for traditional test formats).

Extensive pilot tests must take place. IRT requires a sample size of at least 100–1,000 test takers to be used for pilot testing. The required number depends on the complexity of the IRT model used. More complex IRT models require larger samples.

It is necessary to have experts in psychometrics. For successful deployment, qualified experts are needed, especially for item calibration and IRT analysis, or for the simulation of adaptive testing with a given test set.

Analytical software must be available. IRT analysis software (e.g. freely available ShinyItemAnalysis or commercial equivalents) is required for item calibration.

An IRT-supporting item bank capable of storing IRT item parameters and supporting the design of CATs is essential.

Finally, there is a need to have an appropriate test delivery system. The latter must be capable of adaptive testing based on IRT, at least with appropriate termination criteria and item selection algorithms.

6.8 Use of IRT for Fairness Analysis

By fairness (objectivity) of a test, we mean its ability to measure the studied attribute or construct with equal validity in all subgroups of the tested population. We addressed fairness in the review of test items, mentioning it as one of the proofs of validity. But since the tools of item response theory will come in handy for its ex-post analysis (from the data of a test that has been given), let’s go back to this topic.

We call the item “differentiating” (differential item functioning - DIF) when people with the same latent ability, but from different subgroups, have different probabilities of giving correct answers. The difference in average performance between groups is not necessarily unfair in itself. Unfairness only occurs if the difference in measured performance does not correspond to the actual difference in the latent trait the test is intended to measure.

For example, imagine that we are examining the fairness of a reading comprehension test. In doing so, we find that students with visual impairments achieve worse results. Does this mean the test is unfair to these students? We don't know that yet. It is possible that the students with visual impairments actually have lower reading skills than other students.

Now suppose that we reprint the test with a significantly larger font size and find that the average score of disabled students rises to the level of non-disabled students. This suggests that the reading test in the original small print version was unfair (biased) to disabled students. The result also suggests that the test is fair when presented in the large print version. The small font size introduced a systematic error into the design of the test.

We therefore distinguish between so-called “benign” DIF, where the difference in the probability of the correct answer is related to the measured latent trait, and “unfavorable” DIF, where artifacts in the measurement process, unequal preparation options, language-dependent interpretation of the text, and the like are reflected in the result. There is no definitive quantitative method that can distinguish these two cases from each other. Any time you encounter differential functioning of an item, the item should be carefully reviewed by a team of experts. [129]

Differential item function from an IRT perspective

An analysis based on item response theory can be used to examine the fairness of items in relation to the subgroups of the test population under investigation. Theoretically, problematic items should be eliminated or corrected already during item review, when content and construct validity are verified, but even a careful review may overlook things. Analyzing the responses of test takers including the behavior of items towards different subgroups of test takers can help catch problematic items and improve the quality and fairness of the test in subsequent rounds.

In practice, we check whether the difficulty of the item does not differ for selected subgroups of test takers (e.g. gymnasium school graduates vs. graduates of other secondary schools) who otherwise have the same skills (e.g. measured by the total score). For the given groups, we fit the characteristic curves according to the IRT theory with the measured points and compare these curves with each other. We then take the area between the two curves as an index describing the different functioning of the item for both groups.

In the United States, a verbal analogies question appeared on the SAT: Find a similar relationship:

Runner: Marathon

(a) envoy: embassy

(b) martyr: massacre

(C) rower: regatta

(d) referee: tournament

(e) horse: stable

It is easy to find the correct answer (“rower” and “regatta”) if you are from an environment where the terms “marathon” and “regatta” are used. An analysis of the tests showed that African-American students were noticeably worse at answering this question (22% correct) than their white counterparts (53% correct), although this was not the case for other questions. The question assumed “self-evident” knowledge of the sport widespread among only one of the subpopulations. [130]

Fig. X.XX Illustration of the simplest case of unfair behavior of an item. Characteristic IRT curves for two groups solving the same but differently functioning item (see example above). The size of the area between the curves corresponds to the size of the DIF coefficient. Both characteristic curves are equally discriminating, but show different difficulty for the observed groups. A case where an unfair item gives an advantage to one group of students over another across the entire ability range (as here) is referred to as “uniform DIF”.

Fig. X.XX Illustration of non-uniform differential behavior of an item. The characteristic curves calculated for both observed subgroups show not only varying difficulty of the item for both groups, but also varying discrimination. For the first group (dashed characteristic curve), the item is easier at most skill intervals, except for the highest values, where the item, on the contrary, becomes easier for the second group (solid characteristic curve). This type of differential item behavior is referred to as “non-uniform DIF”. The information functions of this item also have different shapes and values for both groups. The course of the curves is taken from the interactive training section of the ShinyItemAnalysis web application. [131]

Using IRT for fairness analysis provides detailed information that would be difficult to estimate based on fairness reviews. For example, medical school entrance exam questions were found to have differentiating items that women answered significantly better than men. They were mainly questions related to children's diseases.

It is also possible to use other statistical methods for fairness estimates, for example visualization using a graphic representation of the proportions of correct answers, or analysis of contingency tables (Mantel-Haenszel method). Those who are interested can find all the mentioned tools in the ShinyItemAnalysis application.

The issue of fairness exceeds the scope of this text. For those who are interested, we recommend checking out publications, courses and tools that deal with the topic in more depth. [132], [133], [134]

7 Testing Cycle

Like all education, testing is a cyclical process. During the preparation, execution and evaluation of each test, we create (perhaps inadvertently) the simplest basis of the test cycle.

Fig. 7.1 Diagram of the simplest intuitive test cycle.

Once we start preparing tests repeatedly and systematically, we start to project our experience from previous runs into the creation of new items and tests, and this feedback creates the first complete test cycle, at the end of which we are ready to work (better) on a new round of tests.

For tests of great importance, which must meet the standards of validity and reliability, a number of steps need to be implemented that are not explicitly emphasized in the intuitive test preparation. Typically, a test cycle for a test of high importance might look like this:

Fig. 7.2 Diagram of the test cycle for tests of great importance. Loosely based on [135]

The outline of this book follows, in a practical sense, this “big” test cycle. A corresponding chapter exists for most of the steps. Therefore, we will be relatively brief in commenting on the individual steps, as the aim is only to remind you of what is discussed in the relevant chapter.

Defining Learning Objectives

The work on the test should start with the clarification of the objectives. By defining the learning objectives, the teacher defines the scope of the subject that the student should be able to master after completing the course. The teacher specifies the key competencies to be tested.

Test Plan

Test design is another key point of the whole process. It is necessary to establish how many items the test will contain on each thematic area and what types of items will be used. This phase is particularly important if the test is prepared in several versions that are to be compared with each other. The learning objectives will be reflected in the selection of questions and the ratio of representation of individual topics in the upcoming test. Named after the blue colored copies of building plans, this test planning is called blueprinting.

Plan Review

When preparing important evaluations, the influence of the individual preferences of individual educators must be minimized. Therefore, the test plan needs to be opposed by other teachers, so that the representation of topics and the use of test formats is based on the consensus of multiple teachers.

Item Creation

Perhaps the most demanding stage of test preparation is the creation of items. Teachers will design new items in accordance with the topics determined by the blueprint and in the format prescribed by the blueprint.

Item Review

Items usually bear the "handwriting" of the author. That is why we present them to teachers who also know the target group and the subject being discussed. When the questions are answered, the items are submitted for assessment to a group of experts (for example, the test preparation methodology of the Rogō program recommends at least 5-9 people, which is, of course, mostly unrealistic in our conditions), who, according to the prepared form, go through the test items and verify the individual aspects that the new item must meet, and possibly suggest any necessary modifications. An experienced item author must then go through the individual reviews, assess the relevance of the comments, and edit the items if necessary. When supervising reviews, the supervisor can give reviewers credit for good quality reviews. This provides reviewers with feedback and at the same time generates information about their performance and usefulness.

Assembling a Test

The author of the test chooses items from the pool of created items in such a way as to fulfill the intention of the blueprint and at the same time comply with other (often unspoken) requirements. For example, to keep the test's difficulty and time demands reasonable, ensure that the number of calculation items is not higher than in other versions, and so on.

Piloting and Reviewing a Test

If the test is to be of high quality, a pilot run and a test review must also be part of its preparation. To check the behavior of the items and of the test overall, it is advisable to try a pilot run. The analysis of the results of the pilot test can show the (in)ability of the items to differentiate students according to mastery of the material, clarify their objective difficulty, etc. Pilot testing is time- and organizationally demanding, therefore often only the first run of the test is used as a pilot run. In addition to piloting, we have the quality of the test checked by a group of experts who identify and remove the last remaining errors, and ambiguous or problematic wording. Even if, after all these checks, it seems that there can be no more problems in the test, some problems are always discovered!

Setting Boundaries

An important step is to set the borderline for passing the test. If the test is linked to criteria that the participant must meet in order to pass, this step is called absolute standardization and its time comes precisely at this point in the test cycle. Setting the criteria for passing the test in advance gives the test takers confidence that the boundary will not be set on purpose so that a particular participant will still pass. For teachers, setting an objective cutoff is a means of ensuring that only sufficiently competent students pass the test. There are more options for finding “boundaries”, let's recall, for example, the gold standards – the Angoff and Ebel methods.

Test Implementation

As we have already mentioned, a written test can be given in paper or electronic form. In both cases, it is necessary to ensure the creation of “test versions”, distribution of tests to students and the collection of their answers. In addition, for tests where the results will have a significant impact, we must ensure a level playing field for participants, such as supervision during the test and more.

Processing of Responses

This step in the test cycle mainly concerns paper testing, when the answers on the collected answer forms must be optically scanned.

Preliminary Analysis of the Test

For tests of great importance, it is desirable to pre-analyze the test before evaluating the results. We can recognize gross errors such as a mistyped item key or wording errors from the suspicious behavior of items. For such items, the key is modified to make the item work before the test is evaluated, or the item is removed from the tally (all participants receive a point for it). This will prevent a situation where, after the announcement of the results, someone could complain about an incorrect question and it would be necessary to recalculate, say, the order of accepted students.

Grading Students

Student grading is the most important output of the test. During grading, it is possible to compare the number of points (total score) achieved by individual students and thus determine their relative placement. Using expert estimation (e.g. the Ebel or Angoff method) we determine the threshold for the “pass” or “fail” decision (so-called absolute standardization) and by dividing the success interval into the necessary number of parts, we can determine students’ grades in the form of grading scales – grades. The anonymization of the tests before the full evaluation (grading) of the tests contributes to ensuring equal conditions for the participants.

Setting Relative Boundaries

If the test is aimed at comparing performance among the students, then setting the thresholds for passing this test cannot be done before this moment, when the results are known and any problems revealed by a quick analysis are solved. Test takers are ranked according to their test scores, and the dividing line is set either according to some method of relative standardization, or arbitrarily in the event that the test was given to select suitable candidates for a limited number of positions.

Grading of students

In this step, the students' results are converted into a grade assessment and provided to the students.

Analysis of Test Results

After the test round, data is available that can be used to examine in more detail how the test actually performed. While in the quick analysis it was only about identifying, and eliminating, possible problems, in the test analysis we examine the characteristics of the items and the test. This enables us to provide feedback to authors and reviewers as to how well they “hit the mark” in their work. The test is a measuring tool and, like any tool, it has its own characteristics that are important to know. It is optimal to be able to estimate the quality of the test even before its live deployment, i.e. already during pilot testing. The properties of the test then need to be verified on the target group during live use. When using the test repeatedly, it is useful to compare the results between individual test runs.

Evaluation and Reporting

The results of the test analysis are reported as feedback, both to the authors and reviewers, as well as to the relevant responsible persons in the hierarchy of the institution. Since reporting is a common procedure for high-impact tests, many test analysis programs also include tools to prepare the report or parts of it. (Iteman, Xcalibre, Remark Office, ...).

8 Item Banks

The item bank (or “test question bank”) is a repository of tests and test questions and their metadata.

Test item banks are closely related to the test cycle described above, as an item bank in a broader sense can include tools for creating, reviewing and managing items, for planning, creating and delivering tests, and finally for evaluating responses, including analyzing tests and test items. So, they can be used to support the entire test cycle.

Along with tests and questions, metadata is also stored, including psychometric characteristics of items from previous test runs, which can be used as feedback for improving further test runs. If the system also maintains information about the author of the item and the reviewers, it can provide them with feedback on how the item performed in the test, thereby contributing to their education.

Items are usually in a “multiple choice” format, but any format can be used. Items from the bank are used to construct tests distributed in both classic and electronic form.

David Vale described an item bank as “an organized collection of test items”. [136] The simplest item bank may be a shoe box with cards containing test problems. When item statistics were available, they were often written on the back of the cards, along with the date the test took place. If the test writer wanted to create a new test, they manually searched the item box, checked the contents of the items and their statistics, and selected the ones to use in the test. [137]

The difference between an item bank and a simple set of test items is that item banks allow you to track an item throughout its life cycle, in which the item can be used in a series of tests. Information about the behavior of the item in one cycle is stored and used to construct better tests in subsequent cycles. This “reusability” and the improvement of quality due to the evaluation of previous uses of items are among the basic features of item banks.

At the turn of the century, “reusable education objects” were a research topic. In the case of learning materials, the effort to store them in repositories and with metadata for further use did not succeed. Description using metadata was too laborious for busy educators, and its complexity suppressed the effect of savings from reuse. In the case of item banks, the situation is different. Working with metadata is useful and mostly fully automated. The idea of using “reusable education objects” is thus coming to fruition decades later, albeit in a different context than originally planned.

Item banks are an essential part of the quality assessment process. In addition to supporting the creation of test questions, they can do much more: store metadata about items, store psychometric properties of items in tests, track their usage, handle user management, security management, and enforce good workflows that help maintain quality standards. You don't need an item bank if you are doing a small number of tests that are of more of a formative nature. But you can't do without an item bank if you're preparing large-scale and important assessments of student performance.

8.1 Items and Their Metadata

Item banks contain not only the text of each item, but also a range of information regarding its origin, incorporation, use, and psychometric characteristics. Examples of such metadata include:

item text,

creation date,

correct answer,

item format,

assignment to topics,

the author of the item,

item reviewers,

reviewers' statements (for the Angoff Method),

review status (for review, done, rejected, for revision, ...),

item status (e.g. new, reviewed, active, archived, replaced by a new version, ...), characteristics according to Classical Test Theory,

characteristics according to Item Response Theory,

item incorporation in test plans,

relationships to other items.

When building an item bank, it is necessary to consider the requirements of the specific field and already adapt item classification to it, for example. The item can be stored in the database as a whole, or in individual parts – as a stem (short text), the question itself and the offered answers. After this, it is easy to generate derived versions of the items (e.g., with different data for calculation), but it is more difficult to keep the information in order, and keep track of what form the item was used and in which test. The storage method must be chosen at the beginning, since a number of other properties of the item bank depend on it.

8.2 Types of Relationships Between Items on Tests

Various metadata is stored on items. This data includes relationships between items, describing whether and under what circumstances two items can be together on one test.

Companions

Items that must occur together because they rest on a common foundation or use common supporting material.

Close Friends

Closely related items must occur together. If one appears, the other must also be present.

Snobs

Items of the “Close Friends” type that can only be used in a certain order to make them understandable.

Dependent

Items that can only occur with the support of a “Supporter”

Supporters

Items with no separate interaction that provide context for subsequent “Dependent” items. For example, this can be a text that will be referred to in several items.

Antagonists

Items that must not appear close to each other in the test, because one provides a hint for the other.

Enemies

Items that cannot be in the same test because they ask the same thing.

Clone (offspring)

An item that was derived by changing (fixing, improving) the parent item.

8.3 Functionality of Item Banks

We typically expect an Item Bank to ensure and support:

access authentication,

creating items,

saving items,

ordering items,

item review (including review process management), creation of test plans (blueprint),

item management,

test administration,

management of learning objectives for item categorization,

print or take the test electronically,

loading of results forms,

automatic psychometric analyses,

activity logging,

records of item and test exports,

data retention of psychometric data from previous item uses.

The essential feature that distinguishes an item bank from an item repository is precisely that the item bank preserves psychometric characteristics from previous rounds of testing. It thus enables their use for the creation of better items and the construction of better tests.

As the reader must have noticed, automated test generation is not explicitly mentioned in the list of properties. It turns out that the design of the test requires the author to take into account a large number of often unexpressed item parameters and balance their representation in the test. So, even if the system can generate a draft of a set of test items, it is still assumed that this draft will go through a proofreading by the “author” of the test, who is responsible for its balance. The item bank offers the author a number of tools for this, for example, the ability to monitor how items cover the tested area (blueprinting), how many computational items are in the test, and so on.

Comprehensive item banks have other useful mechanisms built in, such as monitoring changes to dependent item parameters. After reworking the item, they can change the “review status” from the value “to be revised” to the value “to be reviewed” etc.

A note on the influence deciding on the necessity of forensic analyses of a test on the structure of the item bank:

The item bank itself does not need to store information on the identity of the tested person. If the creators of the item bank consider it necessary in the given context, they can store information about year, gender and other attributes in the item bank, but there is usually no reason to store a specific identity. The situation changes when it becomes clear that there is a need to use statistical tools to check whether there is any illegal conduct during testing. In that case, it is necessary to work with the identity of the test takers. It is a conceptual question, because then we have to deal not only with the statistics of the test items, but also with the statistics of the examinees. When designing an item bank, it is advisable to remember this circumstance and take it into account in advance.

8.4 Advantages of Item Banks

An item bank should be at the core of any serious testing system. Its use brings a number of advantages [138]:

It enables repeated creation of tests with predictable properties.

It provides an opportunity to objectively determine the specifics of individual authors of items. It may turn out that some item authors systematically prepare easier or, conversely, more difficult items. Some give preference to one thematic area, or one type of item (e.g. computing). Therefore if, for example, you need to supplement the test set with a specific item, you know who to contact.

Regularly working with authors allows you to train them and increase their skills in creating items.

The bank forces a standardized procedure for the preparation of test items (content review, language proofreading, typographic checks, determination of difficulty, post-test evaluation...), which is a guarantee of systematic quality improvement.

Order is maintained in different versions of items. When an item needs to be modified or corrected, either a new version of it or just a modified version is created depending on the extent of the change (e.g., if it's just an issue of correcting typos or modifying typography).

All items are assigned to specific topics and the system allows you to search them according to a number of criteria. This ensures better coverage of the test substance when creating tests, makes it easier to follow the test plan, avoids repetition and prevents problems with items that are of an unknown or unnecessary focus.

Item banks make it possible to assign the results of psychometric analysis of completed tests to items. It is thus possible to sort low-quality items, or to monitor whether an item was not leaked between two tests.

In the item bank, the permissions of individual users are set based on roles. At the same time, all activities are logged, especially changes to items, mass exports and access to finalized tests. All this contributes to increasing the security of the test.

The item bank should allow for easy identification of duplicate items and items that have some type of relationship (enemies, friends, close friends, ...)

An item bank increases testing efficiency and quality by viewing items as reusable objects and supporting the entire test development cycle.

8.5 Examples of Item Banks

An item bank is essentially a simple database, so it can be stored in a database system or even in a spreadsheet environment.

An example of an item bank created in Excel was presented at the Association for Educational Assessment – Europe 2016 conference. [139] The solution was relatively simple and yet completely functional, and we can definitely recommend this step to those potentially interested in item banks. Even if it is only a temporary solution – it will help you clarify what your needs are and what you require from a possible future, more comprehensive, solution.

