Course Book

INTERNATIONAL ASSESSMENT METHODS

DLMIOPIAM01

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

LEARNING OBJECTIVES

With an increasingly globalized work force and the resulting larger pool of able job applicants, it has become imperative to be able to select the most competent person for a vacant position. As a result, it is essential that decision-makers responsible for filling positions and assigning roles are capable of utilizing psychological assessment tools to their full functionality. Beyond this, maintenance of employee well-being is also taking on a more critical role, and similarly, appropriate understanding of relevant diagnostic tools becomes necessary.

The course book **International Assessment Methods**, therefore, aims to familiarize you with a range of diagnostic tools used on an international level, principally drawing on standardized tests from the UK and the US. Particular attention is given to tests of aptitude, intelligence, and personality as well as tests used to assess employee well-being. Beyond engaging with underpinning theory, you will learn how to carry out such standardized assessments and to analyze and interpret them for appropriate occupational decision-making.

UNIT 1

THE DIAGNOSTIC PROCESS

STUDY GOALS

On completion of this unit, you will be able to ...

- understand the diagnostic process.
- describe which data sources can be used to obtain information in the diagnostic process.
- recognize which standards psychological test diagnostics should meet.
- identify the main and secondary quality criteria of psychological test procedures.

1. THE DIAGNOSTIC PROCESS

Case Study

Kim ~~T.~~, junior consultant in a management consultancy firm that specializes in the field of human resources, welcomes her first client. He is a member of the management board of a company that manufactures automobile engines. The customer would like to change the company's application process and use psychological tests in addition to selection interviews. He turned to the management consultancy where Kim works to establish such procedures and says, "I've read that there are psychological tests for all sorts of issues, and I think we should just do as many of them as possible as part of our application process."

Kim explains to her customer: "It's not that easy. There are determining factors that should be adhered to. Also, psychological testing shouldn't be used just because we can. We always need a specific question, a ~~so-called~~ hypothesis, to be answered with the help of psychological tests. For this, it makes sense to carry out a corresponding requirements analysis. Then we can assess what exactly we want to record with the psychological test procedures. It should also be noted that even with the application and use of psychological test procedures, there is no guarantee that we will not make mistakes. That's why it is important not to base our decisions on a single test procedure but to include various valid and objective procedures in our process."

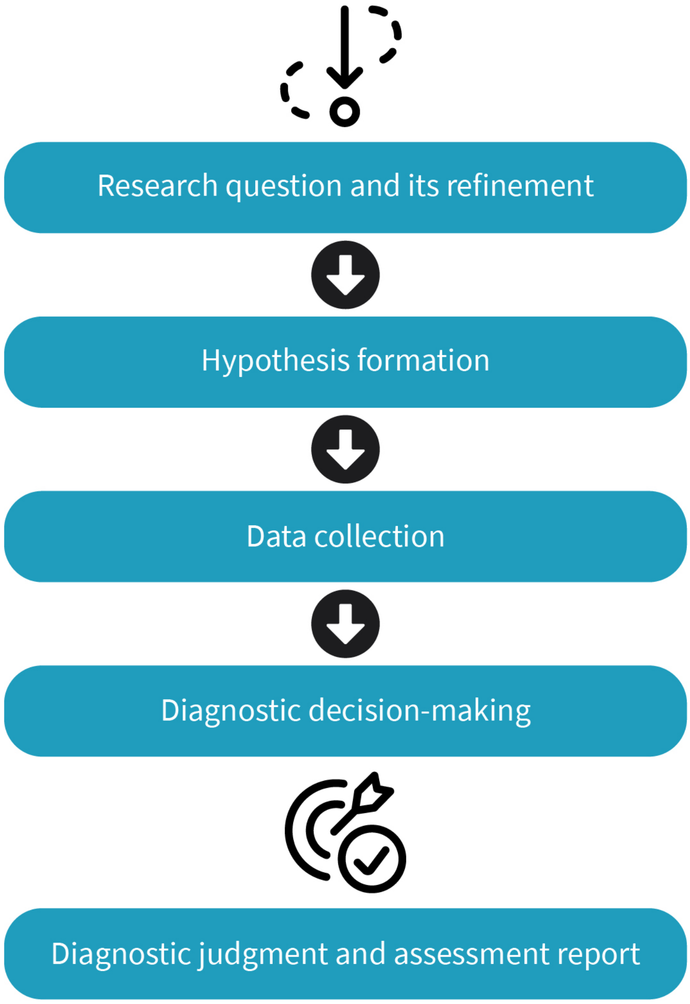
In recent years, the term "psychological testing" has been broadened and developed into psychological assessment. Since psychologists not only use tests for data collection but conduct (semi-structured) interviews and behavioral observations, the "term assessment implies that there are many ways of evaluating individual differences" (Goldstein et al., 2019, p. 4).

1.1 Foundations and Framework Conditions of Psychological Diagnostics

"Possibly the greatest single achievement of the members of the American Psychological Association is the establishment of the psychology of individual differences" (Scott, 1920, p. 85). Even though this argument can be debated, it showcases the profound impact of the assessment of individual differences in the field of psychology. Psychological assessment can be used in many different psychological fields, for example, in clinical psychology, educational psychology, health psychology, and work and organizational psychology. In work and organizational psychology, psychological assessment is used in personnel selection and development as well as career counseling. These different areas of application also imply many different questions psychological assessment has to answer. Murphy (2012) points out that "People differ in many ways. Some of these differences are fleeting (moods), and others are long-lasting and important to some individuals but are not likely to be critical for understanding their behavior in organizations or their effectiveness in particular jobs (e.g., preferences for music)" (p. 31). Work and organizational psychology "focuses on relatively stable individual differences that are relevant for understanding behavior and effectiveness in work organizations" (Murphy, 2012, p. 31). First of all, it is crucial that psychological assessment is always based on a question that is formulated by a client and is, therefore, not causeless. In addition to the assessment of individual characteristics, these questions can also relate to situational patterns of experience and behavior of individuals or a group of people as well as to the context in which they operate. For this purpose, not all possible information is collected at random but specifically that which is relevant to answering the question. These are then interpreted – again with regard to the question. The procedure of data collection needs to conform to scientific standards, and the assessment and interpretation has to be based on psychological expertise (Wright, 2020).

The diagnostic process begins with question from a potential client. This can already be a very specific question (e.g., "For which in-service training is Max suitable in terms of his skills and interests?"), but it can also be a very global question (e.g., "Which applicant should we hire?").

Figure 1: The Diagnostic Process


Research question and its refinement
Hypothesis formation
Data collection
Diagnostic decision-making
Diagnostic judgment and assessment report

Source: Created on behalf of IU (2023), based on Jäger (2006, p. 91–94).

Often, it is necessary to specify the corresponding question in order to process it further, for example, from the question: “Which applicant should we hire?” to the more precise question: “Which applicant is best suited for the advertised position in terms of the characteristics relevant to career success?”

Psychological Hypothesis Testing

Hypothesis
A hypothesis is a theoretically-based assumption that has not yet been tested.

This question or its corresponding specification is usually very complex and cannot be answered without further information gathering. It is, therefore, translated into a psychological **hypothesis** (i.e., an assumption that is then confirmed or rejected in the course of the process and, thus, provides the answer to the underlying question; Wright, 2020). In order to obtain information (diagnostic data), assignments and questions must first be operationalized so that it can be determined which procedures can be used to obtain the data necessary. For example, certain characteristics that are relevant to professional success can be recorded via the result achieved in tests (e.g., that of intelligence in an intelligence test or the characteristic of conscientiousness via the answers given in a personality questionnaire).

All data collected during the diagnostic process is ultimately combined into an overall judgment as part of the diagnostic evaluation. At the end of the diagnostic process, there is a diagnostic judgment that answers the initial question as best as possible. This is then provided to the client in the form of an oral or written report (Wright, 2020).

That being said, the ethical and legal ramifications of assessments must be considered by those who attempt to evaluate individuals for descriptive and predictive purposes. Those who work in the field of Psychology must adhere to the guidelines within their specific jurisdictions, for example, the American Psychological Association in the United States and the British Psychological Society within the United Kingdom. Each respective regulating body has its set of ethical guidelines which lay out the expectation for those assessments (ethical principles of psychologists and code of conduct). Dos Santos et al. (2017) emphasize the importance of ethics in the field of employee selection since “recruitment and selection practices often cause first impressions to be formed, those practices have an impact on employees’ behaviors beyond the time of recruitment and selection [and even] those who are excluded (i.e., not hired) may also be customers and bring to the market the impression they have formed about the organization during the recruitment and selection processes” (p. 92).

When carrying out psychological diagnostics, it is also important to adhere to the relevant legal framework. There are several pieces of legislation for diagnosticians depending on the country you work and operate in. The EU Charter of Fundamental Rights specifies its regulations of protection of personal data through Article 8, stating that:

1. Everyone has the right to the protection of personal data concerning them.
2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning them, and the right to have it rectified.
3. Compliance with these rules shall be subject to control by an independent authority. (European Union, 2012, Article 8)

This is of particular importance insofar as psychological diagnostics could in principle be used to collect information about a person who does not wish to disclose it. For example, as part of a behavioral observation, video recordings could be made without the knowledge and consent of the person (Schmidt-Atzert & Amelang, 2012).

1.2 Data Sources and Methods of Data Collection

As Goldstein et al. (2019) point out: “In recent years, there has been a distinction made between testing and assessment, assessment being the broader concept. Psychologists do not only give tests now; they perform assessments” (p. 4). When considering evaluating individual differences, multiple data sources and methods of data collection are available. The most common way of assessment remains using tests, most often questionnaires. But there are also other ways of gathering information about individuals, such as interviewing, behavioral observations in natural or structured settings (e.g., role playing), or the recording of psychological functioning (e.g., Electroencephalography [EEG] when testing for neurological impairments) used in clinical settings.

The list of possible data sources can be long and needs to correspond to the hypothesis being tested. Potential data sources and methods of data collection can be:

- documents and factual analysis (e.g., the analysis and interpretation of school, university, and work references, curriculum vitae [CVs], etc.)
- interviews both with the applicant/individual, agent, or past employer and colleagues
- Behavioral observation and assessment (e.g., role play, group discussions, presentation exercises, etc.)
- questionnaires (e.g., personality questionnaire)
- tests (e.g., intelligence tests, knowledge tests, written reports, or essays)

Further distinctions are possible, for example, with regard to what is recorded by the various methods. Characteristics such as (personality) traits, experiences, and emotions, as well as attitudes and cognitions, i.e., mental processes, can be considered. However, it is also possible to observe situations and behavior across different dimensions (e.g., the current workplace situation or the behavior shown by the people being tested). In practice, different characteristics are often recorded in one test procedure. So, in a questionnaire designed to examine the conscientiousness of individuals, questions about behavior (“I always come to work on time”) and characteristics (“I am very conscientious”) are combined. Further distinctions are offered by the temporal orientation, whether the perspective of the examined characteristic is directed to the past, present, or future.