Most large companies dealing with testing have developed their own item banks. In the Czech Republic, for example, there is no doubt that SCIO or Cermat have some form of item banks. At some universities, parts of the item bank are integrated directly into the school's information system.

Item Bank of the First Faculty of Medicine of the Charles University

The First Faculty of Medicine of the Charles University has also developed its own separate item bank. It is a comprehensive item bank in the Multiple True/False item format. Similar to the Rogō test system, it is a web application that runs on all major browsers. The bank, whose development dates back to 2014, supports all steps of the test cycle, from blueprinting to item creation, review and version management. The bank enables the creation of tests, their review, print preparation, import of results, itemized analysis and reporting of test results. There are about 10 thousand items in the bank, including metadata about the results of use in previous rounds. Due to its use in important exams, the bank is heavily secured and every access to it is logged. The properties of this item bank are the basis of the general description of item banks given above. And conversely – this bank has all the properties required in the general model.

Many people interested in testing are looking for an economically acceptable way to purchase a commercial item bank. A number of options are available, but their licensing model is usually based on an environment where significantly more resources are allocated for these purposes, which makes these solutions practically unavailable for potential domestic customers.

Experience with TAO of Testing

At the boundaries of commercial products stands the “TAO of testing” system, which caught our attention with the availability of its free version. As it might also attract other potential interested parties so we consider it useful to share our experience.

TAO is, in Chinese philosophy, an expression for the basic principle of the universe, but also a French abbreviation for Testing Assisté par Ordinateur (computer-assisted testing). The TAO platform has been developed at the University of Luxembourg as an open source project. It provides participants in the computer-based testing process with a comprehensive set of features that support the creation, management, and administration of electronic tests. In particular, it covers:

item development and management,

management of the test takers,

creating and managing tests,

management of authors and reviewers,

delivery of tests,

results management.

TAO is an open and modular system based on the assumption that no one solution can fit all, so users are expected to adapt it to their needs. It is a web application that runs on a server and does not require anything to be installed on the user's computer. It supports translations into national languages. It offers authors a WYSIWYG editor for intuitive item creation, including multimedia integration.

Although the system is open source, it is only seemingly “free”. You have two options to use it. Either you use the provider's paid cloud installation, which costs the same as other item banks on the market, or you install TAO on your servers yourself. However, the documentation is insufficient and installation and updates are poorly described. The installation scripts contain errors and the manuals contain references to non-existent scripts and other resources. The developer does offer support, but it is expensive.

The system lacks support for directory services (LDAP), which makes it impossible to use its existing identity verification (and user name and password management) at a specific institution. This is very impractical when deployed in large institutions, which would have to maintain several directories for testing purposes. The TAO system also does not include test and item analysis tools (only export to QTI 2.2. or CSV formats), so you need to use third-party software for test and item analysis.

We implemented and experimentally ran this item bank [140], but problems with documentation and updates, as well as the need to handle student usernames and passwords separately outside of the existing directories (LDAP), led to the termination of this experiment.

8.6 Extensive Banks of Test Items

In some cases, the normal work with the item bank has exceeded the usual dimensions. Let us mention two such interesting cases. The most interesting of these collaborations is the British Medical Schools Council Assessment Alliance (MSA-AA), which unites all 31 medical schools in Great Britain. [141]

Medical Schools Council Assessment Alliance

This alliance operates a common item bank for them. The common objective of all participants is to improve the evaluation of learning results at medical faculties.

The alliance follows the activities of the Association of Medical Faculties – Universities Medical Assessment Partnership (UMAP), which was founded in 2003 for the purpose of cooperation in the creation and sharing of test items. The association gradually expanded to include other schools and after 2009 was transformed into MSC-AA.

The common item bank mainly contains items in the single-best-answer (SBA) format, but items for OSCE stations and “multiple-mini-interviews” are also being added. The bank is accessible to all participating schools. Questions are created collaboratively and undergo extensive quality testing and standardization. All medical colleges in the UK have agreed to include an agreed proportion of shared questions in the final exams, thereby ensuring psychometrically valid comparability of the “state” exams.

Item Management System

In German-speaking countries, the Item Management System (IMS) item bank is widely used in higher education. It was established in 2006 as a result of the cooperation of the medical faculties of the universities in Heidelberg, Berlin and Munich. The group gradually grew to 77 institutions, mainly in Germany and Switzerland and is covered by the Umbrella Consortium for Assessment Networks (UCAN).

The Item Management System is a database of items that can cooperate with computer testing applications, paper or mobile testing, and in analyzing results and so on. As of March 2021, 700,000 items (without format resolution) were stored in the system, of which 125,000 were shared. [142],[143] The Umbrella Consortium for Assessment Networks (UCAN), which maintains the item bank, is declared as a non-profit organization, but outwardly it functions as a commercial entity. The license is calculated according to the number of test takers. Prices are set for Western European conditions.

9 Testing Security

As much as we tend to believe that everything is developing for the better, the published data seems to show that the tendency to cheat on exams tends to increase in the long term. [144], [145]

Fig. 9.1 Development of the probability of students cheating on an exam over the course of half a century. [146]

According to research, the rate of cheating on tests has increased dramatically over the past century. [147] For example, between 1963 and 1993, the rate of serious cheating on written tests increased from 39% to 64%. [148]

Tests are the gateway to many educational and professional objectives. It is therefore not surprising that the motivation to cheat is high. In an informal survey conducted in 2007 among 30,000 American college students, 60.8% admitted to cheating during their studies. [149] The same survey showed that 16.5% of them do not even feel it is an ethical problem. In other studies, up to 85% of students report cheating on tests at least once during their studies. [150] At the same time, social tolerance for cheating is increasing, especially when it is done via the Internet. [151]

A survey conducted at Fordham University made a surprising finding: it pointed to a significant difference between the academic averages of cheating students and their honest counterparts. Cheaters belong to a group with statistically significantly better academic results than those who do not cheat. It is pertinent to ask what role a purely achievement-oriented motivation plays in the formation of these attitudes.

At the same time, 91% of students considered teachers ignoring cheating as being highly unethical. [152]

Under these circumstances, it is obvious that for tests of great importance, attention should be paid to their security. In the broader context of testing security considerations, we must take into account the sum of values that could be compromised in the event of a security breach. If the cheaters were successful, candidates who the test should have excluded could pass the test. Not only would the work of the teachers who prepared the tests be ruined, but the reputation and credibility of the entire institution that runs the tests would also be threatened. One of the prerequisites for the validity of summative assessment is its credibility and objectivity. At the same time, the importance of ensuring the security of the assessment increases with the importance of the exam.

Curiously, there is less literature on this important topic than we would expect. In part, this may be because the necessary know-how was fragmented and held by test takers. It is also partly because publishing cheating detection procedures undermines their intent. However, in recent decades, texts covering this entire area are starting to appear. [153], [154]

Most evaluations at universities have a periodicity associated with the rhythm of the academic year. As a result, security care should also be repeated regularly. The institution sets the rules (e.g. admissions procedure) and the participants try to achieve the best possible result within these rules. The institution then corrects its rules to optimize selection, while maintaining impartiality and objectivity. From this point of view, testing security is not a steady state, but a cyclical process.

Fig. 9.2 Diagram of the security cycle of deterministic testing (loosely according to [155]).

9.1 Testing Security from a Risk Management Perspective

When assessing the importance (and the rationality of the costs incurred) of individual aspects of testing security, we can use a methodology familiar in the field of risk management. Within the institution and in the context of the upcoming testing, we should think about:

Identifying assets,

Asset valuation,

Identifying threats,

By estimating the probability of the occurrence of individual threats,

By estimating the vulnerability of the asset to the threat,

By estimating the total risk resulting from individual threats to assets.

Asset Identification

Assets are anything that represents value to the organization. Although it may not be obvious from a narrow view of testing, the credibility of the entire institution is among the main assets in education. There are cases where gross errors in the security of the testing process (during the admission procedure) have led to fatal consequences in the form of withdrawal of accreditation.

Asset valuation

When valuing assets, we can estimate, for example, the price of an item bank or its contents. The item bank of the First Faculty of Medicine of the Charles University contains about 10,000 items. The cost of their acquisition was about CZK 1,500 per item. From there we get a price of CZK 15 million for the entire content of the item bank. We can also estimate the cost of the established credibility of the institution. These are values built up over the long term with large costs. If, for example, there were a scandal surrounding the admissions process, then this could mean that PR costs for a period of, say, 5 years were devalued. If the loss of confidence in the regularity of the procedure leads to the withdrawal of accreditation, the faculty will lose income for teaching students, i.e. one of the biggest sources of income, for several years. For a larger faculty, the loss thus reaches hundreds of millions of CZK. However, protected assets also include, for example, the personal data of test participants, which are protected under the GDPR under the threat of draconian fines.

Threat identification

Threats are scenarios in which an organization's assets may be at risk. In testing, this mainly concerns external and internal threats. Among the threats, a special place is occupied by attempts to intentionally influence the results by illegal procedures. There are a number of types of unethical and fraudulent behavior that can compromise the test's informativeness:

Leaking items, or unauthorized acquisition of prior knowledge, can occur if participants in previous runs bring up the wording of questions. Either they memorize the content of the exam, or they copy the question with their mobile phone, or they write it down. The objective is to achieve an advantage in the test over other test takers, thanks to knowledge of specific test questions. Unauthorized access to test questions gives cheaters an unfair advantage over honest test takers.

Unauthorized cooperation. Two or more test takers may attempt to work together on completing a test. For example, copy answers, or share answers during the test via text messages and the like.

Identity confusion. The test's informativeness can be impaired if someone other than the actual candidate takes the test. This “test proxy” can be prevented by maintaining high standards for identity verification. This issue needs to be paid a lot of attention, especially in distance tests, where the options for identity verification are limited.

Unauthorized assistance. The test result can also be distorted by collusion, if the test taker receives help from the staff that organizes or evaluates the tests. Cheating means that the test proctor or test administrator provided unauthorized assistance to the examinee or tampered with the test data or test session in some way. An example of collusion could be when an invigilator allows a test taker to deviate from approved test procedures, gain access to unauthorized resources, or allow the test taker to exceed the approved test completion time. Collusion may also involve tampering with exam records, such as changing an examinee's answers from wrong to right, or adding missing answers.

Prohibited aids and resources. According to a survey among 15-year-old pupils in the Czech Republic, in 2013, the most widespread method of cheating was still the use of cheat sheets, while other, technical, means such as mobile phones trailed behind. [156]

Security threats vary for different types of testing. Paper-based exams may be more prone to copying answers than computer-based exams (especially in the case of adaptive testing), while computer-based testing may be more prone to the use of unauthorized resources. Security policies and procedures should be tailored to suit the type of test.

Every type of security threat must be prepared for:

An estimate of the probability and potential consequences of each of the possible cases

Preventive measures to reduce threats

Follow-up procedures to minimize the impact of extraordinary events.

Although risk assessment is laborious and the reasons for dealing with it may not be apparent at first glance, it serves a fundamental purpose – to help ensure the protection of important values with means that are reasonably commensurate with the protected values.

It is obvious that the prevention and elimination of security threats requires a systematic approach. Therefore, in summative testing of great importance, a test security plan is created, which specifies who should deal with what and when to achieve the necessary level of security. Let's go through such a security plan step by step.

9.2 Test Security Plan

The test security plan describes the values that need to be protected, the known risks, and the procedures to reduce the risks.

A test security plan is a comprehensive set of policies, procedures, and documents that outline and govern actions related to test security. From the development of the test plan to the recapitulation of the results in the security audit, “security” applies to almost every step. The use of test scores to assess candidate performance presupposes confidence in the integrity and objectivity of the test. Without trust, even credibility would be compromised.

What needs to be done to make test scores reliable and interpretable? At a minimum, this requires having and following a reliable test security plan.

Most of the policies and procedures in a test security plan are based on common sense. For example, it is essential to have a communication channel for clear and unambiguous sending of messages to applicants. How else can participants be expected to follow the rules if those rules are not explained, along with the corresponding consequences? Adopted procedures must make sense, fit well with the given testing program, be enforceable, and be legally defensible. The proposed procedures should align with the threats specific to your programs and be tailored to specific needs. While in one case the main problem may be the leaking of items between exam dates, in another arrangement the greatest threat may be identity fraud. Threats vary from program to program, and security plans should address protection against these threats.

Roles, Items, Responsibilities

Preparing for important tests is a collective effort. Even from the point of view of security, credibility would be difficult to ensure if all powers were concentrated in the hands of a single person.

A role-based approach helps to limit these risks. All personnel collaborating on testing should have specified roles and work only within the scope of these roles. Someone may have the role of “item author”, someone the role of “item reviewer”, another “test author”, or “test administrator”. Role security limits will help ensure that, for example, whoever is in charge of managing the list of tested students may never see any test items.

Responsibilities that the security team must provide include, but are not limited to, the protection of internal information from disclosure. As part of the so-called “soft security”, we take care of this confidentiality by selecting responsible employees whose moral integrity indicates that they will not spread classified information unnecessarily or succumb to the temptation to provide this information to someone for a bribe. In the concept of “hard security” (for example, when we do not have enough information about the personnel involved), a test security agreement, sometimes referred to as a non-disclosure agreement (NDA), is used to protect classified information. It is usually a unilateral, legally binding contract between the institution developing the test (or content owner) and another party solving a component item. A non-disclosure agreement usually specifies what information or materials are considered confidential and/or proprietary, what the period of confidentiality is, and what the consequences are for violating the agreement.

Recurring activities include updating legal policies and procedures and training staff on test security. At a time when most tests are produced and stored in an electronic environment, the test team’s job is to secure the test data on local or cloud servers. Access to these servers must be limited to entrusted and vetted workers, and monitored and logged.

It is good practice to require that anyone who has access to test content or other proprietary information be trained and sign a non-disclosure agreement. This includes professionals involved in test development, staff who monitor test administration, staff who process test materials and results, teachers who receive or store test materials, etc. Nondisclosure agreements should be updated annually and kept on file after the period specified in the test security plan (usually at least three years).

Preparation and Administration of Tests

There are a number of security precautions that should be taken before the test itself is given. This includes not only the secure preparation of test content, but also the monitoring of websites and social media. The danger of leaking items is magnified by content sharing technologies. There are specialized sites that collect items for certifications and exams that they have captured from individuals, accumulate them by category, and then offer them to interested parties for a fee. These pages can be found under the keyword “brain dump”. Security preparation therefore assumes that the team preparing the test will monitor social networks, try targeted queries to web search engines in search of leaked items, and monitor blogs commenting on the given exam or certification in order to identify leaked items in time.

When illegal practices are suspected, a technique known as “mystery shopping” can be used. This type of exam security verification assumes that an agreed collaborator of the test team registers as a student to take the exam and reports on the security of the test from the examinee's point of view. This form of security monitoring is expensive, but in case of doubt it can provide very valuable and otherwise unavailable data.

Given the importance, attention must also be paid to the distribution of sensitive materials and access to them. The test plan should therefore describe the procedures for how protected materials are distributed, collected and archived and who has access to them. Viewing and interventions in sensitive materials (e.g. in the wording of sharp tests) must either be recorded by technical means (logging, camera recordings) or carried out by committee (at least two people) and a record should be made of the action. An unchecked risk are individual accesses and interventions for which there is no retrospective evidence.

Finally, we need to address the issue of training. Everyone involved in the test cycle should be trained in test security. Training may cover a variety of topics, including, but not limited to, proper handling of test materials, establishing or maintaining a secure testing environment, critical aspects of a confidentiality agreement, examiner rights and responsibilities, and acceptable test supervision practices. Training may also include “what if” scenarios. Training should be tailored to align with the roles of different team members, including subject matter experts, supervisors, test administrators and coordinators, content developers, psychometricians and management. Third-party workers who collaborate on testing should also complete the training. Ensuring testing security requires the cooperation of the entire team.

Test Day Policy

Another important aspect of security is the so-called test day policy. Is the testing environment secure? Are supervisors sufficiently trained in test security? What are the registration requirements? How do participants identify themselves? How many forms of identification are needed? Is the course of the test recorded by a camera system? Is there a predetermined seating order for the test takers in the room? Is there a safe place to store personal items such as cell phones and study materials? Are calculators allowed? Is a response form supplied? If so, is the form individualized? Are the forms collected at the end of the test? Are screen protectors used for computer monitors? Are workstations separated? Are breaks or restroom breaks allowed during the test?

Communication with examinees starts well in advance of the exam date and continues until the moment the results are announced. The rules must be clearly established and distributed to interested parties and stakeholders. In addition, the consequences of breaking the rules must be clearly defined and announced. Before testing, examiners may be required to confirm that they have read, understood and agreed to follow the required rules.

For regular testing of great importance, we cannot do without some form of comprehensive test system (item bank). This, of course, brings a new kind of risk, because valuable information (the wording of items, but also the wording of prepared live tests) is concentrated here in its final form in one place for a long time, which increases the risk of unwanted exposure. Important security measures related to item banks include the technical provision of permanent logging of risk events, especially associated with test exports, or displaying a larger number of test items, or even entire tests.

Security of results

Another component of security are the procedures for the storage and distribution of sensitive materials (e.g. test assignment) and the retention of test results. This procedure determines how protected materials are distributed, collected and archived. The names and functions of the persons responsible for carrying out these procedures are also stored. Data and signatures from each person involved in testing and supervision are collected and archived as part of the test history. In general, when working with sensitive data, access must either be logged, or done under the checking of two pairs eyes, and logs must be kept of the actions performed.

Rapid Analysis of Test Data

Rapid analysis of test data is an extremely valuable tool for detecting signs of irregularities in tests that have just been completed (but not yet scored). It allows, for example, to detect possible ambiguities in the wording of items, errors in the key determining the correct answers, and the like, even before classifying students. Suspicious items, e.g. with very high or low difficulty, or with very low discriminating ability, are subjected to a content check, and in case of errors, ambiguities or inaccuracies, such an item is excluded from scoring, or the key for its scoring is modified. Authors and reviewers are also notified of problematic items to correct before further use. [157] Similarly, a quick analysis can also pick up some non-standard patterns of behavior pointing to potential security issues.

Evaluation of the Course of the Test Round

To ensure the credibility of tests, the testing organization should have a procedure for reporting incidents and irregularities in test administration and security. Test takers, teaching supervisors, and other testing personnel should be aware of the mechanism for reporting incidents, anomalies, or potential rule violations. The form offered should range from an anonymous notification to a formal message.

Incident Response

The test security plan should set out how incidents will be logged, processed and investigated. It should be clear under what circumstances the achieved score will be invalidated and when any sanctions will be applied.

Test Security Audit

During the security audit of the test, the security team recapitulates the preventive measures that were taken, their effectiveness, the threats that were noted, how they were resolved, and what adjustments to the security rules need to be made before administering subsequent tests.

9.3 Security Analysis of Tests

In the event of a violation of academic integrity, test scores may not reflect the skills and knowledge of test takers. Forensic analysis of tests (educational data forensics, EDF) is a statistical analysis of test results with the aim of detecting deviations that potentially indicate tampering, favoritism, or outright test fraud. Should there be violations of academic integrity at the level of test administrators or item bank administrators, forensic analysis is practically the only tool to systematically uncover this activity.

The analysis should answer questions of the type:

Questions focused on individuals

Is there anything unusual about this individual?

Did they answer each item with a “C”?

Were they answering too quickly?

Did they spend 10 minutes on each of the first 5 items and skip the rest?

Did they get a high score in a suspiciously short amount of time?

Did they noticeably change many incorrect answers to correct ones?

Questions focused on relationships between individuals

Are some participants' answers strikingly similar?

Were these participants sitting near each other? In the same classroom?

Does anything unusual appear when comparing this person to others?

Are there individuals around him or her who have almost the same answers?

Fig. no. 9.3.1 Histogram of test score gains illustrating an example of group-level analysis. In the circle, there is a disproportionately large group of extremely successful respondents who achieved almost the full number of points. A two-peaked distribution of scores always indicates an inhomogeneous group. In this case, these could be participants for whom the test was too easy (but then we would expect the distribution to be more “normal”). This course could therefore correspond to a situation where a limited group of respondents had the text of the test available in advance. Such a “two-peaked” test should always be paid close attention to and examined to see if other indications support possible suspicions.

Group level questions

Are some schools or teachers performing unusually well?

Do some test centers have unusually high pass rates and short test times?

Are similarly answered tests common to a certain group of test takers?

What is the common feature of this group?

Does any group of examinees answer questions from one profile subject significantly better?

Does any group of applicants respond significantly better to questions that are new or, conversely, old?

Or newly reviewed? Reviewed by one reviewer?

Are there significant differences between classrooms?

Are there significant differences between candidates from different rounds of the test?

9.3.1 Statistical Indications of Possible Fraudulent Conduct

There are many different forensic data methods that can be used to detect cheating. [158] Statistical methods for detecting suspected irregularities may include:

Evaluating the similarity of responses between pairs of examinees. The simplest methods use descriptive statistics to summarize the number (or proportion) of jointly correct answers or common errors. For example, the responses in common index (RIC) is the number of questions for which two examinees have the same answer. More complex methods work with probability estimation, whether the similarity of common answers can still be coincidental.

Analysis of changed (deleted) answers tracks the number of changed student answers on answer sheets and test programs. An implausibly large number of changed answers in a class may indicate tampering (e.g. mass copying in the absence of supervision). The number of changes from a wrong answer to a good answer is an extremely strong indicator of fraudulent behavior. [159], [160]

Analysis of predicted vs. current performance: Statistical analysis of test results from the previous year can predict future performance. Unexpectedly successful summary test results may indicate cheating, especially when large gains are not repeated in the following year, or test scores if high success rates are confirmed in subsequent years. The effect of improving results from better teaching is more gradual and long-term.

Analyzing Student Responses: It should be considered suspicious if students fail to answer a large number of easy questions while simultaneously having an unlikely number of difficult questions answered correctly. Similarly, testers can look for other statistically significant similarities across tests.

Comparison of scores between subjects: It is suspicious if there is a significant difference in results for subjects whose results are otherwise highly correlated. For example, students within one test room will score improbably high in one subject.

Mismatch between test scores and prior academic performance: If students with poor prior academic performance score high on tests, this may indicate cheating. The approach using machine learning to detect these anomalies is innovative in this regard. [161]

9.3.2 Forensic Test Analysis Tools

We are looking for a way to identify unlikely states from test data that may indicate possible cheating. Not too many user-friendly software tools for data forensics exist.

9.3.2.1 PerFit

One strategy is that we can create a graph for each student of relative success in answering items ordered by difficulty. Logically, one would expect that the graph should be a monotonically decreasing function with increasing item difficulty. Significant deviations are easily recognizable. For this analysis we can use, for example, the PerFit package in R. [162]

This involves use of the “person-fit” analysis, which shows, with a certain (about 25%) sensitivity and a certain (about 90%) specificity, a non-standard test performance for a given student. It does not have to be directly cheating (copying or knowing the questions in advance), it can also be random guessing, perhaps only on a certain part of the test, etc. Although the sensitivity and specificity of this examination are not self-saving, it can be a valuable way of extracting data that already exists anyway.

The package does not require any external data. It works with a matrix of questions and students where there is only a value of 1 (correct) or 0 (incorrect) as a dichotomous item score. The tool itself calculates the difficulty of the item and the probability of a correct answer for that student. The resulting graphs are based on the raw data from the given test, and nothing more is needed.

The procedure is well-usable for cases where everyone has the same test, or when the data for the same test can be recalculated (e.g. if everyone had the same items, only in a different order and with mixed options). The path from the matrix to the graph is straightforward, just 2-3 lines of code and you will get a graph for that student.

Fig. 9.3.2 Illustration of using PerFit to identify improbable test results. The probability of a correct answer should decrease with increasing difficulty. If it doesn't, as in this case, it's an indication of something out of the ordinary that requires attention.

9.3.2.2 SIFT

SIFT (Software for Investigation of Fraud in Testing) is a tool that uses advanced statistical methods to investigate fraud in testing. It is provided free of charge (for registration) by one of the leading suppliers of commercial test systems – Assessment Systems Corporation (ASC). A user manual and sample data are available for the program, but not support, which can be purchased separately. SIFT calculates a number of indices pointing to different types of cheating (copying, teacher assistance, missed items, etc.) and can aggregate the results by grouping by variables such as classroom, or location of the test taker within that classroom, etc. It supports all three areas of analysis – focused on individuals, on relationships between them and on groups. SIFT provides objectively measured statistics for decision-making, but their interpretation in a given situation is up to the user. [163]

9.3.2.3 CopyDetect

CopyDetect (Zopluoglu, 2016) is a package in the open-source R statistical programming language (R Core Team, 2013) that computes several cheating indices within and beyond the IRT model. Among them are the Omega index, introduced by Wollack [164], K indices [165] and S indices [166]. CopyDetect processes only one examined pair at a time. It is therefore up to the user to write a routine for processing larger amounts of data. Note that R packages are open-source software, so they should be approached with some caution.

Statistical methods allow us to express suspicion of unauthorized cooperation during completion of a test, but we should draw conclusions with caution. Statistical procedures should not be the only evidence of copying, especially when used for general screening purposes. While it is clear that the higher the agreement between responses, the more likely it is that test cheating has occurred, but even a high level of agreement is not conclusive evidence of cheating. There is always a chance that a test match is (albeit highly unlikely) the result of honest test completion. If, on the other hand, someone copies less than 10% of the items, statistical methods are not capable of distinguishing them from random phenomena.

9.4 Examples of Security Incidents

Documented and published examples of security incidents are rare because they threaten the reputation of the institution whose processes were affected by the incident. Institutions tend not to disclose information, and when they do, it is in a non-specific form that is unhelpful to those seeking instruction. This makes the cases where enough information has leaked to the public to make it possible to get an idea of how the integrity of the evaluation process was compromised all the more valuable.

Physical Therapist Exams in the Philippines

In 2007, the American Federation of State Boards of Physical Therapy (FSBPT) faced a problem. During the physiotherapist exams at a remote center in the Philippines, some of the test takers apparently had available questions that had been captured by the examinees during previous tests. These questions were then given to other test takers en masse, probably by the test center in Manila itself. There was a large number examinees, and the test was billed. It would be difficult to force everyone to repeat the test – it would shift the burden of proof to the honest ones, and risk lawsuits for lost profits due to license delays. On the other hand, to resign and pass students with suspicious test results would be to jeopardize the integrity of all testing and the good name of FSBPT.

In this situation, the federation turned to Caveon, a company that specializes in testing security and provides a wide range of services in this area. This company, or rather its subsidiary Caveon Data Forensics, was provided with complete test data from all test sites over the past two years for forensic analysis. The company used three independent statistical indicators to identify differently completed tests. First, performance on compromised test items (known to have been leaked at test preparation centers in the Philippines) was compared to performance on non-compromised test items. Second, the similarity of response patterns between candidates was examined, with higher degrees of similarity indicating the possibility of prior knowledge of the test content. A third analysis calculated the probability that a particular test taker attended the course in which the downloaded items were distributed. By combining the calculated indices, it was possible to detect fraud with a risk of error of less than 1:1,000,000. From the 23,500 tests examined, twenty were selected that had all three monitored indicators deviating from norm. Based on this, the mentioned tests were declared invalid and the remaining ones recognized as valid [167]. Note the cautious approach taken by the FSBPT Federation and Caveon. They limited themselves to invalidating only a small proportion of suspicious tests, on which the certainty of fraud was almost 100%. A suspicious result does not yet prove fraud. However, the accumulation of suspicions makes it possible to draw very relevant conclusions.

Security incidents during admission proceedings at Czech universities

Recent decades have seen several security incidents needing to be dealt with in the Czech university environment.

Case one

In 1999, the integrity of the admission procedure at the Faculty of Law of the Charles University was questioned. Because several year prior, information had been leaked that the admission procedure to this faculty was unfair, journalists called on the public to cooperate. On the day of the replacement round of the admissions procedure, a citizen who remained anonymous brought completed examples of admission tests to the Právo newspaper’s editorial offices. According to unverified information, the completed version could be bought for one hundred thousand crowns. The uncompleted one could be purchased for CZK 50,000.

In response, the faculty admitted that the completed tests had been leaked to the public, but denied responsibility. The university described it as “an organized attack by the gangster mafia against the university”. The reputation of the institution was threatened not only by the leak of the tests itself, but also by sloppy investigation and the associated suspicion that an entire system of bribery was in place at the faculty. [168]

Case two

Four years later, in 2003, there was a problem with the quality of test questions at the same faculty. The rejected students had what they considered an unfair test analyzed, and it turned out that at least nine (but more like thirty) items had factual errors. The errors occurred mainly in logical tests (tests of general academic readiness) and in general overview questions. The results had to be recalculated in order not to damage any of the study applicants. Instead of 650 students, 260 additional students had to be accepted into the first year.

After these experiences, the faculty management radically changed the admissions procedure. While questions relating to general knowledge remained under the competences of the faculty, tests and test questions on logic and general study prerequisites were entrusted to the Scio agency. Immediately before the exam, the test was distributed directly to the individual cities where the exams were held – it was not copied or stored at the faculty itself. [169]

Case three

In 2018, it became clear that since its founding, the Faculty of Medicine of the University of Ostrava (LF OU) had been allowing the bypassing of admission procedure results and admitting students who did not actually pass the test (2011). In 2018, for example, a student was admitted to the first year who scored only 43 out of 90 possible points on the profile subject test, although the threshold for admission was 46. From 2011 to 2018, around five applicants were admitted “out of order” in this way every year. These unsuccessful applicants were admitted by (mis)using the non-transparent appeal process against the result of the admission procedure.

Case four

In 2016, an attempt to defraud the admissions procedure at the 1st Faculty of Economics of the Charles University took place. During a written test in physics, an attentive teacher supervisor in the testing room noticed that one of the answer forms submitted was for a different version of the test than the current test booklet (assignment). At the same time, all the numerical markings of the test version had been scribbled over, so it was not possible to determine which version of the test originally matched the answer sheet, without careful inspection. The problem was that in the stressful situation before the start of the testing, the supervising teachers did not assign seats to the test takers, but let the candidates sit as they wished. An individual without the necessary knowledge, but well acquainted with the procedure, had made arrangements with another, well prepared student. The cheating participant then copied the entire answer form from the colleague sitting next to them. Had it not been for an attentive supervisor who registered the mismatch of versions when the tests were returned, the fraud might not have been discovered at all. The amazing thing was that the fraud had been devised and prepared with a deep knowledge of the processes according to which the testing took place at the time. At the same time, such thorough knowledge cannot be obtained on the basis of a one-time individual experience. Therefore, there is suspicion that the method was prepared by one of the companies that prepare students for entrance exams. The faculty responded promptly to the exposed attempt and changed not only the course of the test day, but also individualized the test booklet and answer sheet and added this experience to the training for supervising teachers.

A case of massive questioning of the results of the admission procedure in the USA

In 2019, a scandal erupted in the United States when it was revealed that a firm officially engaged in college admissions counseling had in fact been organizing fraud since 2011, helping, in exchange for bribes worth $25 million, about 750 students gain admission to a total of 11 elite universities. The organizer of the fraud (William Singer) bribed psychologists to issue a medical certificate regarding the applicant's health handicap, which would buy the candidate more time to fill out the test form (the certificate came to between 4,000-5,000 USD). At least two test centers had evidently collaborated with the organizers of the fraud. In at least 20 cases, fraud occurred through the use of identity confusion, where a highly competent substitute replaced the examinee. The second method of influencing the results of the admissions procedure was the obtaining of false documents regarding the practice of elite sports, which are taken into consideration when admitting students to universities in the USA. 53 people were accused in the case. A documentary film, Operation Varsity Blues: The College Admissions Scandal (2020) was made about the scandal. The scandal pointed not only to the gaps in ensuring the fair admission of applicants, but also to the special role of prestigious universities, by graduating from which students gain not only education, but also the connections and social status necessary to break into the highest levels of financially lucrative fields such as law and finance. [170], [171]

Summary

It has been shown that those with either power or information pose the greatest risk. Concentrated knowledge of the admission procedure (for example, in companies preparing students for the admission procedure) provides a temptation to misuse this knowledge to circumvent the system.

9.5 Prevention of Fraudulent Conduct

For structuring considerations about the factors affecting the probability of cheating, we can use the “fraud triangle”, a model often used to assess the probability of ethical failure in various areas of human action.

Fig. 9.5.1 The “fraud triangle”, is a visualization of the factors that together can lead to fraudulent behavior. They are: external pressure, possibly ambitions that exceed skills, opportunity, or generally a lack of control and rationalization, i.e. the possibility of justifying dishonest behavior to oneself.

The model was created on the basis of the hypothesis that trustworthy persons commit dishonest behavior if they believe that they are in a hopeless situation, and at the same time they have the opportunity to resolve this situation by breaking the rules and can somehow justify their actions to themselves. [172]

Identified factors affecting dishonesty are:

Pressure

Opportunity

Rationalization

Pressure

Let us now look at these areas in more detail to understand the reasons students cheat. [173] If we understand the motivation for fraudulent behavior, we can attempt to reduce it.[174]

Pressure

Pressure is the influence of the environment, or indirectly of one's own psyche, on the achievement of unrealistic objectives. The individual has a sense of a hopeless situation, which “forces” them to reach for an incorrect solution.

Probably the most often stated motivation for cheating is the effort to achieve a better grade than would correspond to the knowledge and skills actually acquired [175]. This can be triggered by grades being elevated from an assessment tool to a learning objective. There is often pressure for the student to get “good grades” regardless of what they actually learn. Such pressure can be created, for example, by parents, classmates, or the scholarship system. In the lower levels of education, even the grade point average can have a major impact on a student's further destiny – for example, admission to secondary school or university may depend on it. When a student's objective is to get a good grade, cheating logically becomes one of the possible ways to achieve that objective.

The real objective of higher education is to acquire the skills and competencies for a profession, job or role. Grading is a tool that merely measures the extent to which this objective is being achieved. Cheating hardly helps one gain knowledge and skills. Fraudulent behavior can thus be prevented by reducing the pressure on grades as such, and on the contrary by clearly defining the learning objectives. Students need to understand why they have to learn specific knowledge and skills and what they will be good for in practice. We should communicate the learning objectives to them in a comprehensible form and motivate them to achieve them – not to achieve a good grade. It must be clear to students that they are learning for themselves, not to just complete the subject.

It seems that even excessive intrinsic motivation can create pressure in the same way an ambitious family background can. The key is apparently the discrepancy between the actual results and the expectations that either the students themselves have or that are placed on them by the environment. The tendency to cheat thus surprisingly increases among the best (most motivated) [176].

Pressure can also result from a lack of time, or the feeling of a lack of time, to master the material. That's why it's so important to communicate with students, make sure they understand the learning objectives, and be able to estimate the time needed to prepare for the exam. The impact of the feeling of a "lack of time" is also evidenced by papers that have shown an increased tendency to cheat among students who are more involved in extracurricular activities. [177] However, it is debated whether in the background of these cases there is rather a tendency to imitate "successful" models and thus facilitate the rationalization of fraudulent behavior. [178]

It is also appropriate to help students understand the environmental pressures that affect them and teach them to rationally evaluate the significance of these environmental pressures. Setting adequate objectives and creating appropriate value rankings will help students find the right motivation.

Opportunity

Careful pedagogical supervision can help reduce opportunities for cheating on a test or exam. Studies show that the tendency to cheat is greatly reduced if students are aware that the exam is proctored. [179] The willingness to cheat is also reduced by the setting of a possible sanction in such a way as to deter subsequent students from fraudulent behavior.

Good organization of the test also reduces opportunities for cheating. The risk of copying is lower if the teacher determines the seating order for the exam than if the students themselves can choose who to sit next to. Likewise, the risk of manipulation of the test results can be reduced by making the test anonymous until it is graded. Only after the points have been assigned will the test forms again be linked to a person’s identity. Thoughtfulness and good organization of the test process can significantly limit the space for unwanted activities.

Rationalization

A cheating individual tries to find a rational justification for their behavior. If, in their mind, they can convince themselves that the school is not treating them fairly, it is easier to justify unethical behavior. Among the frequently cited reasons for increasing the tendency to cheat are exams on topics that are unnecessary and marginal from the students' point of view. [180] Students perceive this as “dishonest” behavior on the part of the school and feel entitled to cheat too.

“Cheating is common after all:” Some students state that they have no qualms about cheating on exams because “everyone does it” [181]. They do not feel that they are doing something wrong and do not perceive the social danger of cheating.

The teacher should make it clear that cheating on exams is unacceptable. Part of education should also include metacognition – understanding how I think and why. Part of metacognition is also the ability to estimate one's own possibilities. Metacognition will help set objectives, strengthen motivation to learn, strengthen academic integrity and moral principles. It is therefore necessary to talk to students continuously and within many different subjects about how and why they study and what they want to achieve.

Other Factors

Helping the weaker: Some students let others copy their exams and tests, or help them in some other illegal way. By doing so, they themselves become participants in fraudulent conduct.

Outside of exams, helping the weaker is socially valued. Even on exams, this type of illicit behavior is not clearly perceived by companies as clearly undesirable. The problem is that, in most cases, exams and tests are individual. At the same time, most university students are preparing for professions that are team-based.

Unauthorized help to another during the exam ceases to make sense when, instead of a purely individual assessment, knowledge and skills begin to be assessed as part of teamwork. In addition, this form of grading can help better prepare students for practice. The problem remains, however, that the educational system demands that we eventually “disassemble” such evaluations into individual grades. It is even necessary that we cleanse the grading of the individual from the influence of other team members. Nevertheless, the grading of teamwork should become a regular part of both formative testing and practical testing.

Cheating is advantageous: Some students choose to cheat because they find it more beneficial than investing in exam preparation. Others come to the test knowing or fearing that they are not sufficiently prepared, and cheating appears as a viable strategy to “deal with” the situation [182].

In both cases, the motivation to cheat is reinforced by the division of education into a phase where the student learns and a phase where he or she receives feedback and is graded.

Students do not tend to cheat if they feel well prepared. Similarly, the motivation to cheat decreases if the examinee has invested a lot of time and effort in preparation. A greater emphasis on active learning and the acquisition of skills during learning, along with frequent feedback, presents itself as an approach to prevent fraudulent behavior. An exam should not be an isolated act at the end of the course. It is more advantageous if the student goes through a large number of partial formative tests during the entire course. Intensive feedback helps in motivation to learn. In addition, before the final summative exam, the student has a good idea of what to expect and can realistically estimate how likely he is to meet the conditions. This reduces test anxiety, sometimes eliminates unnecessary fears of failure, and reduces the tendency to cheat.