Of particular interest in the context of assessment are the theory-based psychometric questionnaires and tests that belong to the psychological test procedures and are often simply called “tests.” Schmidt-Atzert and Amelang (2012) summarize that psychological test procedures are measurement methods with which one or several psychological characteristics are to be recorded. The procedure is standardized and includes the collection of a behavioral sample. The behavior is caused by the specific conditions realized in the test. Its variation is said to be largely due to the variation of the characteristic being measured. The goal is a quantitative and/or a qualitative statement about the characteristic.

Goldstein et al. (2019) emphasize that “testing is now in the computer age” (p. 3) with adaptive testing and assessment through virtual reality applications leading to the significance of psychometrics and statistical sciences. This change from the formerly used paper-pencil tests towards a technology-based assessment offers many advantages, such as the increasingly automated and objectified administration, scoring, and interpretation of tests. The authors expedite that testing and assessment have become a matter of cause in many fields such as clinical, education, and work settings.

1.3 Principles of Multimodal Diagnostics

Multimodal diagnostics is a combination of different diagnostic methods to answer a diagnostic question. This can be useful to secure findings, to improve predictions, to minimize sources of error, and to be able to compare different perceptions (e.g., self-perception/perception by others; Goldstein et al., 2019; Schmitt, 2012). In concrete terms, this means not only using biographical data such as CVs and certificates in personnel selection and development but also integrating various methods, e.g., a behavioral test or a psychological test procedure, in order to be able to make reliable statements or to improve the predictive power of the diagnostic judgment.

Various diagnostic dimensions can be taken into account: the survey dimensions (also data level), the data source, the observer perspectives, target or functional areas, and examination methods.

Table 1: Diagnostic Dimensions

Dimension	Description
Data gathering	Biological, psychological, social
Data source	The subject, test administrator, institution, external parties
Perspectives of observation	Self assessment, external assessment
Objectives and functional areas	Organismic functions, cognitive functions, behavior, perception, social interaction, life quality
Assessment methods	Tests, questionnaires, self assessment, external ratings

Source: Created on behalf of IU (2023), based on Mühlig & Petermann (2006, p. 100).

Three dimensions of selection tools have been established in the human resources area: biography-, test-, and simulation-oriented procedures. In the case of biography-oriented procedures, conclusions are drawn about future career successes from previous experiences. Using test-oriented methods, conclusions are drawn about future career success based on recorded current success-relevant characteristics or properties (e.g., personality traits). In contrast, in the simulation-oriented methods, in which potential future behavior is recorded (e.g., through role play), future professional success is inferred.

If procedures from all three survey dimensions are used, one can speak of a **multimodal** procedure (Schmitt, 2012).

Multimodality
This is a collection of diverse information on different levels.

1.4 Occupational Aptitude Diagnostics

Industrial and organizational psychology
The specialty of industrial-organizational psychology (also called I-O psychology) is characterized by the scientific study of human behavior in organizations and the workplace with a strong focus on personnel development, assessment, and selection.

Schmitt (2012) emphasizes that "employee selection has played a central role in **I-O psychology**; it has arguably been the dominant activity for I-O psychologists throughout the history of the field" (p. 22). As organizations developed in size and complexity, there was a clear need for a methodical approach to choosing eligible candidates. Psychologists' capacity to scientifically demonstrate the value of their work was crucial to their success, and this was made possible by the ongoing development of statistical tools that matched improvements in fundamental measurement and assessment technique. Due to this, early industrial psychology was able to set itself apart from pseudoscientific methods to some extent and establish a niche for the young discipline (Schmitt, 2012). The chosen predictors and criteria have proven to be remarkably robust. Interviews, biographical information, job ability as well as cognitive ability tests, personality tests, and situational tests are common and remain well-liked today. Although it is true that these predictors and our knowledge of them have greatly advanced and that the methods of administration have increased (e. g., computer administration), it is interesting to note that much of the focus has been on refining current procedures rather than creating entirely new categories of predictors. Although psychologists have become more refined in measuring characteristics, the criteria used today have not changed significantly from the previous process. While it is without question that great strides have been made in validation research over the years, there has been an increased emphasis on developing valid theories of job performance, criteria, and the selection process, rather than relying on brute-force empiricism (Schmitt, 2012).

International Recruitment, Ethical, and Legal Considerations

Phillips and Gully (2017) argue that "because it influences the number and types of applicants ultimately available for hire, global recruiting is critical to global talent management and strategic human resource management" (p. 29). As the business world becomes more and more international, many issues regarding employee recruitment need consideration. For example, Ryan and Delany (2010) discuss how wording in job advertisements can suggest preferential treatment for certain groups. While this may lead to a lawsuit in the US, the same preferential treatment may be commonly accepted and, in some cases, even be legally mandated in other countries.

Regarding the critical constrains on personnel selection in the US, Gutman (2012) reviews Equal Employment Opportunity (EEO) case law rulings on hiring, promotion, and termination. He emphasizes that "EEO laws are complex, even to the trained lawyer or practitioner" (2012, p. 686).

In Germany, certain quality standards have been defined on the basis of a DIN 33430 for professional aptitude diagnostics. This was first published in 2002 and has been available in its current version since 2016 (Berufsverband Deutscher Psychologinnen und Psychologen, n.d.). DIN 33430 is a service standard with the aim of recording quality features for aptitude diagnostic procedures. It places demands on the diagnostic process, the test methods used, and the qualifications of the people involved in the process, such as the diagnostician, which are regarded as a prerequisite for quality. However, the standard is not legally binding. These national differences must always be considered in the diagnostic process and the selection of personnel.

Steiner (2012) points out that “despite numerous advances increasing our knowledge of applications of selection research throughout the world, many questions remain unanswered because multinational companies have typically applied North American and European approaches in standardizing their worldwide selection strategies” (p. 741). He also emphasizes that considering cultural factors and differences in common selection practices can be beneficial not only to research in the field but also to its application in personnel selection. In doing so, issues such as adaption of selection instruments to other languages and cultural contexts need to be considered (Steiner, 2012).

This being said, professional, ethical, and legal guidelines should be adhered to when assessing personnel. A collection of guidelines which should be considered were summarized by Bartram and Tippins (2017):

Table 2: Collection of Guidelines

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014)	Standards for educational and psychological testing
Society for Industrial and Organizational Psychology (SIOP) (2003)	Principles for the use and validation of personnel selection procedures
European Federation of Psychologists’ Associations (2013)	EFPA review model for the description and evaluation of psychological and educational tests version 4.2.6
International Organization for Standardization (ISO) (2011)	ISO-10667-2 assessment service delivery – Procedures and methods to assess people in work and organizational settings
International Test Commission (2001)	International guidelines for test use
International Test Commission (2005)	International guidelines on test adaptation
Equal Employment Opportunity Commission (EEOC) (1978)	Uniform guidelines on employee selection standards
American Psychological Association (2010)	APA ethical principles of psychologists and code of donduct
International Task Force on Assessment Center Guidelines (2009)	Guidelines and ethical considerations for assessment center operations

Source: Created on behalf of IU (2023), based on Bartram and Tippins (2017, p. 280).

Regardless of national specifics, some procedures should apply universally. The required characteristics needed by the applicant should be determined in advance via a requirements analysis. At the beginning of the aptitude assessment process, it is, therefore, necessary to determine which characteristics are to be considered and the extent to which a person is considered suitable. These characteristics can be traits, skills, needs, or other psychological traits (see also Breaugh, 2017).

Krause (2017) makes a distinction between three levels of characteristics: the basic job specifications, the personal specification, and the general occupational success-relevant characteristics. The job specification level focuses on the basic requirements that the job places on people. In order to exercise these successfully, a person needs the appropriate skills, abilities, and knowledge. The personal specification level considers a person's interests, needs, and values in order to determine how these come into play in the specific workplace, i.e., the satisfaction potential of the work activity. Under characteristics that are generally relevant to professional success, the future potential of a person is considered on the one hand – also in terms of how an adjustment to future changes in work requirements can be managed – and, on the other hand, characteristics such as conscientiousness and intelligence, which are generally said to be related to professional success, are examined. It must also be considered which development and change potential this work activity has (Krause, 2017; Schmitt, 2012). Ployhart and Schneider (2012) introduce the classical personnel selection model with the main goal being the comprehensive definition of a job and the following "identification of the most critical aspects of performance on the job" (p. 49).

The aptitude diagnostics are dedicated to two core tasks: It supports the selection of suitable employees and accompanies change processes (modification), e.g., in the area of personnel development. The former is referred to as selection diagnostics and the latter as modification diagnostics (Schmitt, 2012; Schmidt-Atzert & Amelang, 2012).

Analysis of Requirements and Profile of Requirements

In order to determine which characteristics are relevant to professional success within the framework of professional aptitude diagnostics, a requirements analysis can first be carried out in order to determine a requirements profile. A requirements analysis is a form of work analysis in which personal characteristics are determined that are necessary for the successful completion of an activity (Krause, 2017; Schmitt, 2012). There are three methodological approaches to creating a requirements analysis (Nerdinger et al., 2019): the experience-based intuitive method, the empirical workplace-analysis method, and the personal-empirical method. In practice, the use of several methods is recommended:

- In the experience-based intuitive method, experts take a holistic view of the activity and use this to estimate which corresponding suitability characteristics must be present.
- The empirical workplace-analysis method is characterized by the fact that the targeted activity is recorded using social-scientific research methods such as questionnaires and interviews.
- In the personal-empirical method, the employees who are currently above-average and below-average in their job are examined in a group comparison to determine which characteristics differ between them.

When carrying out the requirements analysis, three levels can be distinguished: “tasks and results,” the “behavior,” and the “characteristics.” At the description level of the tasks and results, information about task-specifics and requirements are grouped into task inventories. Experts, for example current job holders and supervisors, evaluate these activities with regard to various criteria, such as their importance for the overall performance, their complexity, and their risk potential. With the help of statistical processing of these expert answers, task groups can be formed and weighted according to their importance.

A behavior analysis is carried out on the behavior description level. Here, the requirements analysis considers how work is performed and not just what the corresponding result looks like. Questionnaires, or **critical incident technique**, can be used to analyze behavior. This reflects particularly effective or particularly ineffective work behavior by experts. The behaviors recorded in this way are then weighted according to their importance, analogous to the task level (Nerdinger et al., 2019).

Critical incident technique
This is a technique for capturing particularly effective or ineffective behavior.

When analyzing the characteristics level, properties are recorded that are considered to be relevant to the success of the job as a whole. These are usually recorded using questionnaires or the consideration of theoretical models, in which the importance of various characteristics is to be evaluated by experts in relation to the activity (Schmitt, 2012).