<--! *How Schools Are Preventing Students from Cheating Online-->

10 Tools for Testing and Analysis

10.1 Software for Testing

There are dozens of tools to support formative and summative testing. New ones are constantly being created, some are losing their importance and are being abandoned. The existing ones are part of intensive development, especially through shifting towards mobile platforms. In this chapter, focused on test tools, we will not attempt a comprehensive overview of available products, as the static format of this publication does not allow us to capture dynamic developments, but will instead limit ourselves to conveying our experience with those tools that have worked for us, or on the contrary, have not.

Fig. 10.1.1 Infographic of coverage of the test cycle by Rogō, Moolde and Remark Office tools.

10.1.1 Rogō

Among the tools for electronic testing, the Rogō program occupies a prominent place. The extraordinary position of this system lies in the fact that it is a high-quality, secure and easy-to-use test tool that is also freely distributable.

Advantages (quality)

Each item has its own record of changes and use in tests.

Invited external collaborators have easy and secure access to reviews.

During remote testing, it is possible to set the end of the test after the time allowance has been exhausted.

The items used in the test will be locked.

Items cannot be added or removed from a test that is currently in progress.

Possibility to set the time limit of the test.

Student responses can be analyzed and item properties can be easily evaluated.

The grading can be adjusted according to the performance of the cohort after the end of the test using the Hofstee method.

Modified Angoff and Ebel methods are also available for setting test scores.

Security

Only HTTPS protocol is used for connection. Data is encrypted using 256-bit SSL.

Seamlessly and securely share materials across staff teams to work together on assessments.

Customizable checks and weights to ensure fair grading for all users.

They can use established local authentication systems for student and teacher access.

In the event of an Internet connection failure, only the answer on the currently open test page is lost, not the entire test.

Access to the test can be limited to selected IP addresses.

Applicability

It can run on both Windows and Linux servers.

LDAP compatibility – no need to create additional credentials for users.

Support for language mutations.

Tailored help systems - separate help for users by role.

Rogō is a web application that runs on all major browsers - Chrome, Edge, Firefox, Safari, Internet Explorer.

Adepts with special needs can adjust both the appearance of the test and the time allowance.

License

As far as license is concerned, the Rogō online testing web application is a freely distributable open source program, released under the GPL version 3.0. It is therefore possible to change the code, extend it and thus contribute to the project. In practice, it works in such a way that requests for code modifications, whether related to reporting problems or proposals for new functionality, are written into the request queue and are gradually addressed by the community.

History

The Rogō test system has been developed since 2003 at the University of Nottingham Medical School under the name “TouchStone”. The Rogō system originated at the medical school and is very well adapted especially for learning medicine. Among other things, it allows the use of interactive images in items, on which the student marks the desired object with the mouse. The system then evaluates whether the student marked the desired structure with the required accuracy.

After its success in its home faculty, it was expanded to the entire university, converted to open source software, and on that occasion renamed to avoid confusion with other systems. Rogō means “I ask” in Latin [183]. The Rogō development community received financial support from JISC, which enabled further development of the system, including the ability to translate into national languages. In the Czech Republic, the Rogō system is installed (on the servers of the First Faculty of Medicine of the Charles University) at the address https://www.rogo.cz/ and, in addition to the First Faculty of Medicine of the Charles University, it also serves other faculties of the Charles University. The First Faculty of Medicine of the Charles University has prepared a Czech translation of the environment and is continuously working on the localization of the help section. Thanks to LDAP support, students are automatically imported into the system from SIS (Study Information System of the Charles University) and can authenticate with their CAS account (Central Authentication System of the Charles University).

Specific Properties

Unlike other programs, Rogō covers and supports many steps of the test development cycle, from collaborating on the preparation of test items, to challenging them in terms of difficulty and relevance, creating a test plan, standardization, and evaluating the quality of questions. This kind of comprehensive solution brings significant advantages. For example, it is advisable to invite a number of in-house and external experts to supervise test questions, which is usually time- and organizationally demanding. At the same time, if item proposals circulate among a large number of people, it is very difficult to ensure their secrecy. In Rogō, on the other hand, opponents are prompted to join the system, so the test items never leave the system. Again, comments and suggestions are entered directly into Rogō and item authors can respond to them immediately. After the test has taken place, it is possible to display the descriptive characteristics, a histogram of the total scores of all students, or the difficulty of the items. Rogō automatically calculates discrimination indices for each item of the test, which makes it possible to identify poorly constructed items and exclude them from further use. From the point of view of applying modern procedures in testing, the system is completely unique and its introduction supports the extension of good test practices into the field.

The system makes it possible to distribute both paper and online tests, both for self-assessment and for secure summative testing. [184]

Rogō includes tools to automatically import students and courses they have enrolled into the system. Thanks to the support of the LDAP directory service, which is used by the entire Charles University, students can log in directly with their CAS account and are assigned to all their courses in Rogō.

The system allows educators to create a many types of tests and surveys:

Formative assessment

Summative assessment

Progress tests

Surveys (questionnaires)

e-OSCE (clinical trial)

Offline tests

Peer assessment (student)

Each of these types of tests and surveys can use a variety of item forms:

Area delimitation

Dichotomous items

Multiple choice questions

Extended Matching

Multiple true false

Complete the text

Mark points in the image

Likert scale

Scenario matching test

Text fields

Advantages

Table X.X Advantages of the Rogō system

Low cost of acquisition

Support for the entire testing process

Support for teamwork

High level of security

A large selection of types of test items, including multimedia ones

Disadvantages

Table X.X Disadvantages of the Rogō System

A smaller community maintaining and developing the system

Need for local support and administration

Not quite intuitive operation

Time consuming when learning the program

10.1.2 Moodle

The Moodle learning management system is a globally widespread online learning environment. More than 250 million students use this open-source platform. Moodle is probably the most widespread system for learning management at universities today. It was created in 2002 and is continuously updated. With its open source code, security and privacy, it is an attractive solution for many colleges and universities. A large and active community collaborates on the development of Moodle. Services that go beyond the capabilities of individual administrators are offered by specialized companies with Moodle partner qualifications. These are mainly services such as hosting, customization, support, training, or even comprehensive management of entire projects in Moodle.

Advantages

Price

Extension

Community of developers

Flexibility thanks to many modules

Integration options

Mobile and PC interfaces

LDAP support

Disadvantages

Unclear arrangement

Chaotic user interface

Too many modules

Absence of leadership and unified concept

Server load when multiple users work simultaneously

Upgrades may invalidate previous work

From the point of view of lean modern mobile applications, this is a behemoth

The main advantage of Moodle is its extensions and flexibility. On the other hand, it is necessary to keep in mind that with a large number of simultaneous accesses (a typical situation for testing), the system may become overwhelmed. [185] The solution lies in the appropriate dimensioning of the infrastructure, e.g. distributing the load across multiple servers. [186]

Advantages for testing

LMS Moodle is not specifically focused on testing, but it still provides some interesting options. If you use it for proctored testing, it is possible to use the Safe Exam Browser. An Ada Quiz module is available for adaptive testing. The adaptive quiz guides the student through an itinerary of questions adapted to his knowledge.

Multiple item types

Item mixing

Test timeout

Automatic grading

Localization

Good security

Disadvantages for testing

From the point of view of item banking, it is a shame that Moodle does not collect information on items from previous uses. This would allow Moodle to be used more in the role of an item bank. It is strange that the statistics in test analyses have their own nomenclature, so the teacher sometimes has to experimentally find out what the Moodle authors meant by which designation.

Little support for teamwork

Necessity of teacher training

Many other functionalities outside of testing

Unintuitive interface

Laboriousness of test preparation

However, despite these caveats, Moodle is a widely-used tool, excellent for testing. Test items can be prepared outside the Moodle environment, or they can be taken from the item bank. Moodle supports QTI interoperability standards and a range of import formats.

10.1.3 Remark Office

Testing large groups of students is often done using “pencil and paper”. The advantage is independence from technical means and very good provability in the event of a dispute. However, the evaluation of submitted response forms can be a bottleneck when using this technology. Especially when it comes to the case of a large number of test takers and a high importance of the test, it is necessary to ensure that the process is fast, as far as possible error-free, demonstrable and reproducible. Grading using a translucency does not achieve these parameters. Therefore, optical mark recognition (OMR) programs are used.

One such program is Remark Office OMR. This program is used to recognize scanned answer forms, questionnaires or tests and convert them into electronic form.

The program allows you to load data directly from the scanner or from saved files. Collected data can be exported to various data formats or processed directly. The program also enables the creation of reports on the collected data. A list of available statistics and types of reports can be found on the manufacturer's website (www.gravic.com).

A blank form must first be loaded – a template on which places for optical reading are marked. The template defines the variable type, variable name and description, etc. The software recognizes a blacked-out circle, an empty circle, and even a blacked out but later crossed out circle. It can handle incomplete answers, smudged and damaged forms, multiple marks and other anomalies. Remark reads barcodes and can recognize text (OCR). For handwritten text, it can save the field as a graphic file for further processing.

If the automation encounters a case where it doesn't know what to do, it will provide the operator with a zoomed-in view of the relevant part of the form and ask for help. It is very user-friendly and unrecognized cases are minimal. With a slight exaggeration, the company coined the slogan: If you can read it, we can read it too.

The software is very useful not only for evaluating tests and surveys, but for example for academic elections and other occasions when a large number of paper forms need to be processed.

A permanent license costs approx. CZK 30,000 and the price annual support is around 20% of this price. Considering the functionality, the price does not seem excessive. Fixing a license to a specific computer may seem impractical.

10.1.4 Socrative

Socrative (https://www.socrative.com) is a tool for online tests, quizzes and surveys. Originally, it was mainly used for frontal teaching, in which it replaced the previously used “voting buttons”. Socrative allows you to create multiple choice quizzes (true/false, single best answer and multiple true-false) as well as short answer questions. Students answer using mobile phones, and the teacher can run the quiz from a computer or from a mobile phone.

The advantage of Socrative is a well-crafted interface for teachers, which on the one hand allows for a relatively large variability of tests and quizzes and their easy adaptation to what is needed at a given moment, but on the other hand is simple and clear. Teachers are well versed in it, and working with Socrative does not unduly distract attention from the presentation or communication with students.

After the quiz has been completed, items can be scored automatically, the teacher can adjust the assessment of items with a short answer or score such items completely manually. Visualizations of how the students answered are then available, which again can be easily used in further work with students. It is also possible to download several forms of reports, from a summary of answers for each student to tables with results for teachers.

In addition to complete quizzes, it is possible to ask students individual questions. An interesting tool is available for constructed response questions where the teacher can first have students write their answers and then in a second step have them vote on which answer is best. Other features of Socrative include a feedback questionnaire at the end of the lesson.

The basic version of Socrative is free. Quizzes can be used repeatedly, new versions can be created and saved in different folders. The paid version allows you to create multiple “rooms”, which is useful if one teacher uses the tool for several different courses and has a large number of quizzes in it.

10.1.5 Kahoot!

Kahoot! is an application for creating quizzes, which is mainly used in lower levels of education. It provides a playful, stimulating environment that allows you to create engaging competitions. Again, students answer questions using a mobile phone. In addition to individual quizzes, Kahoot! and competitions between teams. The quizzes you create can be shared and you can find many different competitions, tests and quizzes online in Kahoot!, covering a wide range of topics.

10.1.6 Mentimeter

Another tool in which the teacher asks the students questions and lets them answer using their mobile phone or laptop is Mentimeter. Unlike the previous tools, it is not suitable for testing, nor is it possible to conveniently monitor which answer was given by which student. Mentimeter is particularly useful for asking attitude questions and for stimulating discussion on a certain topic. It contains a number of options to visualize the answers. The world cloud is most often used. Listeners are invited to respond with individual words or phrases (it is possible to set the number of responses that one student can submit). The result is displayed as a “cloud”, in which the word or phrase is larger the more often it occurs in the answers. Surveys using different types of scales, such as Likert scales, are also used, the results of which are again clearly visualized.

The free version has a limited number of questions that can be used in one presentation. However, a larger number of presentations can be created, already created surveys can be used repeatedly, survey results can be downloaded and further worked with.

10.1.7 Interoperability of Test Tools

If we consider the situation where the creation of test items can take place in one environment (e.g. in an item bank) and the administration of the test in another environment, it is important to ensure the portability of test items between platforms. Simple exchange formats often support transfer between only a few specific programs, or support only a few test job formats. On the other hand, these formats are clear and understandable (e.g. Aiken). At the opposite end of the notional scale of complexity are generally accepted standards of interoperability of educational systems, of which QTI is the most widely used.

Standard QTI

Question and Test Interoperability (QTI) is an open standard for the exchange of test items created by the IMS Global Learning Consortium. The QTI question exchange standard was created in an effort to prevent the work that went into item preparation from being devalued when testing technology changes. QTI is based on the XML format and defines test interoperability formats and protocols, from paper to digital, adaptive to proctored. Developers then integrate these standards into their solutions. [187]

To integrate testing tools with the learning environment, the IMS consortium also develops data exchange standards with LTI (Learning Tools Interoperability) e-learning tools. LTI standards enable the transfer of data, e.g. grades from a testing program to the learning environment. LTI complements QTI by providing a way to integrate a testing system with a learning platform such as an LMS or learning information system. [188]

10.2 Test Analysis Software

Test analysis and its reporting are important steps in the testing process. During item and test analysis, we can find out how our test behaves as a whole and what the properties of individual items are. Thanks to this feedback, we can correct and improve the test for the next rounds. There are dozens of commercial tools for analyzing tests and items. [189] There are significantly fewer freely available tools. [190] Some analytical tool modules can be included directly in testing programs (Rogō) or learning management systems (Moodle), but for a really thorough analysis, you need to use specialized tools or a statistical environment with relevant libraries.

Specialized commercial programs tend to be user-friendly, very sophisticated and relatively expensive. Freely available non-commercial solutions usually have a high difficulty threshold.

Since we assume that our readers hail mostly from the academic sphere, where high quality analysis is required and mental capacity is more readily available than financial resources, we will start with an excellent analytical tool – the statistical software R. Those who do not have a problem with entering commands from the command line can use one of the many packages in CRAN's “R library” focused on the area of “Psychometric Models and Methods”. For those who prefer a more user-friendly solution, there is the ShinyItemAnalysis web application derived from the package of the same name in the R library.

ShinyItemAnalysis

The ShinyItemAnalysis application, freely available on the web by Patrícia Martinková and her colleagues, was originally created for the analysis of university entrance tests. It also now offers a wide range of other analyzes in the field of didactic and psychological measurements. [191] It allows you to perform test and item analyses including graphic outputs (distractor analysis, two-color DD graphs, ...). You can use the pre-loaded data to test the analyses on sample data. Many methods are available, though the associated help is relatively terse. However, if you know what you need, it's not that much of a problem.

Uploading data to the system is not completely limitless. If you have a data format error, you won't get a message about the reason for the failure. You can spend a lot of time this way before you discover the problem, but don't be discouraged, it will be better the second time around. Otherwise, it's a really unique tool. You can find it at http://www.shinyitemanalysis.org/

jMetric

jMetrik is free and open source psychometric software. It was developed by J. Patrick Meyer at the University of Virginia. Psychometric methods offered include classical item analysis, reliability estimation, test scaling, differential item functioning, item response theory, Rasch models, and more. There is comprehensive help for the program. However, according to some authors, jMetrik is somewhat cumbersome. [190] A separate IRT illustrator module is now available that allows you to plot various Item Response Theory (IRT) functions. Both jMetrik and IRT illustrator are pure Java applications, working on all operating systems that have a current version of Java. Readers can find more information and the software itself at https://itemanalysis.com/

We would like to mention three of the commercially available analysis tools – Lertap, Iteman and Xcalibre, which we had the opportunity to try.

Lertap

Based in Australia, psychometrician Larry Nelson has developed a number of test analysis programs. The latest in this series, LERTAP5, is a comprehensive software package for test analysis, using Microsoft Excel. It computes analysis of test results, items, including graphic outputs. It offers cheating detection tools. Also, Lertap5 is more oriented towards classical test theory (CTT) methods, it also offers basic Rasch analyses for dichotomous test items.[190] The calculations are not the fastest, which is due to the Excel environment in which they take place. The free version includes all the features but will not process more than 250 data records. A perpetual license (bound to a computer) costs $78, placing this product on the borderline between commercial and non-commercial tools.

Iteman

Iteman is an interesting commercial software program for analyzing items and tests using classical test theory (CTT). It is unique in that it produces comprehensive and professionally processed reports in Microsoft Word format on the quality of the test items, on the test as a whole and on its psychometric properties, including embedded graphics and tables. A description of one of the older versions is provided by Byčkovský [192] Iteman is now available (in version 4) either as a cloud version or as an application for Windows. The cloud version allows you to use the software anywhere. The version for Windows, on the other hand, allows you to store all (potentially sensitive) data on one computer. A non-commercial (academic) license for Iteman costs $1,295 per year. The trial version is limited to 100 students and 100 items. The reader can find more information, a description of the current version and licensing terms on the developer's website http://www.assess.com/.

Xcalibre

Xcalibre 4, from the same manufacturer, is a powerful test analysis tool based on item response theory (IRT). The program has a very user-friendly interface. It provides professional reports summarizing analysis results, including embedded tables and graphs. It allows you to analyze large data sets, perform comparative studies, or refine items using distractor analysis. Control is point-and-click, you don't need to write any code. The non-commercial version of Xcalibre costs $1,495 per year.

11 Appendices

11.1 Evaluation based on direct observation

Note: These are direct testing methods – in contrast to indirect methods, to which most of the text is devoted. Compare the opening paragraphs of Ch. 3.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5295751/

11.1.1 MMI

https://en.wikipedia.org/wiki/Multiple_mini-interview

11.1.2 OSCE

https://rogo-eassessment-docs.atlassian.net/wiki/spaces/ROGO/pages/491548/Functional+Specification

https://rogo-eassessment-docs.atlassian.net/wiki/spaces/ROGO/pages/1049414/OSCE+Stations

11.2 Test Anxiety

A test result can also be influenced by the examinee's emotions. The most talked about in this context is anxiety (test anxiety, exam stress) [193]

Most people experience some level of stress or anxiety before an exam. It can even motivate some individuals to perform better. However, if the level of stress is such that it negatively affects the test-taker's performance during the exam, we’re dealing with test anxiety.

Fig. No. 11.1.1 Yerkes-Dodson law, describing the relationship between stimulation and performance. [194] It seems, meanwhile, that for the performance of more complex activities, a lower level of stimulation is optimal than for simpler activities [195].