The requirement-specific information and characteristics recorded in this way are finally integrated into a requirement profile. The result of such a requirements analysis is formulated in terms of personal characteristics. In doing so, it is not only necessary to list the appropriate suitable characteristics but also to decide how these characteristics should be developed and which test procedures should be used to record them, i.e., how this behavior can be **operationalized**. This profile of requirements is usually broken down into skills (which are defined as behaviors that are necessary for professional success); characteristics (traits necessary for success, such as skills, personality traits, interests); and knowledge (technical knowledge that is necessary to be successful in the position such as professional experience, qualifications; Vautier, 2011).

Operationalization
This is the process of making a theoretical construct measurable.

1.5 Screening and Matching

Screening

“Screening” and “screening procedure” are terms that are largely known from the medical field. In Germany, for example, women over the age of 50 can have a mammography screening for the early detection of breast cancer free of charge every year and men over the age of 45 a free screening examination for prostate cancer from a urologist. Both screening examinations have the same goal: to filter out those people from the general population who have not yet been diagnosed with possible cancer so that they can be diagnosed and treated early. Following the same idea, regular screening methods are also used in psychological diagnostics to filter out people with certain psychological characteristics and symptomatology. This is carried out, for example, within the framework of suc-

Sequential diagnostics
This is a step-by-step approach in which diagnostic information is collected sequentially rather than simultaneously.

cessive, **sequential diagnostics** in order to select the persons relevant to the research question from a more or less broad mass, from whom further diagnostic information is then collected (Barrick & Mount, 2012).

A screening process can be used in personnel selection, for example, to identify applicants for whom it is worth continuing the application process, i.e., it makes sense to deepen the diagnostic process. The use of screening methods, therefore, also makes sense from an economic point of view. In personnel development, screenings can be used to record employees who need further or advanced training and who, therefore, make an intervention seem reasonable.

Matching

The matching process is understood to be the comparison of the tested person with the requirement profile, i.e., the question of the fit between a specific person and a specific job. In an increasingly digitized world, such analytical processes are also becoming more and more digitized: Complex algorithms for data analysis are slowly finding their way into strategic personnel management, although the data protection basis for this has not yet been fully clarified for country specifics (Tippins, 2012).

Example

At this point, the reorientation of the recruitment of trainees in a fictional company should be considered as an example. The company had recently experienced that apprenticeship positions could not be filled. The volume of applicants and the suitability of applicants for the training position steadily decreased. It should now be checked whether and which approaches from the field of "data analytics" can be effective for solving this problem.

First, the existing internal requirement profiles were checked, and a requirements analysis was carried out. Job-specific requirement characteristics were identified. In the next step, the diagnostic procedures were selected, which were assigned to the respective requirement characteristics. All diagnostic procedures were brought together in an assessment center. Subsequently, decision-making mechanisms based on mathematical judgments were used. A data analytics assessment tool was created from this process, the aim of which is to support the selection decisions with learning algorithms. These learning algorithms now generate a proposal for a decision on the applicant's progress in the application process. In order to arrive at this suggestion, the algorithm creates weighted scores based on the requirement profile and calculates an optimal threshold value so that as few suitable candidates as possible are eliminated from the application process. Since this is a learning algorithm, all the results of previous recruitment procedures are included in the database, making the algorithm more and more accurate. The ultimate decision to hire an applicant is always made by humans, not the algorithm. The results of this procedure are seen as positive regarding the reorientation of the recruitment of trainees since more suitable candidates were identified than with the previous procedure. At the same time, the use of data analysis increased the efficiency of the selection process.

SUMMARY

The diagnostic process consists of successive steps. First, the diagnostic question is clarified, based on which psychological hypotheses are then formed and checked. The results and diagnostic information from this hypothesis testing eventually lead to a final diagnostic judgment. It makes sense to take a multimodal approach, i.e., to integrate different data sources into the process. In the field of professional aptitude diagnostics, there is an increasing focus on international recruitment (including the cultural and linguistic adaptation of data collection).

It remains internationally standard to create a requirements analysis in order to answer the diagnostic question in the best possible way. In this requirements analysis, it is recorded which requirements a position entails and which suitability characteristics must be present in which form in order to achieve a fit.

UNIT 2

DATA INTEGRATION AND QUALITY CRITERIA

STUDY GOALS

On completion of this unit, you will be able to ...

- describe which standards psychological test diagnostics should meet.
- understand the rules of data integration in diagnostic decision-making.
- identify what the main and secondary quality criteria of psychological test procedures are.

2. DATA INTEGRATION AND QUALITY CRITERIA

Case Study

“You suggested several psychological test procedures to me,” says the human resources (HR) manager of a medium-sized company, who is a customer of junior consultant Kim ~~T.~~, “but what do you have to pay attention to when choosing a test procedure? Is there such a thing as quality criteria? Is one test value or test result sufficient to base my decision on, e.g., which applicant to hire or what further training is required?”

Wright (2020) emphasizes the importance of data integration by saying: “Perhaps the most mystifying (some say intuitive) stage within the psychological assessment process is integrating the data from multiple, extremely varied sources into a coherent picture of the individual being assessed” (p. 65). The following unit will give an outline on data integration and an introduction to quality criteria of psychological test procedures.

2.1 Rules of Data Integration

McPhail and Jeanneret (2012) point out that “although assessment data may be obtained from a variety of sources, at some point this wealth of information must be integrated into a consistent whole to describe the assessee with respect to the particular requirements and situation” (p. 427).

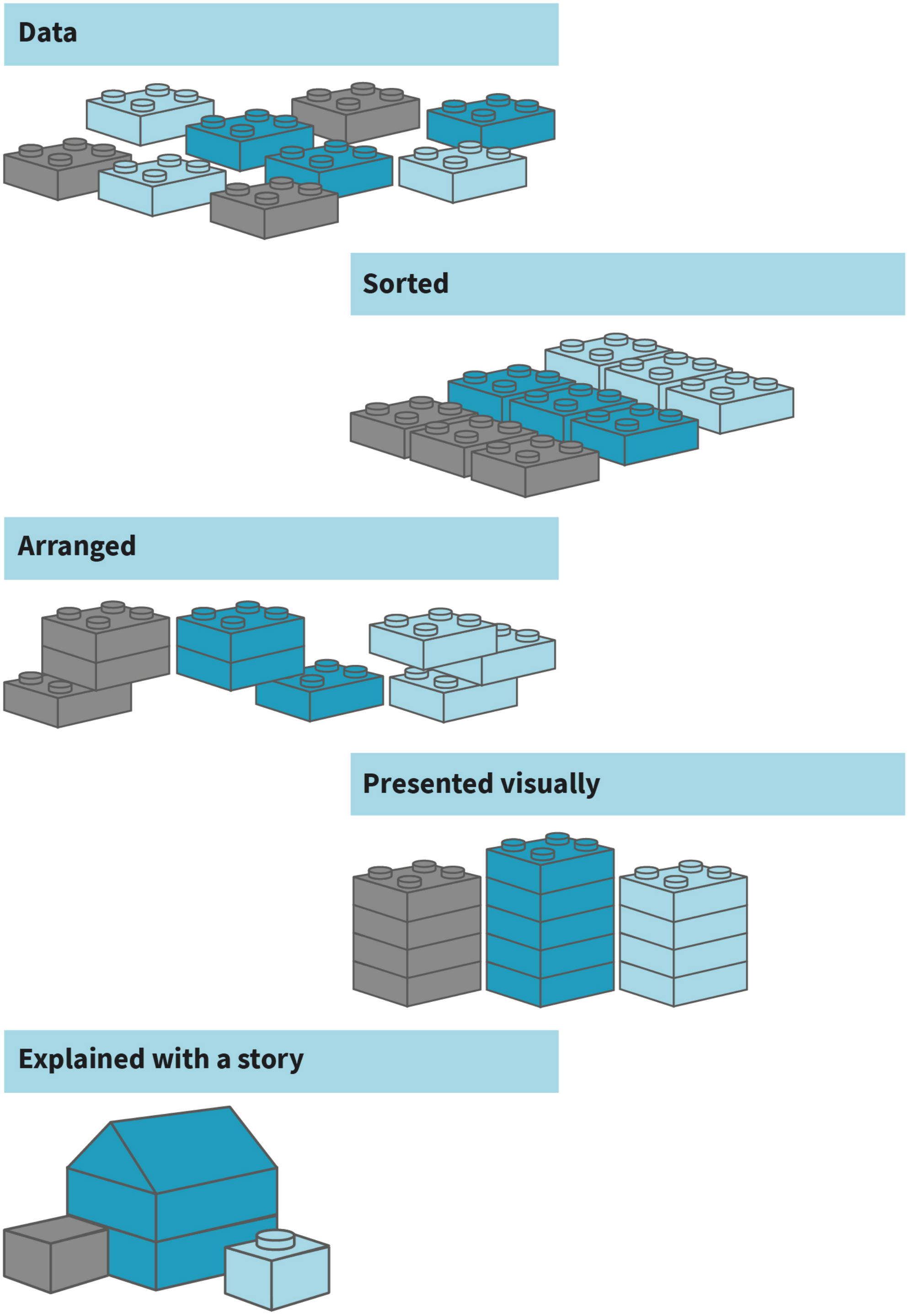
In order to make valid decisions within a diagnostic process, it is important to integrate and weight a large amount of diagnostic information in order to arrive at a reliable judgement. One major reason why, in many countries, only accredited professionals are allowed to conduct psychological assessment is the conceptualization and complexity of integrating data (Wright, 2020). Different approaches to data integration are represented. Wright (2020) recommends a five-step approach:

1. Accumulating the data
2. Identifying themes
3. Organizing the data
4. Finalizing themes
5. Conceptualizing

Wright (2020) argues that the demanding task of assessment can be made more manageable by dissecting the procedure into its constituent parts. The initial step in the procedure is gathering and documenting all your data in one location, including results from tests, the assessment (which can include an interview and other material, such as curriculum vitae (CVs), and behavioral observations. Beginning to classify these facts into a psychological framework is the second stage. The data are organized into basic themes in the third phase, making it simple to assess (a) if there is enough evidence to support each theme and (b) whether the themes make conceptual sense, ensuring that all the data used to support each theme accurately describes the theme. The fourth step is to examine the data to see which themes come together in a coherent and practical way. The fifth and last step is to conceptualize the case utilizing the themes and one of many psychological models as a foundation. Doing so will produce a conceptualization that is very clear and incorporates all the test-related data once this procedure is finished. The conceptualization's narrative framework lends the assessment face validity, making it considerably more probable that the subject would comprehend the results and heed any advice given as a result (Wright, 2020).