Examining the relationship between performance and stress in general, it has been shown that different items require different levels of stimulation for optimal performance. For example, difficult or intellectually demanding items require lower levels of stimulation (to facilitate concentration), while items requiring endurance are better performed with higher levels of stimulation (increased motivation). The Yerkes-Dodson law, which describes the relationship between stimulation (motivation, stress) and performance, suggests that an individual's performance increases only up to a certain degree of excitation. Further increasing the level of motivation and excitement begins to reduce performance, while the threshold level of useful excitation is individual. The mechanism is probably related to the action of stress hormones. Situations that are new, unpredictable, beyond the individual's control, or carry the risk of negative social evaluation (exclusion) lead to the induction of stress. [196] However, changes in performance do not manifest themselves in the same way for all its types. For example, stress improves the memorization of factual data, but at the same time creative functions deteriorate. Speed increases, but accuracy decreases. A load of appropriate intensity can improve performance, or at least some of its components, in the learning phase. However, if we accept the fact that the tests are primarily intended to test understanding and skills, not just the recall of isolated facts, we require the involvement of higher (“more creative”) cognitive functions during the test. Thus, even mild stress appears to impair performance during testing. As much as we may debate the appropriateness of the load during instruction, it is generally accepted that we should keep stressors to a minimum during the time of examination and testing. [197][198][199][200]

Test anxiety is one of the factors that reduce the reliability of the test. It is one of the forms of so-called academic anxiety. It is triggered on the one hand by context-specific stimuli (e.g. briefing before a test) and on the other hand by reactions specific to an academic subject (e.g. anxiety about mathematics). Test anxiety is estimated to affect about 15 to 22% of students. [193]

The degree of test anxiety usually depends on the type and meaning of the exam. The largest is usually for exams of great importance. It is affected by a number of factors. During the more than fifty years that test anxiety has been studied in more detail, a number of theoretical models of the phenomenon have emerged. Most are based on the two oldest concepts. The interference model of test anxiety posits that poorer performance on a test can be explained by factors (e.g., emotions or worries) that interfere with recalling and working with information. In contrast, the deficit model of test anxiety assumes that test anxiety is a consequence of insufficient knowledge and skills, including, for example, the ability to study effectively, the perception of one's own abilities (self-efficacy), motivation, or mastering strategies for completing the test. Neither of these two models can fully explain the variability and dynamics of test anxiety, so other theoretical concepts arise. Newer approaches also include external environmental influences as well as social influences, for example the environment in which education takes place and the relationships between students and teachers and students with each other.

It can be generalized, to a large degree, that text anxiety is milder in students who:

have better study results

had better results on entrance exams

have better cognitive and verbal skills

have greater self-confidence

expect the test to be easier or consider it easier

The connection with motivation is interesting. A student's internal motivation to study reduces test anxiety. On the other hand, test anxiety is increased by external motivation, especially when it comes to negative motivation (e.g. emphasizing the possible consequences of failure).

Similarly, test anxiety is related to problem-solving skills. It is lower in people who, when overcoming obstacles, use strategies aimed at eliminating or overcoming the stressor. In contrast, individuals who choose avoidant strategies show greater levels of test anxiety.

The level of test anxiety also correlates with some demographic predictors. Females tend to have greater test anxiety, although when comparing different studies, it seems that the gender dependence gradually weakens [197] [193]. However, people who perceive themselves as members of a certain minority are significantly more affected by test anxiety.

Prevention and mitigation of the effects of test anxiety

Test anxiety reduces the reliability of the test. It impairs the performance of some test takers and does not allow them to fully utilize the knowledge and skills they have when completing the test. At the same time, it manifests itself to a different extent in different test subjects, so it becomes a source of variability that cannot be overlooked in the final test score. It is therefore desirable to prevent test anxiety, or to minimize its effects.

Strategies for the prevention and management of test anxiety can be divided into measures on the part of the teacher (organizer of the test) and measures on the side of the test taker.

The approaches recommended for students affected by test anxiety are mostly based on sufficient preparation for the test, psychohygiene, relaxation techniques, increasing self-confidence, overcoming unrealistic fears, etc. Some educational institutions organize programs in which they try to intervene with students affected by test anxiety and teach them to overcome test anxiety [201].

The teacher, for their part, must do more than just create the test or prepare for it. The overall learning setting is key. It is essential that the test tests what is being taught, i.e. that its content is not surprising to students [202]. The test must therefore be appropriately planned and valid. It is also important that students understand how the test will be graded and have confidence in the grading.

Supporting a metacognitive approach to learning is effective against test anxiety [203]. Students should understand why they are learning, what the learning objectives are, how and why the learning takes place, what are the components of the educational process, the significance of the test, etc. Social support and the creation of social ties are also important. Test anxiety is greater in students who learn in isolation from others. Including group work in learning reduces test anxiety.

Relatively simple measures that a teacher can take to reduce test anxiety include, for example:

Introducing the topics and scope of the test to the students in advance.

Allow students to try out the test environment in advance, especially if the test is in electronic form.

Familiarize students with the format of the questions and the method of answering them in advance. If the answers are given using a form, explain exactly how to work with the form.

Allow students to take a “mock test”. The mock test can be very short, with only a few questions, but it should contain all the elements of the real test (e.g. workplace preparation, identity verification, the same way of entering and answering questions).

Discuss in advance with the students the topics that are key and will appear in the test. This will reduce the so-called … content uncertainty”.

Help students schedule the time needed to prepare for the exam.

More important tests should be preceded by a series of partial formative assessments that "lead" the student to the summative exam, show him how much he achieves the expected knowledge and skills, what his weak and strong points are, and at the same time gradually prepare him for the content and scope of the summative exam.

During the exam itself, we try to minimize factors that could distract and disrupt students. At the time of the exam, we try to prepare a predictable and friendly environment.

One thing that might stand in the way of reducing test anxiety is the fact that some professions involve high-stress work environments that preclude people who could be thought of as “hothouse flowers” entering them. It may then be a legitimate requirement for the test to verify that the student can work efficiently and accurately even under stress. In this case, however, students should be exposed to pressure mainly in the course of learning and formative assessments, not in a standardized, final, summative exam. Work in stressful situations can also be part of practical testing. However, in most written exams and tests, test anxiety is undesirable, because by reducing the reliability of the test, it ultimately threatens its validity and evaluability.

11.3 Costs of Testing

The preparation of top-notch test questions, and the execution and grading of tests are professionally demanding, laborious and expensive items. The cost of a single test can be expected to drop as the number of students being tested increases (economies of scale), because the fixed costs are spread over multiple units.

For small numbers of examinees, it is definitely less laborious to take oral exams than to prepare, administer and grade tests. There are usually good reasons for deciding to test a small number of students using tests. Assessment of students by means of tests may be the method of choice if we need the results to be demonstrably objective and reproducible, e.g., when there is a risk of appeals from the test takers. From an economic point of view, testing pays off with a large number of test takers, as high acquisition costs will be balanced by low operating costs.

Item cost

The creation of high-quality test questions requires a team of experts in the relevant field, who are also trained in the methodology of test creation. Additional expenses come from reviewing questions and pilot testing with a large enough group of students.

The cost of preparing calibrated items for critical tests is estimated at $1,000 per question and generally speaking, the cost per item does not fall below $300. [68]

Rudner, for example, calculated the total cost of developing one quality item for an adaptive acceptance test. [204] He showed that a good quality calibrated item costs (in the US) USD 1,500–2000. If we compare this figure with Breithaupt's [205] estimate of the required number of items in an item bank for common adaptive testing (about 2,000 items in the bank), we arrive at an astronomical sum of 3-4 million USD for the content of the item bank. [206]

Item costs can be reduced if we can reuse already finished items. These may be items that we have prepared and calibrated in previous rounds of testing, as long as we are sure they have not been exposed by previous use.

Item Cost for the First Faculty of Medicine of Charles University

At the 1st Faculty of Medicine of Charles University, we calculated the costs of a new item in the academic year 2020/21. 230 items were prepared, at a final price of CZK 1,500 per item. The cost of the work of authors and reviewers and, to a lesser extent, the costs of operating and acquiring an item bank are reflected in the price. It is obvious that compared to foreign countries, we have the advantage of qualified labor at lower prices.

Cost Per Examinee

It is interesting to look at the cost of testing per student tested. A large dispersion of data is evident here, due to different conditions and requirements.

In a 1996 report, the Center for Research on Educational Standards and Student Testing (CRESST) calculated the cost of test-based assessment with teacher salaries factored in. His estimates ranged from USD 848 to USD 1,792 per student. [207]

How significant economies of scale can be is shown by the example of the well-known ACT and SAT exams, which are part of the university admissions process in the United States. In 2020, 1.65 and 1.1 million examinees took these exams. Prices for these tests, which cover all costs of test development, scoring, and administration, range from USD 20 to USD 70.

Although cost-benefit estimates are usually context- and scale-dependent, it is clear that the benefits of computer-assisted testing substantially outweigh its costs. [208]

The cost per examinee at the First Faculty of Medicine, Charles University

Again, we can compare these costs with the costs of the admissions process for the First Faculty of Economics of the Charles University in the academic year 2020/21. If we limit ourselves to master's fields, for which there were approximately 4,000 applicants, we get an approximate cost of CZK 250 per examinee. Of this, the so-called “test day costs” make up the largest portion of the total costs, representing more than a third of the total amount. Another third or so goes to printing costs, and the rest is divided into item creation and item bank costs. Here, too, it is evident that we are at a fraction (about a third) of the compared costs of SAT and ACT exams, despite the disparity in the number of test takers. We attribute the lower cost ratio than in the case of question creation, where the ratio was more than 1:20, to the difference in the number of test takers and the incompressibility of printing costs, which are not as geographically sensitive as labor costs.

11.4 Abbreviations in texts on testing

A-level

Advanced Level (General Certificate of Education Advanced Level, as it is fully named) is the designation of the certificate and exams that are part of the state matriculation examination in Great Britain. A-levels are accepted at many universities as one of the important indicators of the suitability of applicants for university study. In order to select students for medical schools, the highest grades in three subjects (A) were required, which had to include chemistry and at least one other natural science or mathematics.

ACT

American College Testing is one of the two most widely used college admissions tests in the US and Canada. Most schools give students a choice of which of the tests to take, and only set the number of points necessary for acceptance on the chosen test. The ACT consists of four parts: English (45 min.), Math (60 min.), Reading (35 min.), and Scientific Thinking (35 min.).

AERA

American Educational Research Association

AIG

Automatic Item Generation – automatic creation of test items.

AMEE

Association for Medical Education in Europe – originally a European, now worldwide, association for the education of doctors.

APA

American Psychological Association

BMAT

BioMedical Admissions Test – this is a used in Great Britain for admission to traditional medical schools with a division of preclinical and clinical education and an emphasis on science education in the first years of study. The BMAT is the successor to the Medical and Veterinary Admissions Test (MVAT). There are discussions about the benefits of this test, or its individual parts.

BTU

Test Item Bank is a database of test items, allowing storing of information with each item, making it possible to also store information on its creation, use and psychometric properties.

CAA

Computer Assisted Assessment

CAT

Computer Adaptive Testing.

CBA

Competency based assessment – assessment based on competencies. It compares individuals with the required standards.

CBA/CBT

Computer-Based Assessment/Testing – assessment (testing) via computer or similar device, e.g. tablet, mobile phone, etc. (opposite to PPT).

CBD

Case-Based Discussion refers to a structured, clinician-led discussion of clinical cases testing clinical reasoning.

CFT

Computerized Fixed-form Tests – Classic test in computerized form.

Class rank

A ranking of a high school student's performance compared to other students in the class. See also high school class rank (HSCR).

CRT

Criterion-Referenced Test – standardized test comparing student performance to predetermined standards required to pass the test (compare with NRT).

CTT

Classical Test Theory – classical test theory is a test analysis tool. It is simpler and easier to use than the more sophisticated IRT (item response theory) test analysis tool.

DIF

Differential Item Functioning – the index of differential functioning of the item indicates different behavior of the test question for groups with the same level of knowledge (ability, academic performance), but different ethnic or gender composition.

DOPS

Direct Observation of Procedural Skills

ECD

Evidence-Centered Assessment Design – an evidence-based approach to the construction of evaluations of learning outcomes.

EDF

Educational Data Forensics – test security analyses.

EMQ

EMQ or also EMI (Extended Matching Question/Item) are “extended matching questions”. These are choice questions with a single best answer, in which the test taker chooses from a larger number (typically around twenty) options. As a rule, the same set of options is used for several items that are consecutive on the test.

ETS

Educational Testing Service is a non-profit educational organization focused on testing and grading. ETS develops various standardized tests for secondary and higher education in the US and also administers international tests including TOEFL language tests.

FYGPA

First Year Grade Point Average – academic averages in the first year of college are used to estimate a student's academic performance in college.

GAMSAT

The Graduate Australian Medical School Admissions Test is a test for the selection of applicants for Masters of Medicine, Dentistry and Veterinary Medicine developed in 1995 by the Australian Council for Educational Research (ACER). It is used to select students for Master's degree studies at health colleges in Australia and, since 1999, and at some schools in Great Britain and Ireland.

GAT

General Aptitude Test – the test of general academic readiness (GAR) is a collective name for ability tests that test applicants' mental skills (in contrast to knowledge tests). In the GAR test, there are usually questions focused on spatial relationships, logical connections and visualization/imagination.

GCSE

General Certificate of Secondary Education – a certificate of secondary education in the relevant field used in England, comparable to a Certificate of Graduation. It is issued to 14-16-year-old students after they pass the relevant exam.

GDPR

General Data Protection Regulation – is a general regulation on the protection of personal data enshrined in EU legislation for the protection of citizens' personal data.

GPA

Grade Point Average – academic averages. Academic averages from secondary school are often included among the selection criteria for admission to some faculties as a good predictor of study success.

HEI

Higher Education Institution – colleges, universities

HSCR

High School Class Rank is a measure of the academic performance of a particular student relative to the performance of others in the class. Another designation for this parameter is Class Rank. It is calculated as a student's rank determined based on GPA and divided by the number of students in the class. The result is the percentile of the best students to which the student belongs. Especially abroad, some secondary schools provide this information. Large public schools provide this figure more often than small private schools. HSCR, along with GPA, is often used to evaluate a student in college admissions

HSGPA

High-School Grade Point Average – high school academic grade average (USA). See also GPA and uGPA.

ICC

Item Characteristic Curve – the characteristic curve of the item expresses the relationship between the measured latent feature of the test taker (knowledge) and the probability of a correct answer to the item. It is used in item response theory. It is another name for ICF.

ICF

Item Characteristic Function – the characteristic function of the item that expresses the relationship between the measured latent trait of the test taker (knowledge) and the probability of a correct answer to the item. It is used in item response theory.

IDEAL

IDEAL Consortium (International Database for Enhanced Assessments and Learning) – a voluntary association of 23 universities from around the world sharing medical test items in English.

IMS

Item Management System – item bank in the broader sense of the word – a system for creating, storing, sharing and delivering items.

IRT

Item Response Theory – a modern test analysis tool that allows estimating item properties for different proficiency levels.

IRM

Item Response Models – refers to a group of mathematical models that attempt to explain the relationship between latent traits and their manifestations (i.e., observed outcomes, responses, or performance) using item response theory (IRT).

JISC

Joint Information Systems Committee – the joint committee for information systems is a non-governmental public institution in the UK whose mission is to support higher education through the implementation of information and communication technologies.

LMS

Learning Management System – systems for organizing learning, such as: Moodle, BlackBoard, Adobe Connect, etc.

LTI

Learning Tools Interoperability – standard for collaboration of e-learning tools and environments.

MCAT

Medical College Admission Test – a standardized “computer-based” admissions test for US medical schools, established by the Association of American Medical Colleges in the US.

MCAT(R)

In 1991–2, the MCAT was again revised and restructured, and its new form bears the above designation.

MCQ

Multiple Choice Question – multiple-choice question, the most general category that includes all forms of multiple-choice test problems.

MEFANET

MEdical FAculties NETwork – a voluntary association of Czech and Slovak medical and health faculties cooperating on electronic education support.

MEQ

Modified Essay Question is a format of open-ended questions in which a constructed answer is expected to be longer than in the SAQ, but considerably shorter than an essay. The test taker gradually answers partial questions, among which he obtains various additional information. Analytical reasoning, data interpretation and critical decision-making can be assessed using this type of question.

MeSH

Medical Subject Headings – a controlled dictionary of descriptors for indexing in medicine and biology.

MRQs

Multiple Response Questions denotes multiple-choice questions in which more than one answer is correctly (and must be marked) offered – a synonym for MTF.

MSC-AA

Medical Schools Council Assessment Alliance is an organization of medical schools in Great Britain working together to assess the outcomes of undergraduate learning. Its predecessor was UMAP

MSF

Multisource Feedback – (also referred to as 360-degree feedback) This is a multisource assessment method of assessment in which feedback is provided to an individual by an imaginary circle of respondents who come into contact with him and compare with his self-assessment. The benefit of this assessment is that it provides information about how the individual is doing in the eyes of others.

MTF

Multiple True/False question. Multiple choice question, several of which may be correct. For each offered answer, the test taker considers whether it is correct or incorrect. In practice, it is often confused with the much more general term MCQ.

NBME

National Board of Medical Examiners is an independent, non-profit organization that deals with assessing the quality of education of healthcare workers. The NBME develops and administers the USMLE (National Medical Licensing Examination).

NCME

National Council on Measurement in Education – The National Council on Educational Measurement is an American professional organization for individuals involved in assessment, evaluation, testing, and other aspects of measuring educational outcomes. It publishes a quarterly magazine, The Journal of Educational Measurement (JEM).

NDA

Non-disclosure agreement (confidentiality agreement).

NRT

Norm-Referenced Testing is a standardized test in which an individual’s performance is evaluated in comparison to the performance of the relevant (compare with CRT).

OMR

Optical Mark Recognition – optical character recognition used in machine evaluation of answer forms.

OSCE/OSPE

Objective Structured Clinical/Practical Examination – objectively structured clinical/practical examination is a way of objectively evaluating the results of learning clinical/practical skills. The test is usually organized in the form of 5-10 minute stops at stations where the examinee solves the relevant item.

P&P

Using Paper-and-Pencil

PPT

Paper-and-Pencil Testing

QTI

Question and Test Interoperability specification – international standard for interoperability of testing systems.

RIR

Item Rest Correlation – The RIR Index (or just RIR) is the correlation coefficient between success on a given test item and the total number of points for the test when the given item is excluded. The RIR coefficient takes on values from -1 to 1 and is used to evaluate the discriminating ability of an item. A well-discriminating item should achieve an RIR value of at least 0.3. Significantly smaller or negative values indicate that the item is not discriminative or discriminates in the opposite way to the test

RIT

Item Test Correlation – The RIT Index (or just RIT) indicates the correlation coefficient between success in a given test question and the total number of points on the test. The RIT coefficient is used similarly to the RIR index.

RIC

The Responses in Common index is the number of items to which two examinees gave the same answer.

RTE

Response Time Effort – Response time for completion of the question

SAQ

A Short-Answer Question is an open-ended question to which the respondent has to create a very short answer (one word, word combination). There may be several answers that are evaluated as correct.

SAT

Standardized Admissions Tests – together with ACTs, are two of the most widely used tests for determining the college preparedness of high school students in the USA. In its current form valid since 2005, the SAT lasts three and three-quarter hours and consists of three parts: critical reading, math and writing. Up to 800 points can be achieved for each part. An “experimental” section is included in the test, which is not used to evaluate the student's ability, but to evaluate the question itself, for possible future use in SAT tests.

SBA

Single Best Answer Question. A question with a choice of usually three to five answer options. The test taker chooses one of the offered answers. The other options (distractors) are either incorrect or (more often) qualitatively significantly less suitable answers to the question.

uGPA

Undergraduate grade point average – high school grade point average, often used to predict success in studies at university. See also GPA and HSGPA.

UKCAT

UK Clinical Aptitude Test – is a test created in 2006 by a consortium of British medical and dental faculties to test applicants' mental skills. UKCAT is designed to test skills and attitudes, not academic achievement – which is well predicted by A-levels, GCSEs or GPA. The test therefore focuses on critical logical thinking and the ability to draw conclusions. The benefit of this test is debatable and is stimulating a discussion in the UK about the suitability of psychological testing of applicants for the selection of medical students.

ULI

Upper-Lower index is an index for the assessment of sensitivity, or the discriminative ability of an item.

UMAP

Universities Medical Assessment Partnership – formerly a voluntary association of medical schools in Great Britain founded in 2003 to work together to create and share test questions. In 2009, the association transformed into the current MSC-AA.

UMAT

The Undergraduate Medicine and Health Sciences Admission Test is used to select high school applicants to medical school in Australia and New Zealand. After completing a bachelor's degree, applicants for a follow-up "master's" degree are selected using the GAMSAT

USMLE

United States Medical Licensing Examination is an official test for graduates in medicine for entrance into postgraduate programs in clinical medicine in the USA.

VLE

Virtual Learning Environment – for example, Moodle.

VSP

The General Academic Readiness Test is a collective name for aptitude tests that test applicants' mental skills. GAT tests usually include questions focused on spatial relationships, logical connections and imagination. See also GAT.

12 References

Citation in Harvard format

Schindler, R., 2006. Rukověť autora testových úloh, Praha: Centrum pro zjišťování výsledků vzdělávání.

ŠTUKA, Čestmír, Patrícia MARTINKOVÁ, Martin VEJRAŽKA, Jan TRNKA a Martin KOMENDA. Testování při výuce medicíny: konstrukce a analýza testů na lékařských fakultách. Praha: Karolinum, 2013, 155 s. ISBN 978-80-246-2369-6.

The aims of higher education. 1. Pretoria: Council on Higher Education, 2013. ISBN 978-1-919856-84-1

SOOZANDEHFAR, Seyyed Mohammad Ali a Mohammad Reza ADELI. A Critical Appraisal of Bloom’s Taxonomy. American Research Journal of English and Literature: An Academic Publishing House [online]. 2016, 2016(2), 1-9 [cit. 2021-11-16]. ISSN 2378-9026. Available at: https://www.arjonline.org/papers/arjel/v2-i1/14.pdf

ROLAND, Case. The unfortuate consequences of Bloom’s taxonomy. Social Education [online]. National Council for the Social Studies, 2013, 77(4), 196-200 [cit. 2021-10-11]. ISSN 0037-7724.

TUTKUN, Omer, DILEK GÜZEL, MURAT KOROĞLU a HILAL İLHAN. Bloom's Revized Taxonomy and Critics on It. In: MURAT, Koroğlu a İlhan HILAL. The Online Journal of Counselling and Education [online]. 2012, pp. 23-30 [cit. 2021-10-11]. ISSN 2146-8192. Available at: https://www.researchgate.net/publication/299850265_Bloom's_Revized_Taxonomy_and_Critics_on_It

BERGER, Ron. Here’s What’s Wrong With Bloom’s Taxonomy: A Deeper Learning Perspective. Education Week [online]. 2018, 14.3.2018 [cit. 2021-11-16]. Available at: https://www.edweek.org/education/opinion-heres-whats-wrong-with-blooms-taxonomy-a-deeper-learning-perspective/2018/03

Criticisms of Bloom's Taxonomy: Educational theorists have criticized Bloom’s Taxonomy on a few grounds. Teachers commons: A place for teachers to share [online]. 24.4.2008 [cit. 2021-11-16]. Available at: http://teachercommons.blogspot.com/2008/04/bloom-taxonomy-criticisms.html

O´NEIL, Geraldine a Feargal MURPHY. UCD DUBLIN. Assessment: Guide to Taxonomies of Learning [online]. Dublin: UCD TEACHING AND LEARNING, 2010 [cit. 2021-10-11]. Available at: https://www.ucd.ie/t4cms/ucdtla0034.pdf

MALAMED, Connie. Alternatives to Bloom’s Taxonomy for Workplace Learning. The eLearnin Coach: Helping you design smarter learning experiences [online]. 2020 [cit. 2021-11-16]. Available at: https://theelearningcoach.com/elearning_design/alternatives-to-blooms-taxonomy/

MILLER, G. E. The assessment of clinical skills/competence/performance. Academical Medicine. 1990, vol. 65, no. 9 Suppl, s. 63-7, ISSN 1040-2446. PMID: 2400509

PEÑALVER, Elena Alcalde. Financial Translation. CUI, Ying a Wei ZHAO, ed. Handbook of Research on Teaching Methods in Language Translation and Interpretation [online]. IGI Global, 2015, 2015, pp. 102-117 [cit. 2021-11-16]. Advances in Educational Technologies and Instructional Design. ISBN 9781466666153. Available at: doi:10.4018/978-1-4666-6615-3.ch007

KREVIČ, Nataša. Katalyzátor změn vyučování?: Inovace v hodnocení žáků. Pro vzdělávání: Školské poradenské zařízení a zařízení pro další vzdělávání pedagogických pracovníků [online]. Praha: Národní ústav pro vzdělávání, 2019 [cit. 2021-11-16]. Available at: http://provzdelavani.nuv.cz/clanky/ze-zahranici/katalyzator-zmen-vyucovani-inovace-v-hodnoceni-za

CRUESS, Richard L., Sylvia R. CRUESS a Yvonne STEINERT. Amending Miller’s Pyramid to Include Professional Identity Formation. Academic Medicine [online]. 2016, 91(2), 180-185 [cit. 2021-11-16]. ISSN 1040-2446. Available at: doi:10.1097/ACM.0000000000000913

Definitions: Workplace-based assessment (WPBA). Assessment Department [online]. London: The Royal College of Pathologists, 2019 [cit. 2021-11-16]. Available at: https://www.rcpath.org/trainees/assessment/workplace-based-assessment-wpba.html

PRAKASH, Jyoti, K CHATTERJEE, K SRIVASTAVA, VS CHAUHAN a R SHARMA. Workplace based assessment: A review of available tools and their relevance. Industrial Psychiatry Journal [online]. 2020, 29(2) [cit. 2021-11-16]. ISSN 0972-6748. Available at: doi:10.4103/ipj.ipj_225_20

KLENOWSKI, Val, Sue ASKEW a Eileen CARNELL. Portfolios for learning, assessment and professional development in higher education. Assessment & Evaluation in Higher Education [online]. 2006, 31(3), 267-286 [cit. 2021-11-16]. ISSN 0260-2938. Available at: doi:10.1080/02602930500352816

SEIFERT, Kelvin. Advantages and disadvantages. Educational Psychology [online]. OpenStax CNX, 2011, pp. 318-320 [cit. 2021-10-11]. Available at: https://www.opentextbooks.org.hk/ditatopic/6468

HERMAN, Joan L. a Stephen A. ZUNIGA. Assessment: Portfolio Assessment. Education Encyclopedia [online]. [cit. 2021-11-16]. Available at: https://education.stateuniversity.com/pages/1769/Assessment-PORTFOLIO-ASSESSMENT.html

STITT-BERGH, Monica a Yao HILL. What is a portfolio?: Using Portfolios in Program Assessment. Assessment and Curriculum Support Center: Learning outcomes assessment for improvement [online]. Manoa, Hawaii: University of Hawaiʻi at Mānoa [cit. 2021-11-16]. Available at: https://manoa.hawaii.edu/assessment/resources/using-portfolios-in-program-assessment/

DRIESSEN, Erik, Jan VAN TARTWIJK, Cees VAN DER VLEUTEN a Val WASS. Portfolios in medical education: why do they meet with mixed success? A systematic review. Medical Education [online]. 2007, 41(12), 1224-1233 [cit. 2021-11-16]. ISSN 03080110. Available at: doi:10.1111/j.1365-2923.2007.02944.x

VLEUTEN, Cees van der. OSCEs by Cees van der Vleuten. Maastricht University [online]. 2019 [cit. 2021-11-13]. Available at: https://www.maastrichtuniversity.nl/news-events/newsletters/article/+5u+DZKHLUQtFBjwefD8Tg

The Limits of Standardized Tests for Diagnosing and Assisting Student Learning. FairTest [online]. Jamaica Plain: National Center for Fair & Open Testing, 2007 [cit. 2021-11-16]. Available at: https://fairtest.org/limits-standardized-tests-diagnosing-and-assisting

FRANZ, Riffert. The Use and Misuse of Standardized Testing: A Whiteheadian Point of View. Interchange [online]. Salzburg: University of Salzburg, 2005, 36(1-2), 231-252 [cit. 2021-10-12]. ISSN 0826-4805. Available at: doi:10.1007/s10780-005-2360-0

Herman JL. A Practical Guide to Alternative Assessment. Association for Supervision and Curriculum Development. Alexandria, VA: 1992. on-line https://eric.ed.gov/?id=ED352389

VAN DER VLEUTEN, C P, G R NORMAN a E DE GRAAFF. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ [online]. 1991, vol. 25, no. 2, pp. 110-8, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/2023552>. ISSN 0308-0110.

ABDELLATIF, Hussein a Abdullah M. AL-SHAHRANI. Effect of blueprinting methods on test difficulty, discrimination, and reliability indices: cross-sectional study in an integrated learning program. Advances in Medical Education and Practice. 2019, 10, 23-30. ISSN 1179-7258. Available at: doi:10.2147/AMEP.S190827

PATIL, SunitaY, Manasi GOSAVI, HemaB BANNUR a Ashwini RATNAKAR. Blueprinting in assessment: A tool to increase the validity of undergraduate written examinations in pathology. International Journal of Applied and Basic Medical Research. 2015, 5(4), 76-. ISSN 2229-516x. Available at: doi:10.4103/2229-516X.162286

CHVÁL, Martin, Ivana PROCHÁZKOVÁ a Jana STRAKOVÁ. Hodnocení výsledků vzdělávání didaktickými testy. 1. Plzeň: Česká školní inspekce, 2015. ISBN 978-80-905632-9-2.pg. 113.

KUBISZYN, Tom a Gary BORICH. Educational Testing and Measurement. - vydání. Wiley, 2000. 530p. ISBN 9780471364962.

JOLLY, Brian. Written examinations [online] . In SAWANWICK, Tim. Understanding medical education: Theory and practice. 1. vydání. Oxford : Wiley-Blackwell, 2010. 464 s. s. 208-230. Dostupné také z <http://dx.doi.org/10.1002/9781444320282.ch15>. doi: 10.1002/9781444320282.ch15. ISBN 978-1-4051-9680-2

Testwise. APA Dictionary of Psychology [online]. Washington DC: American Psychological Association, 2020 [cit. 2021-11-16]. Available at: https://dictionary.apa.org/testwise

Testwiseness and Guessing: What is testwiseness and guessing? [online]. Lawrence: The University of Kansas [cit. 2021-11-16]. Available at: http://www.specialconnections.ku.edu/?q=assessment/quality_test_construction/teacher_tools/testwiseness_and_guessing

BERK, Ronald A. Humor as an instructional defibrillator: Evidence-based techniques in teaching and assessment. Sterling, Va.: Stylus, 2002, 268 s. ISBN 1579220630

AL-FARIS, Eiad A, Ibrahim A ALORAINY a Ahmad A ABDEL-HAMEED, et al. A practical discussion to avoid common pitfalls when constructing multiple choice questions items. J Family Community Med [online]. 2010, vol. 17, no. 2, s. 96-102, dostupné také z <https://doi.org/10.4103/1319-1683.71992>. ISSN 1319-1683 (print), 2229-340X.

DRASGOW, Fritz, Richard M LUECHT a Randy E BENNETT. Technology and testing. In Brennan, Robert L. Educational measurement. 4. vydání. Praeger Publishers, 2006. 779 s. Washington, DC: American Council on Education. ISBN 0275981258, 9780275981259

GIERL, Mark J a Thomas M HALADYNA. Automatic item generation: Theory and practice. 1. vydání. New York : Routledge, 2012. 256 s. ISBN 978-0-415-89750-1.

GIERL, Mark J. a Hollis LAI. The Role of Item Models in Automatic Item Generation. International Journal of Testing [online]. 2012, 12(3), 273-298 [cit. 2021-9-26]. ISSN 1530-5058. Available at: doi:10.1080/15305058.2011.635830

GIERL, Mark J, Hollis LAI a Simon R TURNER. Using automatic item generation to create multiple-choice test items. Medical Education [online]. 2012, 46(8), 757-765 [cit. 2021-10-4]. ISSN 03080110. Available at: doi:10.1111/j.1365-2923.2012.04289.x

FIŘTOVÁ, Lenka. Klonování úloh jako cesta k vyrovnání obtížnosti různých variant testu? In: Konference Psychologická diagnostika. Brno: MUNI FSS, 2021

VAN DER VLEUTEN, Cees. Automatic Item Generation by Cees van der Vleuten [online]. Maastricht University, 2019 [cit. 2021-10-4]. Available at: https://www.maastrichtuniversity.nl/news-events/newsletters/article/NyJydZFCFpcpCYHi4Fadew

KOSH, Audra E., Mary Ann SIMPSON, Lisa BICKEL, Mark KELLOGG a Ellie SANFORD-MOORE. A Cost–Benefit Analysis of Automatic Item Generation. Educational Measurement: Issues and Practice [online]. 2018, 38(1), 48-53 [cit. 2021-10-4]. ISSN 0731-1745. Available at: doi:10.1111/emip.12237

Davier, M.V. (2019). Training Optimus Prime, M.D.: Generating Medical Certification Items by Fine-Tuning OpenAI's gpt2 Transformer Model. ArXiv, abs/1908.08594.

Educational testing service. EETS Standards for Quality and Fairness [online] . Educational Testing Service, 2014. Dostupné také z <https://www.ets.org/about/fairness>.

ALDERSON, J, Caroline CLAPHAM a Dianne WALL. Language test construction and evaluation. New York, NY, USA: Cambridge University Press, 1995, 310 p. ISBN 0-521-47255-5.

KOMENDA, Martin a Andrea POKORNÁ. Benefity a úskalí elektronického testování [online]. Brno: Masarykova univerzita, 2011, dostupné také z <https://www.mefanet.cz/res/file/publikace/benefity-uskali-elektronickeho-testovani.pdf>.

TAVAKOL, Mohsen a Reg DENNICK. Post Examination Analysis of Objective Tests. 1. vydání. AMEE, 2011. AMEE guide; sv. 54. ISBN 978-1-903934-91-3.

ADESOPE, Olusola O., Dominic A. TREVISAN a Narayankripa SUNDARARAJAN. Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. Review of Educational Research. 2017, 87(3), 659-701. ISSN 0034-6543.

Yang BW, Razo J, Persky AM. Using Testing as a Learning Tool. Am J Pharm Educ. 2019 Nov;83(9):7324. doi: 10.5688/ajpe7324. PMID: 31871352; PMCID: PMC6920642.

DUBOIS, Philip H. A History of Psychological Testing. Michigan: Allyn and Bacon, 1970.

EGARTER, Saskia, Anna MUTSCHLER, Ara TEKIAN, John NORCINI a Konstantin BRASS. Medical assessment in the age of digitalisation. BMC Medical Education [online]. 2020, 20(1) [cit. 2021-11-26]. ISSN 1472-6920. Available at: doi:10.1186/s12909-020-02014-7

DENISON, Alan, Emily BATE a Jessica THOMPSON. Tablet versus paper marking in assessment: feedback matters. Perspectives on Medical Education [online]. 2016, 5(2), 108-113 [cit. 2021-11-27]. ISSN 2212-2761. Available at: doi:10.1007/s40037-016-0262-8

DENNICK, Reg, Simon WILKINSON a Nigel PURCELL. Online eAssessment: AMEE Guide No. 39. Medical Teacher [online]. 2009, 31(3), 192-206 [cit. 2021-11-27]. ISSN 0142-159X. Available at: doi:10.1080/01421590902792406

DENDIR, Seife a R. Stockton MAXWELL. Cheating in online courses: Evidence from online proctoring. Computers in Human Behavior Reports [online]. 2020, 2 [cit. 2021-11-14]. ISSN 24519588. Available at: doi:10.1016/j.chbr.2020.100033

Diedenhofen, B., Musch, J. PageFocus: Using paradata to detect and prevent cheating on online achievement tests. Behav Res 49, 1444–1459 (2017). https://doi.org/10.3758/s13428-016-0800-7

NIGAM, Aditya, Rhitvik PASRICHA, Tarishi SINGH a Prathamesh CHURI. A Systematic Review on AI-based Proctoring Systems: Past, Present and Future. Education and Information Technologies [online]. 2021, 26(5), 6421-6445 [cit. 2021-11-14]. ISSN 1360-2357. Available at: doi:10.1007/s10639-021-10597-x

SAM, Amir H., Michael D. REID a Anjali AMIN. High-stakes, remote-access, open-book examinations. Medical Education [online]. 2020, 54(8), 767-768 [cit. 2021-11-27]. ISSN 0308-0110. Available at: doi:10.1111/medu.14247

Uzbekistán vypnul internet a SMS: Pro někoho možná radikálním krokem se rozhodli zakročit v Uzbekistánu proti korupci a podvodu. [online]. 2014, 8.8.2014 [cit. 2021-11-13]. Available at: https://www.esemes.cz/magazin/uzbekistan-vypnul-internet-a-sms/

DURNING, Steven J., Ting DONG, Temple RATCLIFFE, Lambert SCHUWIRTH, Anthony R. ARTINO, John R. BOULET a Kevin EVA. Comparing Open-Book and Closed-Book Examinations. Academic Medicine [online]. 2016, 91(4), 583-599 [cit. 2021-11-13]. ISSN 1040-2446. Available at: doi:10.1097/ACM.0000000000000977

ZAGURY-ORLY, Ivry a Steven J. DURNING. Assessing open-book examination in medical education: The time is now. Medical Teacher [online]. 2021, 43(8), 972-973 [cit. 2021-11-14]. ISSN 0142-159X. Available at: doi:10.1080/0142159X.2020.1811214

JOHANNS, Beth, Amber DINKENS a Jill MOORE. A systematic review comparing open-book and closed-book examinations: Evaluating effects on development of critical thinking skills. Nurse Education in Practice [online]. 2017, 27, 89-94 [cit. 2021-11-27]. ISSN 14715953. Available at: doi:10.1016/j.nepr.2017.08.018

DOLEJŠ, Martin, Michal MIOVSKÝ a Vladimír ŘEHAN. Testová příručka ke škále osobnostních rysů představujících riziko z hlediska užívání návykových látek: (SURPS - substance use risk profile scale). 1. vydání. Praha: Klinika adiktologie, 1. lékařská fakulta Univerzity Karlovy v Praze a Všeobecná fakultní nemocnice v Praze ve vydavatelství Togga, 2012. 84 s. ISBN 978-80-87258-81-1.

BAUMGARTNEROVÁ, Gabriela a Andrea KAPUSTOVÁ. Metodický materiál pro hodnotitele písemných prací z českého jazyka a literatury. Centrum pro zjišťování výsledků vzdělávání, 2013. 37 s.

Čeština pro cizince. Pokyny k organizaci zkoušky z českého jazyka pro trvalý pobyt v ČR. 2010.

AYERS, William. To teach: The journey of a teacher. 2. vydání. New York: Teachers College Press, 2001. 151 s. s. 116. ISBN 08-077-3985-5.

DAVIDSON, Cathy N. Now you see it: How the Brain Science of Attention Will Transform the Way We Live, Work and Learn: [object Object]. 1. vydání. Viking Adult. 2011. 342 s. ISBN 9780670022823.

NORCINI, John J. Setting standards on educational tests. Medical Education [online]. 2003, 37(5), 464-469 [cit. 2021-11-18]. ISSN 0308-0110. Available at: doi:10.1046/j.1365-2923.2003.01495.x

JEŘÁBEK, Ondřej a Martin BÍLEK. Teorie a praxe tvorby didaktických testů. Olomouc: Univerzita Palackého v Olomouci, 2010. ISBN 978-80-244-2494-1.