The following visualization of a data-integration process depicts the flow of accumulating and organizing data with a final step of conceptualizing or explaining the data.

Figure 2: Data Integration Process



Source: Created on behalf of IU (2023).

In order to collect and integrate data in a sufficient way, Schmitt and Gschwendner (2006) propose six questions of central importance:

1. According to which aspects should diagnostic data be collected?
2. According to which rules should a limited selection be made from a wealth of data?
3. According to which rules should data be linked?
4. When does it make sense to consider diagnostic information simultaneously, when should it flow sequentially into the judgment?
5. How should the individual diagnostic data be weighted?
6. What types of correct and incorrect diagnoses are there and how can their probabilities be evaluated and influenced?

Where these approaches do differ in their specifics, the main goal is mutual and distinguishes the process of assessing, reducing, and combining data into a coherent picture of, for example, an applicant making this process valid and reliable rather than arbitrary. When selecting diagnostic information, it is important to select the data that can clarify the diagnostic question in the best possible way. It is especially helpful if this data can be collected as economically as possible while at the same time being rich in content regarding the aimed characteristic (Wright, 2020). It is also not uncommon for more information to be available than is ultimately needed for decision-making. Information should be prioritized in order of quality and relevance. McPhail and Jeanneret (2012) also indicate the issue of data interpretation by stating:

> Two key issues must be resolved: (1) selection, from among available group norms, one that is most relevant for the current situation and (2) deciding whether the data should be interpreted by comparison to the scores produced by others (normatively) or by comparison of scores within the individual assessee (ipsatively). (p. 428)

There are various ways in which diagnostic information can be linked together. In the case of the conjunctive linking of two characteristics, only those persons are to be classified as suitable who achieve a previously defined level in both characteristics. With the additive linking of these two characteristics described, the non-achieved expression of one characteristic can be compensated by the above-average expression of another characteristic. In the case of disjunctive linking, all test subjects are assessed as suitable as soon as they achieve the previously defined level in one of the two characteristics – regardless of whether they also do this regarding the other characteristic (Schmitt & Gschwendner, 2006). Since these rules are very rigid, combinations of these rules are often used in practice.

Diagnostic data are not always collected or taken into account at the same time. When selecting applicants, for example, the documents that have been sent, such as CV and certificates, are first checked to determine whether the formal requirements for the advertised position have been met. Those who meet the requirements are then analyzed regarding their application documents, such as the cover letter and CV. On this basis, a decision is made as to which applicants are invited to the aptitude test and which are not. Those applicants who pass the aptitude test will then be asked to an application interview. This is referred to as sequential consideration of diagnostic information (Schmitt & Geschwender, 2006). The sequential sequence is usually more economical than the simultaneous consideration and collection of diagnostic information, i.e., letting all applicants go through all the steps directly, even if they are perhaps not suitable for the advertised position in terms of their formal requirements (Schmitt & Gschwendner, 2006). There are situations in which it can make sense to weight diagnostic information differently in the decision-making process. Studies have shown that in many cases of additive linking, this is not necessary or does not create any added value. Under certain conditions, however, weighting can make sense in order to increase the accuracy of the data obtained. Such a weighting is based on a statistical analysis (Schmitt & Gschwendner, 2006).

The data integration usually ends with the decision or the diagnostic judgment. This is mostly a binary decision, e.g., whether an applicant is hired or rejected or whether an employee is promoted to a managerial role or not (McPhail & Jeanneret, 2012). This decision can be right or wrong. Hence it is important to be able to assess the probability of a correct or incorrect diagnosis. For this purpose, these judgements can be divided into four diagnostic decisions: A suitable applicant is identified as suitable, and an unsuitable applicant is identified as unsuitable. Both decisions are correct in this case. Wrong decisions are made when a suitable applicant is recognized as unsuitable and an unsuitable applicant as suitable. We refer to the proportion of people who exceed the critical suitability value as the suitability rate. This suitability value should be checked or tested, for example by means of a psychological performance test. Based on the test result, a decision is then made as to whether someone is considered suitable or not. The proportion of those applicants who achieve this cut-off value is called the selection quota. The selection and suitability rates can both be reduced as well as increased, thereby systematically changing the proportions of correct and incorrect decisions. Diagnosis quotients provide us with knowledge about the frequency of right and wrong decisions as well as the selection and diagnosis quotas (Schmitt & Geschwendner, 2006).

Table 3: Diagnosis Quotient

Predictive accuracy	Percentage of correct diagnoses among all diagnoses made
Sensitivity	Percentage of correct positives among those who are suitable
Specificity	Ratio of correct negatives to incorrect negatives
Positive predictive value	Percentage of correct positives among all positives
Negative predictive value	Percentage of correct negatives among all negatives

Source: Created on behalf of IU (2023).

2.2 Test Standards and Test Economy

There are different test standards, i.e., quality requirements, for different areas of assessment. Bartram and Tippins (2017) allocated a comprehensive list of guidelines with the most important being as follows (p. 218):

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014): Standards for educational and psychological testing.

Society for Industrial and Organizational Psychology (SIOP) (2003): Principles for the use and validation of personnel selection procedures
European Federation of Psychologists' Associations (2013): EFPA review model for the description and evaluation of psychological and educational tests version 4.2.6.
International Organization for Standardization (ISO) (2011): ISO-10667-2 Assessment service delivery – Procedures and methods to assess people in work and organizational settings
International Test Commission (2001): International guidelines for test use
International Test Commission (2005): International guidelines on test adaptation
Equal Employment Opportunity Commission (EEOC) (1978): Uniform guidelines on employee selection standards

All these guidelines hold information on test standards and criteria accounting for high test standards within the field of individual assessment. Considering specific test (procedures), Loewenthal and Lewis (2021) describe characteristics of a good psychological measure which should feature (p. 2):

- a statement of what the scale measures
- justification for the scale – its uses and advantages over existing measures
- a description of how the preliminary pool of items was drawn up
- a description of the sample used for testing
- descriptive statistics (norms) such as means, standard deviations, ranges, different subscales (if any)
- reliability statistics
- validity statistics
- the scale itself (instructions, items or examples of items)
- the scale's construction and use that must follow appropriate ethical guidelines.

In Germany, the Diagnostic and Test Board (DTK), which evaluates the quality of test procedures, deserves special mention. With the test evaluation system, published by the Diagnostics and Test Board of the Federation of German Psychological Associations, questionnaires and test procedures from all areas of psychology can be evaluated using a checklist regarding the completeness of the information provided (*DTK Test Information Standard*). In addition, the system serves as a guide for the development of high-quality questionnaires and tests as well as for the design of manuals and handbooks (Diagnostik- und Testkuratorium, 2018).

Table 4: Excerpt From the Guidelines of the DTK for the Assessment of Tests to Record Human Behavior and Experience

1. General information about the test through the procedural notes and description of the test and its diagnostic objective

- target audience (age range, limitations of applicability)
- test structure (subscales, number of items, answering mode, test forms)
- information on the implementation (time required for implementation and evaluation, required qualifications of the test leader)
- evaluation and interpretation (procedure, available tools)
- information regarding empirical studies

2. Theoretical foundations as a source for test construction

Presentation of the theoretical background:

- precise information on the measurement characteristic
- description of the construct and the underlying theory
- similarity to other tests
- differentiation and added value
- derivation/justification of the items

3. Objectivity

- implementation objectivity (standardization of the test, precise instructions, clear instructions for test administrators, e.g., on how to deal with questions, sample items)
- evaluation objectivity (precise instructions on the use of templates; information on how to deal with unanswered items; how to deal with different observation results or assessments; in the case of evaluations that cannot be completely standardized, measures to ensure the best possible objectivity; in the case of computer-based tests, the evaluation should be able to be checked)
- interpretation objectivity (case descriptions in the manual, information on taking test experience into account, etc., information on the required expertise)

4. Standardization

Are standards available for all specified diagnostic goals?

- representativeness of the norming sample for the target groups
- information on how the data was collected
- size of the norming sample (in relation to the measurement accuracy)

Data integration and quality criteria

- appropriateness of the scale (standard values such as T-scores) in relation to the ability of the test to differentiate
- key user expertise

5. Reliability

Are the characteristic values estimated for the population(s) for which the test is to be used according to the diagnostic objective?

- Consider different types of reliability.
- Note the homogeneity of the sample.
- Is a very high internal consistency due to extremely similar items?
- Assessing the speed component when estimating reliability
- adequacy of the retest interval
- In the case of tests based on the item response theory, specification of the standard error of estimation

6. Validity

The validity of the interpretation of the results obtained with the test is decisive.

- Have the validity coefficients for the population(s) for which the test is diagnostically intended to be used been estimated?
- Has a survey under test conditions that correspond to those in the field of application been implemented?
- Has the validity determination been guided by the diagnostic objective?
- Consider the appropriateness, validity and psychometric quality (e.g., reliability) of the criteria used for validation.
- Assess validity evidence in its entirety.

7. Other quality criteria (susceptibility to failure, immutability and scaling)

Susceptibility to the situational conditions of test execution and the current condition of the test person

- Is falsification of test results ("faking good" and "faking bad") possible?
- Is the relationship between the number of test values and the behavior (scaling) checked or at least discussed?

Source: Created on behalf of IU (2023), based on Schmitz-Atzert & Ameland (2012, p. 131).

Test Economy

The test economy is a secondary quality criterion. Wright (2020) points out that "most important are the time and cost associated with the use of the tests under consideration. A balance must always be struck between getting enough data from tests and creating an assessment protocol that is not overly cumbersome and ultimately prohibitive" (p. 51).

There are various options that make a test more economical. The possibility of a short form of an existing test is often used. For example, the scales used are shortened, i.e., recorded with fewer items. It is important to note that the quality criteria of reliability and validity of the long form cannot be adopted unchecked for the short form. The short form must not simply change the factorial structure of the test procedure. It is, therefore, of essential importance that the short form actually brings the same diagnostic value with reduced effort. Another option of working more ecologically is, following Wright (2020), the use of single sub-tests or scales of extensive measurements (e.g., intelligence tests). As mentioned before, using abbreviated forms calls for the need to check for information value and explanatory power.

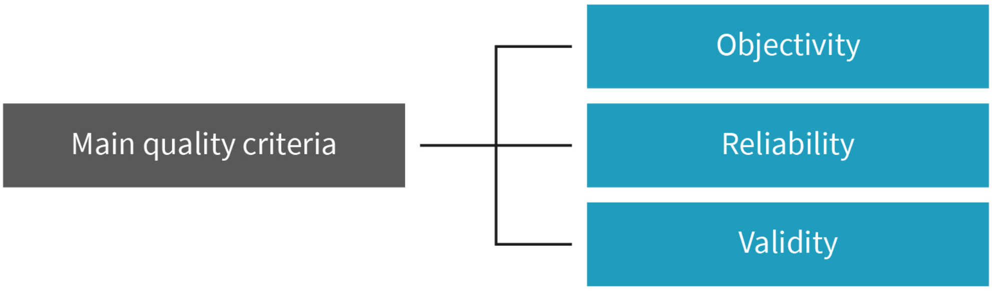
Adaptive testing is also a test economy option in which a test is shortened. The items are presented depending on the response or performance behavior. The item difficulty is, therefore, adjusted to the performance shown by the test participants (Meijer & Nering, 1999).

2.3 Quality Criteria

Quality criteria form the quality features of a test which can be used to evaluate the quality of the same. The main quality criteria are objectivity, reliability, and validity. There are also other quality criteria that are referred to as secondary quality criteria.

The Main Quality Criteria

Figure 3: Main Quality Criteria

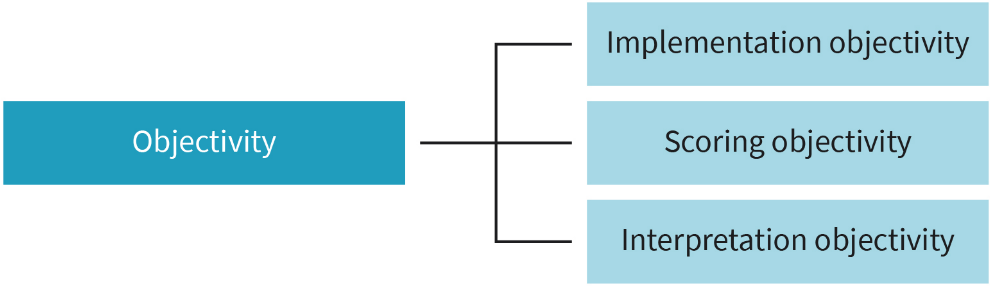


Source: Created on behalf of IU (2023).

Objectivity

The objectivity of a test means that the test result is obtained regardless of who conducts, scores, and interprets the examination. It, therefore, raises the question of how much the result depends on who conducts, scores, and interprets the test (Schmidt-Atzert & Amelang 2012; Tavakol & Dennick, 2011). A distinction is made between different types of objectivity: implementation, scoring, and interpretation objectivity. Good objectivity can be achieved through a high level of standardization. Greene and Ollendick (2019) point out that “the emphasis on objectivity also necessitates consideration of developmental and cultural norms and has ramifications for the selection of assessment procedures and for the types of conclusions one may draw from the information obtained through the assessment process” (p. 435).

Figure 4: The Main Quality Criterion Objectivity



Source: Created on behalf of IU (2023).

For example, when it comes to the objectivity of the implementation, it is important to ensure through precise instructions that the implementation is the same for all people tested. This can be ensured, for example, by the test administrator reading the task carefully, by the fact that everyone is presented with the same material, and by ensuring that all people tested are subject to the same implementation conditions (Schmidt-Atzert & Amelang, 2012).

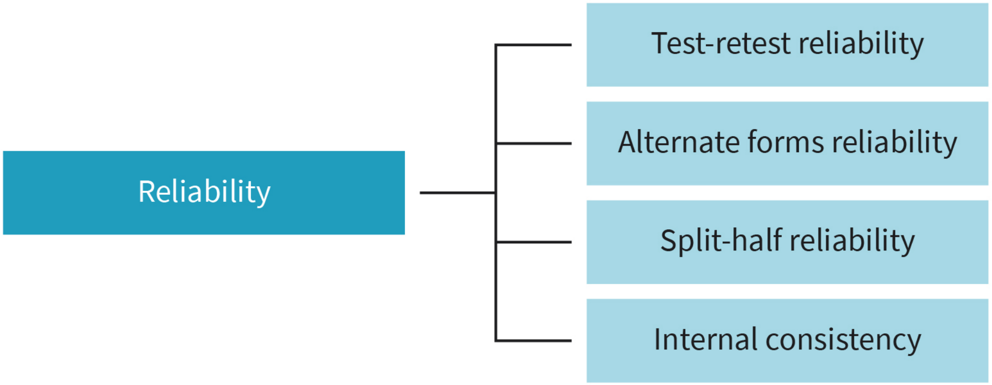
The objectivity of scoring is given if the test result only comes about through the instructions in the manual, e.g., by using an evaluation template, or the evaluation is carried out by a computer program. To do this, it is necessary to specify clear rules as to how results are to be evaluated and how missing answers are dealt with.

Interpretational objectivity occurs when different test users independently reach the same conclusions and interpretations of people with the same test score. This requires information from the manual about which characteristic is measured and which characteristics can be determined in the test person (Tavakol & Dennick, 2011).

Reliability

Reliability reveals how well items measure the underlying constructs. The reliability, thus, indicates the measurement accuracy of the test, whereby perfect reliability would mean that there are no measurement errors at all. However, according to the basic axioms of classical test theory (CTT), this is never the case since unsystematic measurement errors have an impact on every test execution (DeMars, 2018). It is desirable that this measurement error is kept as small as possible so that the accuracy of a test is as high as possible.

Figure 5: The Main Quality Criterion Reliability



Source: Created on behalf of IU (2023).

Test-retest reliability

Loewenthal and Lewis (2021) define reliability as consistency. Test-retest reliability is also called test repeat reliability. The same test is presented to the same person at two different times. The two test results are then correlated with each other. This determination of reliability is always useful when it is theoretically assumed that the characteristic to be recorded is stable over time, especially over short periods of time, as is the case with personality traits, for example. So, if the conscientiousness of a test person is measured today using Test A, the test result should be as similar as possible if this Test A is presented to the same person again a week later. The differences in the test results that occur represent the measurement error and not true feature changes. When interpreting, it should be noted, however, that the retest reliability decreases the further apart the test dates are, as this increases the probability of a change in the true value. Theoretically, it cannot be ruled out that a person's conscientiousness changes over a period of e.g., two years (Guttmann, 1954; Polit, 2014).

For prognostic purposes in the context of personnel assessment, the test-retest reliability is particularly interesting over a longer period of time. When trainees' ability to concentrate is tested, it is hardly interesting how it turns out three weeks later but, instead, how well the test measures the ability to concentrate over a three-year period (Schmidt-Atzert & Amelang, 2012).

Alternate forms reliability

When testing alternate forms reliability, test takers are presented with parallel versions of a test that measures the same construct but uses differently worded items. These should record the same true values, which means that the correlation of the test results should be correspondingly high. Due to the high construction effort of parallel test versions, this reliability is rarely used in practice (Loewenthal & Lewis, 2021; Schmidt-Atzert & Amelang, 2012).

Split-half reliability

The split-half reliability is also called "test-halving reliability." A test is divided into two test halves that are as parallel as possible, and the correlation of these two test halves is then recorded. This correlation must then be extrapolated to the entire length of the test using statistical means in order to be able to provide information about the reliability. For this, however, it must be possible to assume that the homogeneity and number of items also allow such a division into two halves (Loewenthal & Lewis, 2021).

Internal consistency

Internal consistency, often referred to as Cronbach's alpha, is a generalization of split-half reliability. The test is not divided into two halves but into as many parts as there are items in the test. The Cronbach's alpha value, thus, corresponds to the reliability of the test value variables. The stronger the positive correlation between the items on a scale, the higher Cronbach's alpha and reliability. This is the most common reliability estimate and "considered most desirable" (Loewenthal & Lewis, 2021, p. 8).

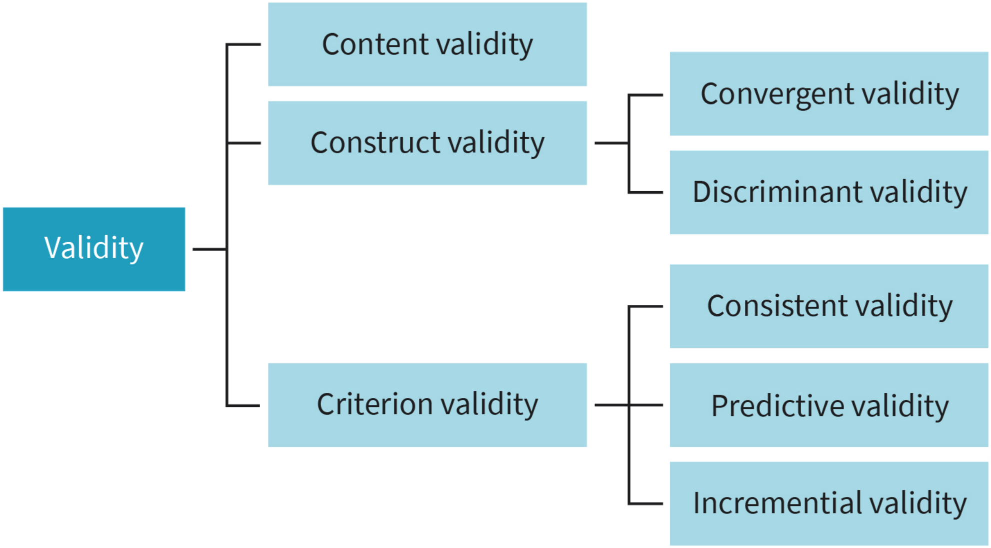
Validity

"A valid test is one that measures what it is supposed to measure" (Loewenthal & Lewis, 2012, p. 13). Only if a test is valid can assumptions be made about the value of the characteristic recorded in the test outside of the test situation. Due to the validity, the test values, therefore, gain real significance beyond the test situation (Schmidt-Atzert & Amelang 2012). The validity of a test is its most important quality criterion because it ensures that the postulated characteristic is measured and not a similar, different characteristic. Accordingly, the most independent and accurate measurement (objectivity and reliability) mean nothing if the correct construct is not captured. There are usually three types of validity: content, construct, and criterion validity (Sackett et al., 2012).

Ryan and Sackett (1998) identified multiple complications which impact the conduct of research and application in individual assessment and suggested improved and different validation strategies. McPhail and Jeanneret (2012, p. 414) collected these five main issues:

1. Clear definition of the predictor and the purpose of evaluating validity
2. The impact of measurement considerations, such as restriction in range, sample representativeness, and potential criterion contamination on the results
3. How assessment results are subsequently integrated into the organization's decision-making process
4. How the assessment data are to be considered, for example, as dimension ratings, integrated judgments, narrative descriptions, or overall recommendations and
5. The role and validity of predictions as separate from descriptions of the assessee

Figure 6: The Main Quality Criterion Validity



Source: created on behalf of IU (2023).

Content validity

The content and face validity of a test is ensured by the fact that the test authors describe their approach to constructing the test and its items. It is not a fixed numerical value that is recorded or calculated, but rather logical and theoretical justifications that are judged by the expertise and authority of experts. A given content validity is, thus, assumed if the psychological characteristic to be recorded is representatively recorded by the test and its items (Loewenthal & Lewis, 2021).