DOWNING, Steven M a Thomas M HALADYNA. Handbook of test development. 1. vydání. Mahwah: Lawrence Erlbaum Associates, 2006. 778 s. ISBN 9780805852653.

HAMBELTON, Ronald K a Barbara S PLAKE. Using an extended Angoff procedure to set standards on complex performance assessments. Applied measurement in education. 1995, roč. 8, vol. 8, no. 1, s. 41-55, ISSN 0895-7347 (Print), 1532-4818 (Online). DOI: 10.1207/s15324818ame0801_4.

MCKINLEY, Danette W. a John J. NORCINI. How to set standards on performance-based examinations: AMEE Guide No. 85. Medical Teacher [online]. 2014, 36(2), 97-110 [cit. 2021-11-29]. ISSN 0142-159X. Available at: doi:10.3109/0142159X.2013.853119

BURR, Steven Ashley, Daniel ZAHRA, John COOKSON, Vehid Max SALIH, Elizabeth GABE-THOMAS a Iain Martin ROBINSON. Angoff anchor statements: setting a flawed gold standard? MedEdPublish [online]. 2017, 6(3) [cit. 2021-11-14]. ISSN 23127996. Available at: doi:10.15694/mep.2017.000167

The Angoff Analysis Tool: A free spreadsheet to set cutscores that are legally defensible, using the modified-Angoff method. Assessment Systems Corporation (ASC) [online]. Available at: https://assess.com/angoff-analysis-tool/

How the Angoff Analysis Tool makes it easy to set defensible cutscores. YouTube: Assessment Systems [online]. 22.3.2018 [cit. 2021-11-14]. Available at: https://www.youtube.com/watch?v=CQh6hJpDfl8

LAFAVE, M, L KATZ a D.J BUTTERWICK. Development of a content-valid standardized orthopedic assessment tool (SOAT). Advances in health sciences education: theory and practice. 2008, roč. 13, vol. 13, no. 4, s. 397-406, ISSN (Print) 1382-4996, (Online) 1573-1677. PMID: 17203268.

CANTOR, Jeffrey A. A Validation of Ebel's Method for Performance Standard Setting through its Application with Comparison Approaches to a Selected Criterion-Referenced Test. Educational and Psychological Measurement. 1989, roč. 49, vol. 49, no. 3, s. 709-721, ISSN (Print) 0013-1644; (Online) ISSN: 1552-3888.

VIOLATO, Claudio, Anthony MARINI a Curtis LEE. A validity study of expert judgement procedures for setting cutoff scores on high stakes credentialing examinations using cluster analysis. Evaluation and the Health Professions [online]. 2003, roč. 26, vol. 26, no. 1, s. 59-72, dostupné také z <http://www.internationalgme.org/Resources/Pubs/Validity%20Cutoff%20Scores%20-%20Violato.pdf>. ISSN (Print) 0163-2787; (Online) 1552-3918. PMID: 22973420.DOI: 10.1177/0163278702250082.

AZIZ, Saman. A Modified Ebel Standard Setting Method for a Medical School Clinical Skills Assessment. Chicago: University of Illinois, 2005. 162 s.

BUTTERWICK, D. J, D.M PASKEVICH a A.L VALLEVAND, et al. Development of content-valid technical skill assessment instruments for athletic taping skills. Journal of Allied Health. 2006, roč. 35, vol. 35, no. 3, s. 149-157, ISSN (Print) 0090-7421, (Online) 1945-404X. PMID: 17036669.

VIOLATO, Claudio, Lanree SALAMI a Sylvia MUIZNIEKS. Certification Examinations for Massage Therapists: A Psychometric Analysis. Journal of Manipulative Physiological Therapeutics [online]. 2002, roč. 25, vol. 25, no. 2, s. 111-115, dostupné také z <http://www.jmptonline.org/article/S0161-4754(02)70455-7/fulltext>. ISSN 0161-4754. DOI: 10.1067/mmt.2002.121413.

HOMER, Matt, Jonathan DARLING a Godfrey PELL. Psychometric characteristics of integrated multi-specialty examinations: Ebel ratings and unidimensionality. Assessment & Evaluation in Higher Education [online]. 2012, 37(7), 787-804 [cit. 2021-11-20]. ISSN 0260-2938. Available at: doi:10.1080/02602938.2011.573843

CLAUSER, Brian E., Janet MEE, Su G. BALDWIN, Melissa J. MARGOLIS a Gerard F. DILLON. Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. Journal of Educational Measurement [online]. 2009, 46(4), 390-407 [cit. 2021-11-20]. ISSN 00220655. Available at: doi:10.1111/j.1745-3984.2009.00089.x

BOURQUE, Jimmy, Haley SKINNER, Jonathan DUPRÉ, Maria BACCHUS, Martha AINSLIE, Irene MA a Gary COLE. Performance of the Ebel standard-setting method in spring 2019 Royal College of Physicians and Surgeons of Canada internal medicine certification examination consisted of multiple-choice questions. Journal of Educational Evaluation for Health Professions. 2020/04/20, 17, 12. Available at: doi:10.3352/jeehp.2020.17.12

HOMER, Matt a Jonathan C. DARLING. Setting standards in knowledge assessments: Comparing Ebel and Cohen via Rasch. Medical Teacher [online]. 2016, 38(12), 1267-1277 [cit. 2021-11-20]. ISSN 0142-159X. Available at: doi:10.1080/0142159X.2016.1230184

JØRGENSEN, Morten, Lars KONGE a Yousif SUBHI. Contrasting groups’ standard setting for consequences analysis in validity studies: reporting considerations. Advances in Simulation. 2018, 3(1), 1-7. ISSN 2059-0628. Available at: doi:10.1186/s41077-018-0064-7

COHEN-SCHOTANUS, Janke a Cees P. M. VAN DER VLEUTEN. A standard setting method with the best performing students as point of reference: Practical and affordable. Medical Teacher [online]. 2010, 32(2), 154-160 [cit. 2021-11-12]. ISSN 0142-159X. Available at: doi:10.3109/01421590903196979

NORCINI, John J. Setting standards on educational tests. Medical Education [online]. 2003, 37(5), 464-469 [cit. 2016-03-18]. ISSN 03080110. Available at: doi:10.1046/j.1365-2923.2003.01495.x

BOWERS, John J. a Russelyn Roby SHINDOLL. A Comparison of the Angoff , Beuk , and Hofstee Methods for Setting a Passing Score. ACT Research Report Series 892. 2014.

COHEN-SCHOTANUS, Janke a Cees P. M. VAN DER VLEUTEN. A standard setting method with the best performing students as point of reference: Practical and affordable. Medical Teacher [online]. 2010, 32(2), 154-160 [cit. 2021-10-13]. ISSN 0142-159X. Available at: doi:10.3109/01421590903196979

KOLEN, Michael J, Robert L BRENNAN a Michael J KOLEN. Test equating, scaling, and linking: methods and practices. 2nd ed. New York: Springer, c2004, xxvi, 548 p. ISBN 0-387-40086-9.

JELÍNEK, Martin a Petr KVĚTON. Testování v psychologii: Teorie odpovědi na položku a počítačové adaptivní testování. 1. vydání. Praha: Grada, 2011. 160 s. ISBN 978-802-4735-153.

HAMBLETON, Ronald K., Hariharan SWAMINATHAN a H. Jane ROGERS. Fundamentals of item response theory. Newbury Park, Calif.: Sage Publications, c1991. ISBN 0803936478.

A Practitioner's Introduction to Equating: With Primers on Classical Test Theory and Item Response Theory [online]. Washington: Council of Chief State School Officers, 2021 [cit. 2021-10-1]. Available at: https://ccsso.org/resource-library/practitioners-introduction-equating

Han, K. T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. Applied Psychological Measurement, 33(6), 491-493.

ALBANO, Anthony D. Equate: An R Package for Observed-Score Linking and Equating. Journal of Statistical Software [online]. 2016, 74(8) [cit. 2021-10-1]. ISSN 1548-7660. Available at: doi:10.18637/jss.v074.i08

JEŘÁBEK, Ondřej a Martin BÍLEK. Teorie a praxe tvorby didaktických testů. Olomouc: Univerzita Palackého v Olomouci, 2010. ISBN 978-80-244-2494-1.

MEZERA, Antonín. Školní měření a evaluace výsledků vzdělávání ve škole: Studijní materiál pro interní potřebu učitelů základních a středních škol [online]. [cit. 2021-11-18]. <http://www.pppraha7a8.cz/files/zaklady%20skolniho%20mereni.pdf>.

MCLACHLAN, John C a Susan C WHITEN. Marks, scores and grades: scaling and aggregating student assessment outcomes. Medical Education [online]. 2000, roč. 34, vol. 34, no. 10, s. 788-797, dostupné také z <http://doi.wiley.com/10.1046/j.1365-2923.2000.00664.x>. ISSN 0308-0110. DOI: 10.1046/j.1365-2923.2000.00664.x.

JACOBS, Lucy Cheser a Clinton I. CHASE. Developing and using tests effectively: a guide for faculty. San Francisco: Jossey-Bass Publishers, c1992. ISBN isbn-1-55542-481-3.

JAMES, Richard. A comparison of norm-referencing and criterion-referencing methods for determining student grades in higher education. JAMES, Richard, Craig MCINNIS a Marcia DELVIN. Assessing learning in Australian universities: Ideas, strategies and resources for quality in student assessment. Melbourne, Vic: Centre for the Study of Higher Education, 2002. ISBN 9780734029027.

Grading systems by country. Wikipedia [online]. 2021 [cit. 2021-11-03]. Available at: https://en.wikipedia.org/wiki/Grading_systems_by_country

THOMPSON, Nathan. What are the possible transformations for scaled scoring? Assessment Systems Corporation (ASC): Psychometrics [online]. [cit. 2021-11-03]. Available at: https://assess.com/2019/07/13/what-are-the-possible-transformations-for-scaled-scoring/

Radek Šindler, Rukověť autora testových úloh, Praha 2006, ISBN 80-239-711-5

KLINE, Paul. The Handbook of Psychological Testing. 1995

Van der Vleuten, C.P.M. 1996. The assessment of professional competence: developments, research and practical implications; Advances in health science education, 1, 41-67.

Van der Vleuten, C.P.M. and Schuwirth, L.W.T. 2005 Assessing professional competence: from methods to programmes; Medical Education, 39, 309-317.

Murphy, K. R. and C. O. Davidshofer (2005). Psychological testing: principles and applications. Upper Saddle River, N.J., Pearson/Prentice Hall. ISBN 0-13-189172-3

Standars for Educational and Psychological Testing: AERA, APA & NCME (2014). Washington: American Educational Research Association, 2014, ix, 230 s. ISBN 9780935302356.

SCHUWIRTH, Lambert W. T. a Cees P. M. VAN DER VLEUTEN. General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher [online]. 2011, 33(10), 783-797 [cit. 2021-9-26]. ISSN 0142-159X. Available at: doi:10.3109/0142159X.2011.611022

CROCKER, Linda M. a James ALGINA. Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston, c1986. ISBN 0030616344

TAVAKOL, Mohsen a Reg DENNICK. Making sense of Cronbach's alpha. International Journal of Medical Education [online]. 2011, 2, 53-55 [cit. 2021-10-30]. ISSN 20426372. Available at: doi:10.5116/ijme.4dfb.8dfd

STREINER, David L. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. Journal of Personality Assessment [online]. 2003, 80(1), 99-103 [cit. 2021-10-30]. ISSN 0022-3891. Available at: doi:10.1207/S15327752JPA8001_18

ŠTUKA, Čestmír, Patrícia MARTINKOVÁ a Karel ZVÁRA, et al. The prediction and probability for successful completion in medical study based on tests and pre-admission grades. The New Educational Review [online]. 2012, roč. -, vol. 28, no. 2, s. 138-152, dostupné také z <http://www.educationalrev.us.edu.pl/vol/tner_2_2012.pdf>. ISSN 1732-6729.

BYČKOVSKÝ, Petr a Karel ZVÁRA. Konstrukce a analýza testů pro přijímací řízení. 1. vydání. Praha: Univerzita Karlova v Praze, Pedagogická fakulta, 2007. 79 s. ISBN 978-80-7290-331-3.

ZVÁRA, Karel. Regrese. 1. vydání. Praha: MATFYZPRESS, vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze, 2008. 254 s. ISBN 978-80-7378-041-8.

BYČKOVSKÝ, Petr a Karel ZVÁRA. Konstrukce a analýza testů pro přijímací řízení. 1. vydání. Praha: Univerzita Karlova v Praze, Pedagogická fakulta, 2007. 79 s. ISBN 978-80-7290-331-3.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529–540.

Položková analýza testů studijních předpokladů jako součást zkvalitňování procesu přijímání na vysokou školu. In: MAIEROVÁ, Eva, Lenka ŠRÁMKOVÁ, Kristýna HOSÁKOVÁ, Martin DOLEJŠ a Ondřej SKOPAL. PHD EXISTENCE 2015: česko-slovenská psychologická konference (nejen) pro doktorandy a o doktorandech. Olomouc: Univerzita Palackého v Olomouci, Filozofická fakulta, 2015, s. 75-84. ISBN 978-80-244-4694-3.

SWERDLIK, Mark, Edward PROFESSOR a Ronald COHEN. Psychological Testing and Assessment: An Introduction to Tests and Measurement. - vydání. McGraw-Hill Education, 2012. 752 s. ISBN 9780078035302.

Hodnotící zpráva Matematika+ 2018: Pokusné ověřování obsahu, formy, organizace a hodnocení výběrové zkoušky ze středoškolské matematiky. CERMAT: Centrum pro zjišťování výsledků vzdělávání [online]. Praha, 2018 [cit. 2021-11-16]. Available at: https://data.cermat.cz/files/files/Matematika/MA-PLUS_hodnotici_zprava_2018.pdf

A Practitioner's Introduction to Equating: With Primers on Classical Test Theory and Item Response Theory. Washington: Council of Chief State School Officers

HAMBLETON, Ronald K., Hariharan SWAMINATHAN a H. Jane ROGERS. Fundamentals of item response theory. Newbury Park, Calif.: Sage Publications, c1991. ISBN 0803936478.

TAVAKOL, Mohsen a Reg DENNICK. Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide. Medical Teacher. 2013, 35(1), e838-e848. ISSN 0142-159x. Available at: doi:10.3109/0142159X.2012.737488

STEMLER, Steven E. a Adam NAPLES. Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. Practical Assessment, Research, and Evaluation. 26(11). ISSN 1531-7714. Available at: doi:10.7275/v2gd-4441

KEAN, Jacob, Erica F. BISSON, Darrel S. BRODKE, Joshua BIBER a Paul H. GROSS. An Introduction to Item Response Theory and Rasch Analysis: Application Using the Eating Assessment Tool (EAT-10). Brain Impairment [online]. 2018, 19(1), 91-102 [cit. 2021-11-21]. ISSN 1443-9646. Available at: doi:10.1017/BrImp.2017.31

BOONE, William J., Amity NOLTEMEYER a Gregory YATES. Rasch analysis: A primer for school psychology researchers and practitioners. Cogent Education [online]. 2017, 4(1) [cit. 2021-11-21]. ISSN 2331-186X. Available at: doi:10.1080/2331186X.2017.1416898

CÍGLER, Hynek. Jak začít s Teorií odpovědi na položku?: S pomocí knihy „Applying The Rasch Model: Fundamental Measurement in the Human Sciences“. Testfórum [online]. 2014, 2014, (3) [cit. 2021-9-28]. Available at: https://testforum.cz/article/download/TF2014-3-15/10487

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. The R Journal, 10(2), 503-515. doi: 10.32614/RJ-2018-074.

Martinková, P., Drabinová, A., & Houdek, J. (2017). ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů. TESTFÓRUM, 6(9), 16-35. doi: 10.5817/TF2017-9-129.

BRESLAU, Joshua, Kristin N. JAVARAS, Deborah BLACKER, Jane M. MURPHY a Sharon-Lise T. NORMAND. Differential Item Functioning Between Ethnic Groups in the Epidemiological Assessment of Depression. Journal of Nervous & Mental Disease [online]. 2008, 196(4), 297-306 [cit. 2021-11-11]. ISSN 0022-3018. Available at: doi:10.1097/NMD.0b013e31816a490e

CULBERTSON, John C. An Essay Review: The Bell Curve: Class Structure and the Future of America. Education Policy Analysis Archives [online]. 1995, vol. 3, no. 2, s. 1-12, dostupné také z <http://epaa.asu.edu/ojs/article/view/645/767>. ISSN 1068-2341.

MARTINKOVÁ, Patrícia, Adéla DRABINOVÁ a Jakub HOUDEK. ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů. TESTFÓRUM [online]. 2017, 6(9), 16-35 [cit. 2021-11-09]. ISSN 1805-9147. Available at: doi:10.5817/TF2017-9-129

VLČKOVÁ, Katarína. Férovost didaktických testů a jejich položek. Praha, 2014. Diplomová práce. MFF UK. Vedoucí práce Patrícia Martínková.

CÍGLER, Hynek. Férovost a zkreslení při testování: Přednáška 8. Fakulta sociálních studií MU: Katedra psychologie [online]. Brno: MUNI, 2020, 24. 11. 2020 [cit. 2021-11-06]. Available at: https://is.muni.cz/el/fss/podzim2020/PSYn4790/um/PSYn4790_2020_P08.pdf?lang=en

MARTINKOVÁ, Patrícia, Adéla DRABINOVÁ a Jakub HOUDEK. ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů. TESTFÓRUM [online]. 2017, 6(9), 16-35 [cit. 2021-11-09]. ISSN 1805-9147. Available at: doi:10.5817/TF2017-9-129

SCHUWIRTH, Lambert W. T. a Cees P. M. VAN DER VLEUTEN. General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher [online]. 2011, 33(10), 783-797 [cit. 2021-10-31]. ISSN 0142-159X. Available at: doi:10.3109/0142159X.2011.611022

Computerized Item Banking. DOWNING, Steven M. a Thomas M. HALADYNA. Handbook of test development. Mahwah, N.J.: L. Erlbaum, 2006, s. 261-286. ISBN 9780805852646.

WEISS, David J., ed. Item Banking, Test Development, and Test Delivery. GEISINGER, Kurt F. The APA Handbook on Testing and Assessment in Psychology. Washington DC: American Psychological Association, 2011. ISBN 978-1-4338-1227-9.

WEISS, David J., ed. Item Banking, Test Development, and Test Delivery. GEISINGER, Kurt F. The APA Handbook on Testing and Assessment in Psychology. Washington DC: American Psychological Association, 2011. ISBN 978-1-4338-1227-9.

VERSCHOOR, Angela a Caroline JONGKAMP. Item banking for optimal tests: AEA Europe pre-conference workshop. Prague, 2016.