Construct validity

"Construct validity is achieved if you have formulated your test in the context of a theory that makes predictions about behaviour in relation to the test" (Loewenthal & Lewis, 2021, p. 15). Construct validity is, therefore, the most important of the validities. It is present when the underlying psychological characteristics can be inferred from the test results. It is important when estimating construct validity to compare the construct with equal constructs while at the same time distinguishing it from similar constructs. This is distinguished as convergent and discriminant (also divergent) validity.

Convergent validity is estimated by correlating a test result with the results of other tests measuring the same characteristic. If a new intelligence test actually measures the characteristic intelligence and not a similar construct (e.g., ability to concentrate), the test results of the new intelligence test should correlate positively with the test results in other construct-valid intelligence tests. In order to differentiate the construct of a test procedure from similar constructs, it is checked whether it differs from tests that measure the similar characteristic in a construct-valid manner. The discriminant validity of a new intelligence test can be recorded via the correlation with the test values of a concentration test (Sackett et al., 2012).

Criterion validity

Criterion validity captures the relationship between the test result and specific performance or behavior outside of the test situation. Psychological testing procedures are constructed to serve a specific purpose, such as measuring (prognostic) performance, behavior, or traits (Loewethal & Lewis, 2021). A test procedure should, therefore, prove its validity for this criterion. The particular criterion must, therefore, be observable and measurable (Sackett et al., 2012).

In order to be able to validly predict school success using an intelligence test, the test results of the intelligence test are correlated with measurable variables, such as school grades. The criterion examined can either relate to a parameter that is given at the time of the test (consistent validity) or a parameter that only develops in a period of time after the test has been carried out (predictive validity). If later career success is to be predicted or forecast, the predictive validity of the test procedure is particularly important (Loewenthal & Lewis, 2021).

In psychological diagnostics, there is a great deal of interest in clarifying the relevant criterion (e.g., professional success) as comprehensively as possible. Therefore, several test methods (e.g., interviews, questionnaires, tests) are often used in a diagnostic process. The incremental validity describes the increase in clarification, i.e., the improvement in the prediction of a criterion, which is created by using an additional test procedure. In order to determine the incremental validity, a **semi-partial correlation** is used, which can determine the added value of the information (Schmidt-Atzert & Amelang 2012).

Semi-partial correlation
This is the correlation of two variables, in which the influence of other variables is controlled.

Translation

“Because globalization typically requires translations and adaptations of test materials, the deployment of tests internationally poses problems for establishing their validity” (Bartram & Tippins, 2017, p. 282). Many organizations make the assumption that a test that is valid enough for selection purposes in one country will also be valid enough when translated and used in other countries, but this assumption may not be accurate if test-takers are unfamiliar with the test’s format or content or if the test’s construct is altered during the translation and adaptation processes. It is common knowledge that translation alone has issues. The original meaning of test content is frequently severely distorted by translations and backtranslations, which is likely to cause serious issues for test users. This is especially clear when using personality testing. Professional standards emphasize the value of employing efficient adaptation and translation techniques (Bartram & Tippins, 2017). Establishing the equivalence of different test versions helps to promote the idea that validity is consistent throughout multiple test versions, even though this is frequently technically impossible. Even when test-takers have the skill being tested, the familiarity of various populations with different item types and content might affect a test's validity in some cultural contexts: "For example, analogies seem to be more familiar to American applicants than to other foreign nationals" (Bartram & Tippins, 2017, p. 282). Due to the fact that not all countries use the the metric system, some items that need calculations may be more accurate in one nation than in another.

Secondary Quality Criteria

In addition to the main quality criteria, there are many other quality criteria that are referred to as "secondary quality criteria." The standardization of a test is defined as the creation of a reference system with the help of which the results of a test person can be clearly classified and interpreted in comparison to the characteristics of a representative sample of test participants. These are means, standard deviations, and ranges (norms). This is particularly important if we want to use test diagnostics for individual diagnostics (Loewenthal & Lewis, 2021). For example, in order to be able to assess how solving 15 correct tasks should be evaluated in a performance test, it is necessary to consider how many other test subjects also managed 15 correct solutions. The standardization, thus, serves as a frame of reference and is particularly important for the interpretation of test values (Schmidt-Atzert & Amelang, 2012). Norms are often reported separately according to age and gender and in the performance area also according to education (e.g., after school graduation), so that a test person can be compared with people of their gender, their age, and their highest educational qualification.

The economy of a test or questionnaire describes the profitability of a test, i.e., it is measured by the necessary costs (financial and time expenditure) in relation to the diagnostic knowledge gain (Schmidt-Atzert & Amelang, 2012). A test is considered useful if there is practical relevance for the characteristic measured, and the decision made based on this characteristic indicates that more benefit than harm can be expected. A test is described as reasonable if its implementation does not represent a disproportionate burden for the person being examined in terms of time, mental, and physical effort. One aspect of reasonableness is the perceived acceptance of the test, i.e., the extent to which the test subjects accept the test with regard to its items relating to the measured construct (Schmidt-Atzert & Amelang, 2012). This can influence the motivation and willingness to perform during the examination if, for example, the tasks to be processed in a personality test as part of the professional aptitude diagnostics have no visible connection with the profession, such as questions about spiritual world views (Truxillo et al., 2017).

It is important that a test cannot be falsified because this ensures that the design of the test alone makes it impossible or almost impossible for the test person to deliberately falsify it. There are situations in which it is important to the test persons that their answers are particularly good or particularly bad (Furnham, 2017). In the case of performance tests in professional aptitude diagnostics, this is usually not a problem, since one cannot distort one's own performance upwards; with personality tests, which are designed to be easy to understand, it is certainly possible to present yourself more positively.

A test is fair if the test results do not lead to any systematic discrimination against certain people because of their ethnic, socio-cultural, or gender-specific groups, i.e., if discrimination based on these factors can be ruled out. In this context, it should be made clear that a test is not inherently fair or unfair. The unfairness arises at the moment when it is used in an environment that consists of the correspondingly disadvantaged people (Schmidt-Atzert & Amelang, 2012).

2.4 Meta Analysis

The basic idea behind meta-analysis in clinical research is to systematically find and, when applicable, statistically combine the findings of all studies that have addressed a certain research issue. Given the growth of information in clinical research, it makes perfect sense to base research reviews on precise quantitative collection of study results and systematic searching (Naylor, 1997). Previously, local validity studies were prioritized before validity generalization research. As a result, it was challenging to improve theory and collect knowledge about the relationship between predictors and employee's performance outcomes (Banks & McDaniel, 2012).

Situational specificity theory
This is the conclusion that an employment test's validity in one organization did not seem to transfer to another organization.

The **situational specificity theory** was disputed in the late 1970s by Schmidt and Hunter (1977) who argued that statistical artifacts caused variations in the validity of personnel selection procedures. They suggested that employment test validity is typically consistent across organizations. When researchers adjusted for variance in study results brought on by statistical artifacts, this stability became clear. The conclusion was that local validation studies are not always required if a company wants to use selection techniques. Banks and McDaniel (2012) emphasize this finding by stating: "One of the major contributions of this work was the observation that much of the variation across applications in the validity of the personnel selection methods was caused by simple random sampling error (sampling error is one type of statistical artifact)" (p. 158). In order to account for variance between studies brought on by sampling error, measurement error, and range restriction, Schmidt and Hunter (1977) established procedures which made it possible to estimate the population or the true validity of employment tests.

"When the variability in population validity indicated that most validities would be positive in future applications, the employment test was considered to have validity generalization. This indicated that the validity would generalize across most applications in which the test might be used" (Banks & McDaniel, 2012, p. 158). The method proposed by Schmidt and Hunter (1977) is now referred to as psychometric meta-analysis. Its application to demonstrating the extent and relative stability of validity across conditions is known as validity generalization. Another prominent type of meta-analysis is in the tradition of Hedges and Olkin (1985) and differs from the psychometric meta-analysis in some assumptions regarding statistical artifacts.

These two methodological strategies center on estimating the population distribution of studies. Both meta-analysis methods acknowledge that random sampling error causes correlations (and other effect sizes) to vary from study to study. Meta-analyses in the Hedges and Olkin tradition typically do not take other statistical artifacts into account, whereas psychometric meta-analysis explicitly takes these into account (Banks & McDaniel, 2012). Now that statistical artifacts can be controlled and corrected, researchers can more precisely measure the validity of employment tests. In conclusion, validity generalization has made significant contributions to the development of personnel selection theory and practice. Researchers were unable to build knowledge and advance ideas regarding the validity of employment assessments before the introduction of validity generalization (Banks & McDaniel, 2012).

SUMMARY

Data integration is mostly gradual. After the diagnostic information has been selected, it is reduced. When linking this diagnostic information, a distinction is made between conjunctive, additive, and disjunctive linkage, which can lead to different diagnostic decisions in each case. The diagnostic information can be considered simultaneously or sequentially and weighted according to their relevance. The integration of the psycho-diagnostic data ultimately leads to a diagnostic decision.

Quality standards for psychological testing procedures are of great importance. The main quality criteria include objectivity (independence of the test from the test administrator), reliability (trustworthiness of the test), and validity (meaningfulness of the test). In the ideal case, this means a test that reliably measures the characteristic that it claims to measure independently of possible sources of interference. Another quality feature are meta-analyses. Meta-analysis integrates and analyzes the results of various individual studies in a research area as systematically, representatively, and objectively as possible in the form of quantitative variables. In view of the increasingly high number of publications and sometimes contradictory research results, reviews are becoming increasingly important.

UNIT 3

METHODS OF ITEM AND TEST ANALYSIS

STUDY GOALS

On completion of this unit, you will be able to ...

- understand which axioms encompass classical test theory.
- recognize the difference between classical test theory and item response theory.
- describe what types of factor analysis exist.
- recognize why profile analyses are helpful for diagnosticians.
- identify the purpose for which a multitrait-multimethod analysis is carried out.

3. METHODS OF ITEM AND TEST ANALYSIS

Case Study

The next customer that Kim T. receives in her job (junior consultant for the HR department in a management consult firm) asks how one sees or determines characteristics such as intelligence or social skills, i.e., how one can determine a psychological characteristic. “We can’t see a person’s intelligence,” Kim replies, “but we can detect and measure behavior that suggests a person is intelligent, for example, if one scores well in an intelligence test. However, we must always assume that, in this way, we do not record the true intelligence value of a person and that such a measurement also contains measurement errors that are unknown to us.”

If we imagine a subject correctly answering one math question, can we conclude that they have good math reasoning skills? Certainly not! It is intuitively clear to us that this information base is not sufficient to be able to decide whether a subject has mathematical reasoning skills. Many questions remain unanswered, e.g., how much time it took the subject to answer this one question correctly, whether they might just have answered it correctly by chance, and whether they would be able to solve a similar or even more difficult task. A psychological test usually consists of several tasks or questions of varying difficulty that a subject must solve or answer. The result of the test is a score of correctly answered or affirmed items from which various conclusions can be drawn.

The question of the requirements that a test must meet in order to be able to draw conclusions about an actual expression of the tested characteristic based on a test result is the subject matter of test theory.

3.1 Classical Test Theory

The classical test theory (CTT), which is based on a natural science model, assumes that the test result corresponds directly to the true degree of expression of the examined characteristic. The classical test theory is, thus, deterministic. In order to understand CTT, it is important to realize that every psychological test procedure has a certain susceptibility to error. Such measurement errors, or the failure to take them into account when interpreting test values, may lead to incorrect diagnostic decisions (DeMars, 2018; Bortz & Döring, 2015).

Classic test theory assumes that, regarding a certain characteristic, every person has a stable “true value” that describes their actual parameter-value. Such a value is called “latent” because it is not directly observable. This can be, for example, the actual intelligence of a person or a personality characteristic such as extraversion. Since these characteristics are not directly observable, psychological test procedures try to measure them. In the beforementioned cases, this can be done through intelligence or personality tests. But every time a psychological test is used, it not only assesses the “true value” but also a certain **measurement error**. These unsystematic measurement errors can occur in the design, implementation, and evaluation of a test (DeMars, 2018).

Measurement error
This is an unsystematic error whose exact size is unknown and that varies through each measurement.

This means that whenever we use a psychological test procedure, we assess the observable value, which is composed of the true value and the measurement error. The measurement error is not a fixed quantity but varies from measurement to measurement and remains unknown. Since it is impossible to mathematically eliminate the measurement error from a test result, no psychological test result is error-free. Therefore, the main goal of psychological testing is to measure as accurately as possible. The measurement accuracy of a test is described as reliability. The reliability should be as high as possible for the assessed characteristic to be recorded as precisely as possible. The classic test theory makes basic assumptions, known as axioms, which it presupposes *a priori*. They are, therefore, not empirically proven. These axioms make assumptions about the true value and the measurement error in order to estimate the measurement accuracy (Gulliksen, 1950; Novick, 1966):

- The test result is made up of the true value and the measurement error.
- With repeated test application, error compensation occurs. Therefore, the mean of the measurement error is zero.
- The measurement error is independent of the value of the tested characteristic. The true value and the measurement error are uncorrelated.
- The measurement error is independent of the value of other personality traits. For example, the measurement error of an intelligence test should not correlate with test anxiety.
- The measurement errors of different test applications are independent of each other. The error values are independent.

These axioms were originally postulated by Gulliksen (1950) and have had a strong influence on the development and understanding of psychometric test procedures in psychology.

Classic test theory is deterministic in nature. Separate from the measurement error, the test result corresponds directly to the characteristic value. A probabilistic test model, conversely, determines those characteristic values that are most likely for different types of item responses.

3.2 Item Response Theory

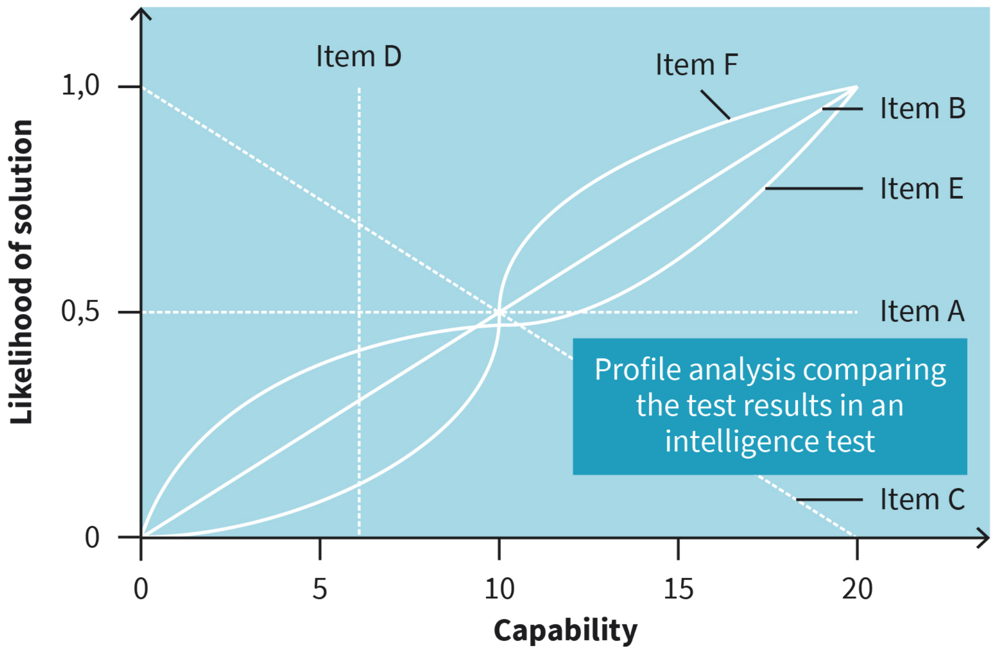
In contrast to the CTT, the basic idea of the probabilistic test theory (item response theory; IRT) is based on the assumption that the probability of a specific answer to each individual item depends on the expression of a latent characteristic dimension. A person with good math reasoning skills is more likely to solve a math problem than a person with poor math reasoning skills. The IRT is, therefore, not subject to a deterministic but to a probabilistic test model and determines those characteristics that are most likely for different types of item responses (DeMars, 2018).

The IRT includes numerous statistical, measurement-theoretical, and psychological models, which can only be touched on exemplary here. An overview of the basics, recent developments, and applications can be studied in Fischer and Molenaar (1995), van der Linden and Hambleton (1996), and Irwing et al. (2018).

Item Characteristic Curve

In probabilistic test theory, probabilities of solving items depending on the ability of the person being tested are of primary interest. The type of relationship that links the probability of an item's solution to the person's ability is called "item characteristic curve" (ICC; Bortz & Döring, 2015). "In three-parameter models, the items may differ in whether the curve starts at zero for people of very low ability (no guessing) or above zero (if guessing correct is likely); how fast the curve rises (discrimination); and whether the curve is to the left (easy), in the middle, or to the right (hard) (Rindskopf, 2001, p. 13023). Only discrimination and difficulty are employed in two-parameter models; guessing is set to zero. Additionally, only difficulty is employed in one-parameter (Rasch) models, with guessing set to zero and discrimination set to one for all items.

Figure 7: Item Characteristic Curve (ICC)



Source: Created on behalf of IU (2023), based on Bortz & Döring (2015).

Dichotomous Logistic Model

Probably the most frequently used probabilistic test model goes back to Rasch (1960). The dichotomous logistic model was developed to analyze tests with dichotomous responses. According to this approach, the number of possible monotonic function types is significantly reduced if a test satisfies the following assumptions (Bortz & Döring, 2015):

1. The test consists of a finite set of items.
2. The test is homogeneous in the sense that all items measure the same characteristic.
3. The item characteristics are monotonically increasing.
4. Local stochastic independence is assumed: Whether someone solves an item or not depends solely on their ability and the difficulty of the item.
5. The number of items solved is an exhaustive statistic of a person's ability: It does not matter which item was solved, just how many.

Based on the dichotomous logistic model, personal parameters (abilities) and task parameters (difficulties) can be determined. Comparisons of people lead to identical results, regardless of the items on which they are based. According to Rasch (1960), they are specifically objective. Conversely, comparisons between different items are also independent of the sample.

Besides its initial goal to analyze tests with dichotomous models, numerous new developments have been established that allow analysis with practically any possible answering format.

Adaptive Testing

A special application variant of IRT is adaptive testing. In conventional tests, the subject processes all items one after the other. This is rather uneconomical because a lot of redundant information is obtained. A subject with medium ability will be able to solve very easy items with a high probability and very difficult items with a low probability. This is avoided in adaptive testing.

If nothing is known about the ability of the person to be tested, adaptive testing begins with an item of moderate difficulty. Then, depending on whether the item was solved, you continue with a more difficult or an easier item. After answering the first two items, a provisional estimate of the personal parameter is possible. This is then successively specified by further items with maximum information. Items with maximum information have a probability of solution of 50 percent. As a result, the difficulty of the items to be processed successively corresponds to the last determined ability (Bortz & Döring, 2015; Meijer & Nering, 1999).

In clinical psychology as well as personnel selection and development, the assessment of change in individuals holds a primary importance. This can either be done following the methodologies of CTT or IRT. Jabrayilov (2016) revealed that IRT is superior to CTT in individual change detection, provided that the tests consist of at least 20 items.

3.3 Factor Analysis

Factor analysis is a statistical analysis with which common underlying (latent) variables are suggested/drawn by observable (manifest) variables (Mulaik, 2018). For example, factor analysis can show how the different items of a personality test are explained by their underlying personality dimensions. Factor analyses are calculated using statistical programs such as SPSS, Mplus, or R. Furthermore, two types of factor analysis are distinguished: the exploratory factor analysis and the confirmatory factor analysis.

Exploratory Factor Analysis

Exploratory factor analysis is used when investigators have not made a prior hypothesis about how many factors underly the model or how the individual items are assigned to the various factors. The number of factors is, therefore, only determined by the exploratory factor analysis. An example from personality psychology is the well-established model of the Big Five, the five main personality traits (extraversion, agreeableness, openness, conscientiousness, and neuroticism). The approach used is the lexical approach. It argues that essential personality traits are reflected through language. With the help of lists of over 10,000 adjectives, explorative factor analyses were used to identify five very stable, independent, and largely culture-independent factors, the Big Five (see Booth, & Murray, 2018).

Confirmatory Factor Analysis

The confirmatory factor analysis, conversely, is a hypothesis-testing procedure, which means that it is theoretically determined, before the examination, on how many factors exist and how they relate to the observable variables (Mulaik, 2018). The confirmatory factor analysis can, therefore, be used to test whether there is sufficient agreement between the postulated theoretical model and the empirically collected data that can then either confirm or reject the model. This principle is often used in testing intelligence. For example, the underlying model of the Intelligence-Structure Test 2000-R (IST 200-R; Liepmann et al., 2007) was constructed based on confirmatory factor analyses. The assumed structure of verbal, numerical, and figural intelligence as well as the retentiveness was confirmed based on confirmatory factor analysis.