Tao of Testing: instalace na 1. LF UK [online]. 2017 [cit. 2021-11-25]. Available at: https://tao.lf1.cuni.cz/

MSC Assessment Alliance. Medical Schools Council [online]. London: Medical Schools Council [cit. 2021-11-26]. Available at: https://www.medschools.ac.uk/our-work/assessment/msc-assessment-alliance

INSTITUTE FOR COMMUNICATION AND ASSESSMENT RESEARCH. Umbrella Consortium for Assessment Networks [online]. [cit. 2021-11-27]. Available at: https://www.ucan-assess.org/

HOCHLEHNERT, Achim, Konstantin BRASS, Andreas MÖLTNER, Jobst-Hendrik SCHULTZ, John NORCINI, Ara TEKIAN a Jana JÜNGER. Good exams made easy: The item management system for multiple examination formats. BMC Medical Education [online]. 2012, 12(1) [cit. 2021-11-27]. ISSN 1472-6920. Available at: doi:10.1186/1472-6920-12-63

BERNARDI, Richard A., Ania V. BACA, Kristen S. LANDERS a Michael B. WITEK. Methods of Cheating and Deterrents to Classroom Cheating: An International Study. Ethics & Behavior [online]. 2008, 18(4), 373-391 [cit. 2021-10-7]. ISSN 1050-8422. Available at: doi:10.1080/10508420701713030

CIZEK, Gregory J. a James A. WOLLACK. Handbook of Quantitative Methods for Detecting Cheating on Tests. New York and London: Routledge, 2017. ISBN 978-1-138-82180-4.

SIMS, Randi L. The Relationship Between Academic Dishonesty and Unethical Business Practices. Journal of Education for Business [online]. 2010, 68(4), 207-211 [cit. 2021-10-7]. ISSN 0883-2323. Available at: doi:10.1080/08832323.1993.10117614

Richard A. Bernardi , Ania V. Baca , Kristen S. Landers & Michael B. Witek (2008) Methods of Cheating and Deterrents to Classroom Cheating: An International Study, ETHICS & BEHAVIOR, 18:4, 373-391, DOI: 10.1080/10508420701713030

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. Ethics & Behavior, 11(3), 219–232.

Online Education Database (OEDb). 8 Astonishing Stats on Academic Cheating [online]. ©2010. Poslední revize 2010-12-19, [cit. 2012-11-25]. <http://oedb.org/library/features/8-astonishing-stats-on-academic-cheating>.

DE KLERK, Sebastiaan, Sanette VAN NOORD a Christiaan J. VAN OMMERING. The Theory and Practice of Educational Data Forensics. Theoretical and Practical Advances in Computer-based Educational Measurement. Cham: Springer International Publishing, 2019, 2019-07-06, , 381-399. Methodology of Educational Measurement and Assessment. ISBN 978-3-030-18479-7. Available at: doi:10.1007/978-3-030-18480-3_20

VRBOVÁ, Jana. „Co mi ve škole vadí víc, podvádění, či klamání?“ Postoje žáků k nečestnému chování ve škole v kontextu školního podvádění. Studia paedagogica [online]. 2013, 18(2-3) [cit. 2021-10-7]. ISSN 1803-7437. Available at: doi:10.5817/SP2013-2-3-6

KEITH-SPIEGEL, Patricia, Barbara G. TABACHNICK, Bernard E. WHITLEY JR. a Jennifer WASHBURN. Why Professors Ignore Cheating: Opinions of a National Sample of Psychology Instructors. Ethics & Behavior [online]. 1998, 8(3), 215-227 [cit. 2021-10-7]. ISSN 1050-8422. Available at: doi:10.1207/s15327019eb0803_3

FOSTER, David. Security Issues in Technology-Based Testing. WOLLACK, James A. a James J. FREMER. Handbook of Test Security. London: Routledge, 2013, s. 39-83. ISBN 978-0415805643.

CIZEK, Gregory J. a James A. WOLLACK. Handbook of Quantitative Methods for Detecting Cheating on Tests. New York and London: Routledge, 2017. ISBN 978-1-138-82180-4.

DE KLERK, Sebastiaan, Sanette VAN NOORD a Christiaan J. VAN OMMERING. The Theory and Practice of Educational Data Forensics. Theoretical and Practical Advances in Computer-based Educational Measurement. Cham: Springer International Publishing, 2019, 2019-07-06, , 381-399. Methodology of Educational Measurement and Assessment. ISBN 978-3-030-18479-7. Available at: doi:10.1007/978-3-030-18480-3_20

VRBOVÁ, Jana. „Co mi ve škole vadí víc, podvádění, či klamání?“ Postoje žáků k nečestnému chování ve škole v kontextu školního podvádění. Studia paedagogica [online]. 2013, 18(2-3) [cit. 2021-10-7]. ISSN 1803-7437. Available at: doi:10.5817/SP2013-2-3-6

MARTINKOVÁ, Patricia, Lubomír ŠTĚPÁNEK, Adéla DRABINOVÁ, Jakub HOUDEK, Martin VEJRAŽKA a Čestmír ŠTUKA. Semi-real-time analyses of item characteristics for medical school admission tests. In: . 2017-9-24, s. 189-194. Available at: doi:10.15439/2017F380

Cizek, Gregory J. and James A. Wollack , "Handbook of Quantitative Methods for Detecting Cheating on Tests" (Abingdon: Routledge, 2016).

Maynes, D.; Educator cheating and the statistical detection of group-based test security threats. In WOLLACK, James A. a John J. FERMER. (Eds.), Handbook of test security (pp. 187–214). New York, Routledge, Psychology Press, 2013. ISBN 978-0-203-66480-3.

Ranger, J., Schmidt, N., & Wolgast, A. (2020). The Detection of Cheating on E-Exams in Higher Education-The Performance of Several Old and Some New Indicators. Frontiers in psychology, 11, 568825. https://doi.org/10.3389/fpsyg.2020.568825

Kamalov F, Sulieman H, Santandreu Calonge D (2021) Machine learning based approach to exam cheating detection. PLoS ONE 16(8): e0254340. https://doi.org/10.1371/journal.pone.0254340

TENDEIRO, Jorge N., Rob R. MEIJER a A. Susan M. NIESSEN. PerFit: An R Package for Person-Fit Analysis in IRT. Journal of Statistical Software [online]. 2016, 74(5), 1-27 [cit. 2021-10-7]. ISSN 1548-7660. Available at: doi:10.18637/jss.v074.i05

THOMPSON, Nathan. SIFT: A new tool for statistical detection of test fraud: SIFT: Software for Investigating Test Fraud. Assessment Systems Corporation (ASC) [online]. 2016 [cit. 2021-11-16]. Available at: https://assess.com/sift-new-tool-statistical-detection-test-fraud/

WOLLACK, James A. A Nominal Response Model Approach for Detecting Answer Copying. Applied Psychological Measurement [online]. 1997, 21(4), 307-320 [cit. 2021-10-6]. ISSN 0146-6216. Available at: doi:10.1177/01466216970214002

VAN DER LINDEN, Wim J. a Leonardo SOTARIDONA. Detecting Answer Copying When the Regular Response Process Follows a Known Response Model. Journal of Educational and Behavioral Statistics [online]. 2006, 31(3), 283-304 [cit. 2021-10-6]. ISSN 1076-9986. Available at: doi:10.3102/10769986031003283

SOTARIDONA, Leonardo S. a Rob R. MEIJER. Two New Statistics to Detect Answer Copying. Journal of Educational Measurement [online]. 2003, 40(1), 53-69 [cit. 2021-10-6]. ISSN 0022-0655. Available at: doi:10.1111/j.1745-3984.2003.tb01096.x

FSBPT. Forensic Analysis Conducted to Investigate Effect of Trafficking in Recalled Test Items Leads to Invalidation of 20 Candidate Test Scores [online]. The Federation of State Boards of Physical Therapy, ©2012. [cit. 2012-12-13]. <https://www.fsbpt.org/forfaculty/yourquestions/index.asp#InvalidatedNPTEScores>.

ČULÍK, Jan. Jak se prodávají zkoušková zadání na právnické fakultě. [online]. In: . Praha, 1999, 14.7. 1999 [cit. 2021-10-22]. Available at: http://www.ceskaskola.cz/1999/07/jan-culik-jak-se-prodavaji-zkouskova.html

Výroční zpráva o činnosti za rok 2005: Univerzita Karlova v Praze, Právnická fakulta [online]. Praha, 2005 [cit. 2021-10-23]. Available at: https://www.prf.cuni.cz/dokumenty-download/1404044551/

WINTER, Tom. College cheating ringleader says he helped more than 750 families with admissions scheme. NBC NEWS [online]. 13.3.2019 [cit. 2021-11-04]. Available at: https://www.nbcnews.com/news/us-news/college-cheating-mastermind-says-he-helped-nearly-800-families-admissions-n982666

BAKER, Vicky. Celebrity parents and the bizarre 'cheating' scandal: US college admissions scandal. BBC News [online]. Washington DC: BBC, 2019, 15.3.2019 [cit. 2021-11-04]. Available at: https://www.bbc.com/news/world-us-canada-47585336

CRESSEY, Donald Ray. Other people's money: A study of the social psychology of embezzlement. 1953, 191 s.

SIMMONS, Andrew. Why Students Cheat—and What to Do About It: A teacher seeks answers from researchers and psychologists. GEORGE LUCAS EDUCATIONAL FOUNDATION: CLASSROOM MANAGEMENT [online]. 2018, 2018 [cit. 2021-11-16]. Available at: https://www.edutopia.org/article/why-students-cheat-and-what-do-about-it

FOLTÝNEK, Tomáš. Akademická integrita a jak ji utvářet: European Network for Academic Integrity. In: Akademická integrita [online]. Slovenská akreditačná agentůra pre vysoké školstvo, 2021, 2021 [cit. 2021-11-17]. Available at: https://saavs.sk/wp-content/uploads/2021/06/Akademicka-integrita_Foltynnek.pdf

https://theconversation.com/motivation-is-a-key-factor-in-whether-students-cheat-155274

MCCABE, Donald L., Kenneth D. BUTTERFIELD a Linda K. TREVIÑO. Cheating in college. The Johns Hopkins University Press, 2012. ISBN 9781421407166.

MA, Yuchao, Donald L. MCCABE a Ruizhi LIU. Students’ Academic Cheating in Chinese Universities: Prevalence, Influencing Factors, and Proposed Action. Journal of Academic Ethics [online]. 2013, 11(3), 169-184 [cit. 2021-11-03]. ISSN 1570-1727. Available at: doi:10.1007/s10805-013-9186-7

VOWELL, Paul R. a Jieming CHEN. Predicting Academic Misconduct: A Comparative Test of Four Sociological Explanations. Sociological Inquiry [online]. 2004, 74(2), 226-249 [cit. 2021-11-03]. ISSN 0038-0245. Available at: doi:10.1111/j.1475-682X.2004.00088.x

DYER, Jarret, Heidi PETTYJOHN a Steve SALADIN. Academic Dishonesty and Testing: How Student Beliefs and Test Settings Impact Decisions to Cheat:

How Student Beliefs and Test Settings Impact Decisions to Cheat. 2020/04/28.

BURNETT, Audrey J., Theresa M. ENYEART SMITH a Maria T. WESSEL. Use of the Social Cognitive Theory to Frame University Students’ Perceptions of Cheating. Journal of Academic Ethics [online]. 2016, 14(1), 49-69 [cit. 2021-11-03]. ISSN 1570-1727. Available at: doi:10.1007/s10805-015-9252-4

KENNEDY, Robert. Why Students Cheat and How to Stop It. ThoughtCo: World Largest Education Resource [online]. Dotdash, 16.11.2019 [cit. 2021-11-16]. Available at: https://www.thoughtco.com/cheating-basics-for-private-schools-2773348

Why Do Students Cheat? UNT Teaching Commons: Center for Learning Experimentation, Application, and Research [online]. University of North Texas [cit. 2021-11-16]. Available at: https://teachingcommons.unt.edu/teaching-essentials/academic-integrity/why-do-students-cheat

WILSON, Scott. Rogō: an open source solution for high-stakes assessment [online]. OSS Watch team blog: open source software advisory service, ©2012. [cit. 2012-12-09]. <http://osswatch.jiscinvolve.org/wp/2012/09/13/rogo-an-open-source-solution-for-high-stakes-assessment/>.

BAYLEM, N.J, S WILKINSON a R DENNICK. Would the MRCS Written Papers Benefit from Computerisation? The University of Nottingham Experience. Bulletin of The Royal College of Surgeons of England. 2011, roč. -, vol. 93, no. 1, s. 1-5, ISSN 14736357. DOI: 10.1308/147363511X546545.

KORVINY, Petr, Roman FOLTYN a Robert KEMPNÝ. LMS Moodle na více serverech [online] . 1. vydání. 2009. 443 s. s. 239-244. Dostupné také z <http://korviny.cz/clanky_pdf/smm2009-korviny_foltyn_kempny-clanek.pdf>. Proceedings of International Conferences: ICT Bridges, Sunflower 2009, Silesian Moodle Moot 2009. ISBN 978-80-248-2117-7

KORVINY, Petr a Roman FOLTYN. LMS Moodle v clusteru. In EUNIS-CZ. Open Source na vysokých školách: sborník příspěvků ke konferenci: Špindlerův Mlýn 23.-25.9.2012. 1. vydání. Západočeská univerzita, 2012. 71 s. ISBN 8026101499, 9788026101499

BOUSSAKUK, Mohammed, Ahmed BOUCHBOUA, Mohammed EL GHAZI, Moulhime EL BEKKALI a Mohammed FATTAH. Designing and Developing e-Assessment Delivery System Under IMS QTI ver.2.2 Specification. International Journal of Emerging Technologies in Learning (iJET) [online]. 2021, 16(01), 219-233 [cit. 2021-11-27]. ISSN 1863-0383. Available at: doi:10.3991/ijet.v16i01.16257

Igniting Digital Assessment Innovation. IMS Global: Learning Consortium [online]. [cit. 2021-11-27]. Available at: http://www.imsglobal.org/activity/qtiapip

Psychometric software. Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2021-11-28]. Available at: https://en.wikipedia.org/wiki/Psychometric_software

NELSON, Larry Richard. Item analysis software for classes: Measurement Classes with Lertap 5, jMetrik, SAS University, BILOG-MG, and Xcalibre [online]. In: . Curtin University, 2017 [cit. 2021-11-28]. Available at: doi:10.13140/RG.2.2.32532.71049

MARTINKOVÁ, Patrícia, Adéla HLADKÁ a Jan NETÍK. Psychometrická analýza interaktivně a v R: Co je nového v ShinyItemAnalysis. In: Konference Psychologická diagnostika. Brno, 2021.

BYČKOVSKÝ, Petr a Marie MARKOVÁ. Využití software ITEMAN k položkové analýze a analýze výsledků testů. In: 11. konference ČAPV - Sociální a kulturní souvislosti výchovy a vzdělávání.: Sborník referátů. [online]. Pedagogická fakulta, Masarykova Univerzita, 2003 [cit. 2021-11-28]. Available at: http://www.ped.muni.cz/capv11/5sekce/5_CAPV_Byckovsky.pdf

VON DER EMBSE, Nathaniel, Dane JESTER, Devlina ROY a James POST. Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. Journal of Affective Disorders [online]. 2018, 227, 483-493 [cit. 2021-11-16]. ISSN 01650327. Available at: doi:10.1016/j.jad.2017.11.048.

Yerkes RM, Dodson JD (1908). "The relation of streng th of stimulus to rapidity of habit-formation". Journal of Comparative Neurology and Psychology 18: 459–482. doi:10.1002/cne.920180503.

Nakonečný, Milan. 1992. Motivace pracovního jednání a její řízení. Praha: Management Press.

Lupien SJ, Maheu F, Tu M, Fiocco A, Schramek TE (2007). "The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition". Brain and Cognition. 65 (3): 209–237. CiteSeerX 10.1.1.459.1378. doi:10.1016/j.bandc.2007.02.007. PMID 17466428. S2CID 5778988

Hembree, Ray. “Correlates, Causes, Effects, and Treatment of Test Anxiety.” Review of Educational Research 58, no. 1 (1988): 47–77. https://doi.org/10.2307/1170348.

Andrews, B. and Wilding, J.M. (2004), The relation of depression and anxiety to life-stress and achievement in students. British Journal of Psychology, 95: 509-521. https://doi.org/10.1348/0007126042369802

https://doi.org/10.1111/j.1467-1770.1982.tb00522.x

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Flink.springer.com%2Fcontent%2Fpdf%2F10.3758%252FBF03330210.pdf&clen=867852&chunk=true

WEEMS, Carl F., Brandon G. SCOTT, Leslie K. TAYLOR, Melinda F. CANNON, Dawn M. ROMANO, Andre M. PERRY a Vera TRIPLETT. Test Anxiety Prevention and Intervention Programs in Schools: Program Development and Rationale. School Mental Health [online]. 2010, 2(2), 62-71 [cit. 2021-11-28]. ISSN 1866-2625. Available at: doi:10.1007/s12310-010-9032-7

https://www.researchgate.net/publication/274734273

DOI: 10.1080/87567555.1992.10532238

RUDNER, Lawrence M.. Implementing the graduate management admission test computerised adaptive test [online] . In van der Linden, Wim J.; Glas, Cees A.W. Elements of Adaptive Testing. 1. vydání. New York: Springer, 2010. s. 151-165. Dostupné také z <https://link.springer.com/chapter/10.1007/978-0-387-85461-8_8>. ISBN (Print) 978-0-387-85459-5, (Online) 978-0-387-85461-8

BREITHAUPT, Krista, Adelaide A ARIEL a Donovan R HARE. Assembling an inventory of multistage adaptive testing systems [online] . In van der Linden, Wim J.; Glas, Cees A.W. Elements of Adaptive Testing. 1. vydání. New York: Springer, 2010. s. 247-266. Dostupné také z <http://link.springer.com/chapter/10.1007%2F978-0-387-85461-8_13>. DOI:10.1007/978-0-387-85461-8_13. ISBN (Print) 978-0-387-85459-5, (Online) 978-0-387-85461-8

GIERL, Mark J, Hollis LAI and Simon R TURNER. Using automatic item generation to create multiple-choice test items. Medical Education [online]. 2012, roč. -, vol. 46, no. 8, pp. 757-765, dostupné také z <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2012.04289.x/full>. sv. DOI:. ISSN 1365-2923. PMID: 22803753.DOI: 10.1111/j.1365-2923.2012.04289.x.

PICUS, Lawrence O., Alisha TRALLI a Suzanne TACHENY. Estimating the Costs of Student Assessment in North Carolina and Kentucky: A State-Level Analysis. In: CSE Report 408: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). [online]. Los Angeles: University of California, 1996 [cit. 2021-11-17]. Available at: https://cresst.org/publications/cresst-publication-2780/

PHELPS, Richard. Estimating the Costs and Benefits of Educational Testing Programs. In: The Education Consumers Consultants Network: Issues in Public Education: Research and Analysis from the Education Consumers Foundation [online]. Arlington: Education Consumers Foundation, 2002 [cit. 2021-11-18]. Available at: http://education-consumers.org/research/briefs_0202.htm

13 Index

14 Afterword

Today, testing is an obvious part of higher education. The authors have attempted to provide an initial orientation for teachers interested in this area. The aim was more the popularization of procedures and methods than a detailed investigation into these procedures and methods.

Here, we would like to thank everyone who helped us orient ourselves in the subjects, those who encouraged and inspired us.

We believe that every interest stimulates curiosity in people, and we would be happy if you would seek answers that go beyond this text. We wish everyone who goes this route to have as much fun as we do.

And to conclude, a favorite memento:

Test results are to the assessment of a student what lab results are to a diagnosis.

Category: Book

The project “Zvýšení kvality vzdělávání na UK a jeho relevance pro potřeby trhu práce” (reg. no. CZ.02.2.69/0.0/0.0/16_015/0002362) is supported with funds from the Operational Programme Research, Development and Education (OP RDE).

Personal data protection About StaTest Disclaimer of Liability Developers Statistics Cookies

```
Statest is a Mefanet project.
```

288,122

13,000

2,500