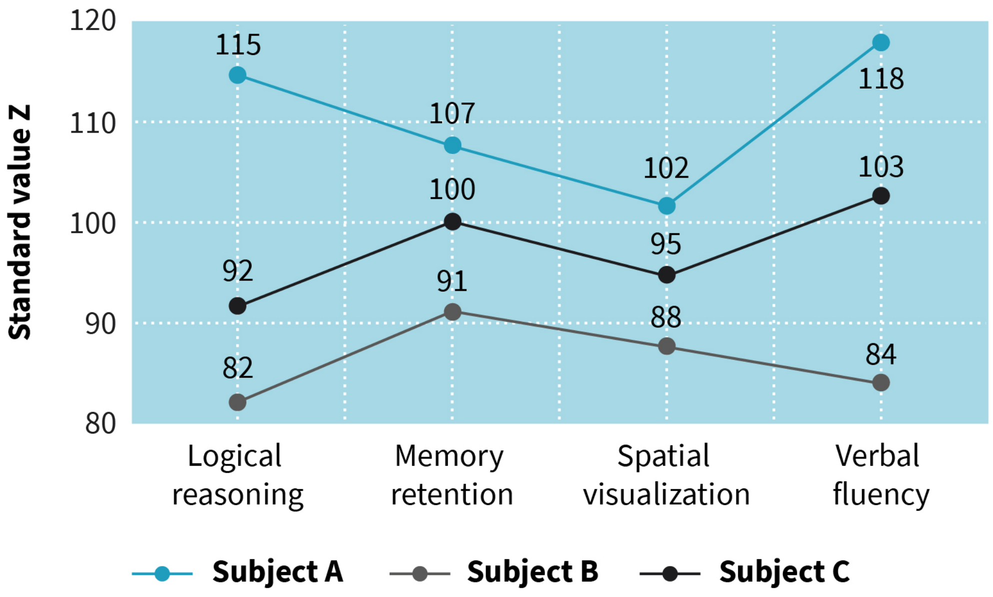
3.4 Profile Analysis

Watkins et al. (2005) assert that “profile analysis has typically been applied for two major purposes: (1) diagnostically discriminating average and exceptional children, and (2) identifying specific cognitive strengths and weaknesses” (p. 251). In this context, we understand profile analysis as a method of analyzing test profiles. Cronbach and Gleser (1953) first mention the need to discuss similarity only with respect to specified dimensions and, therefore, developed D, the sum of the squared deviations of corresponding scores.

Profile analysis includes the summary of results from several independent individual tests using a graphic representation or the comparison of an individual profile with a reference profile. A profile analysis can help diagnosticians, clinicians, or researchers to identify whether participants show significantly different profiles. The analysis can be carried out across groups or across scores for one person. With profile analysis, patterns of tests, subtests, or assessments can be uncovered.

By summarizing results in a graphical representation, a comprehensive and clear diagnostic picture of various psychological characteristics is possible. A test profile does not always refer to the results of different tests. It can also summarize and offer a graphical representation of partial results of one single test when consisting of different subtests, such as an intelligence test. The representation in a test profile can support the content evaluation of one or more tests or the comparison between test-takers by visualizing the results. The profile analysis can display strengths and weaknesses clearly or make deviations from an ideal profile visible. This can help determine an intra-individual fit or help compare inter-individual results of two or more people.

Figure 8: Profile Analysis Comparing the Test Results in an Intelligence Test



Source: Created on behalf of IU (2023), based on Geyser & Eid (2006, p. 319).

When creating test profiles, it is important to ensure that these different variables (e.g., the different facets of intelligence) are comparable. For this purpose, the individually assessed test values (raw data) need to be transformed into standard values (z-values or t-values), so that they are comparable. Here is an example of z-transformation: Firstly, subtract the mean value from the corresponding raw value. The difference between the raw value and the mean is then divided by the standard deviation. This value is called the z-score. In other words, they indicate how many standard deviations a person's test value is from the mean value (see Watkins et al., 2005).

We can describe test profiles by three different characteristics: the amount/height/strength of the profile, the variation of the profile, and the shape of the profile. The height of a profile is defined as the mean of a person across all variables included in the test profile. The profile variation indicates the deviation of the individual test variables from this individual profile mean. Excluding the height and the variation from the analysis leads to the profile shape.

Based on the newly created test profile, different comparisons can be made with the help of statistical analysis. For example, the comparison of two test persons with each other, or the comparison of the test profile to a profile of requirements.

The profile analysis should be carried out with caution. The use of profile analysis with the Wechsler Scales (Wechsler, 2017) in children and adolescents, especially, has been criticized. The main concerns relate to psychometric (e.g., reliability, standard error of measurement, ipsative measurement) limitations and a questionable validity when studies "failed to provide support for the belief that profile or individual subtest scores on intelligence tests are meaningful predictors" (see Bray et al., 1998, p. 214).

3.5 Multitrait-Multimethod Analysis

The multitrait-multimethod analysis is a method for checking the convergent and discriminant validity. The multitrait-multimethod analysis, which goes back to Campbell and Fiske (1959; also see Sullivan & Feldman, 1979) represents a special type of construct validation. This validation strategy requires that multiple constructs (multitrait) be captured through multiple data gathering methods (multimethod).

A systematic, rule-based analysis of the reciprocal relationships between construct and methods allows the level of construct validity to be estimated. In the multitrait-multimethod approach, a distinction is made between two components of construct validity: convergent and discriminant validity (Koch et al., 2018).

Convergent validity
This occurs when several methods consistently measure the same construct.

Convergent validity pictures the correlation of the test results of a Test Procedure A with another Test Procedure B, which purports to record the same psychological characteristic or construct as Test Procedure A. For example: In some studies, subjects are asked to directly indicate how lonely they feel on a rating scale (ranging from 1 = not at all to 6 = strongly). In other studies, participants are presented with a complete questionnaire that addresses multiple aspects of loneliness and, as a result, provides a global score for loneliness. Both the questionnaire and the rating scale are intended to measure the intensity of loneliness. If they represent valid operationalizations, they must correlate with one another and, therefore, be convergent on the construct of loneliness.

Discriminant validity
This requires the target construct to be different from other constructs.

Discriminant validity is used to verify that the postulated construct is captured. This is done by the application of test procedures that are similar but not capturing the same constructs. The discriminant validity of an intelligence test could be checked, for example, by discriminating against a concentration Test. A thorough theoretical preparatory work and a precision of the target construct is required.

With the help of the multitrait-multimethod technique, both discriminant and convergent validity can be systematically assessed using measures of association. The reciprocal relationships between characteristics and methods are presented in a special correlation matrix (Multitrait-Multimethod Matrix; MTMM Matrix; see Koch, et al., 2018).

Multitrait-Multimethod Matrix

Imagine a newly developed questionnaire is designed to measure a psychological characteristic (Trait 1) through a self-assessment questionnaire (Method 1). In order to validate this questionnaire, studies on convergent and discriminant validity are carried out and recorded in a MTMM Matrix. To assess the convergent validity of the self-report questionnaire (Method 1) of a psychological trait (Trait 1), the psychological trait (Trait 1) is also measured using two other methods: peer rating by close friends (Method 2) and peer rating by acquaintances (Method 3). Each of these three methods measures with a certain level of accuracy, i.e., its reliability.

Table 5: Example of a Multitrait-Multimethod Matrix

		Method 1			Method 2			Method 3		
		Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
Method 1	Trait 1	(Rel.)								
	Trait 2	**A**	(Rel.)							
	Trait 3	**A**	**A**	(Rel.)						
Method 2	Trait 1	**B**			(Rel.)					
	Trait 2		**B**		**A**	(Rel.)				
	Trait 3			**B**	**A**	**A**	(Rel.)			
Method 3	Trait 1	**B**			**B**			(Rel.)		
	Trait 2		**B**			**B**		**A**	(Rel.)	
	Trait 3			**B**			**B**	**A**	**A**	(Rel.)

Notes:

- "Diagonal of reliability": The main diagonal displays the reliabilities of the procedures being assessed. Each element on the diagonal represents the reliability of an individual measure or item within a test or assessment.
- "**A**" = "heterotrait-monomethod": Assessing the validity of multiple traits (heterotrait) using a single method of measurement (monomethod).
- "**B**" = "monotrait-heteromethod": Assessing the validity of a single trait (monotrait) by comparing it with multiple methods of measurement (heteromethod).
- All white boxes underneath the "diagonal of reliability" = "heterotrait-heteromethod": Assessing the validity of multiple traits (heterotrait) using multiple methods of measurement (heteromethod).
- All boxes above the "diagonal of reliability" remain empty.

Source: Created on behalf of IU (2023), based on Schmidt-Atzert & Amelang (2012, p. 153).

This is recorded in the MTMM Matrix in the main diagonal (rel.). The test results of the psychological property (Trait 1) of these three methods should be positively correlated with each other, as all three claim to measure the same construct, Psychological Trait 1. The convergent validity shows up as a diagonal in the matrix and is indicated by the letter "B." Here, a characteristic (Trait 1) is measured with three different methods (1, 2, and 3).

To determine discriminant validity, low correlations between the psychological property Trait 1 and similar but different properties to Trait 2 and Trait 3 are expected. Characteristics Trait 2 and Trait 3 are also recorded using a self-assessment questionnaire (Method 1) and two external assessments, one by close friends (Method 2) and one by acquaintances (Method 3). Each of these test methods also measures with a corresponding accuracy (rel.), which is reflected in the main diagonal. The discriminant validity is now shown in the fields with the letter "A," forming a triangle. The different properties (1, 2 and 3), which are each recorded with the same method (1, 2 and 3), are considered here.

To assess the correlations recorded in this way, the findings on convergent and discriminant validity are compared. The correlations of the convergent validity should have higher values than the correlations of the discriminant validity. In the other fields, all located under the main diagonal and without assigned letter, the correlations between different properties (1, 2 and 3), which were measured with different methods (1, 2, and 3), are recorded. So, there should be no specific connection here. All fields above the reliability diagonal are not paid attention to.

SUMMARY

The classic test theory assumes that each person has a "true value" regarding a certain characteristic. However, this "true value" can never be measured, since an unsystematic measurement error is incorporated in every measurement, which varies from measurement to measurement and the actual size of which is never known. Item response theory, on the other hand, considers the characteristics examined as latent dimensions and the individual test items as indicators of these latent dimensions.

Factor analysis is used to infer latent variables from corresponding manifest variables.

A profile analysis can help to visualize and facilitate the interpretation and comparison of test results. It is possible to compare the test profiles of two people with each other as well as the test profile of a person with a requirements profile.

The multitrait-multimethod analysis is a method of analyzing the validity of a corresponding procedure which tests the convergent and discriminant validity in a correlation matrix.

UNIT 4

APTITUDE TESTING

STUDY GOALS

On completion of this unit, you will be able to ...

- identify psychological performance tests.
- describe their application and use.
- understand and evaluate the performance tests test of everyday attention (TEA), Wisconsin card sorting test (WCST), and Wonderlic.

4. APTITUDE TESTING

Case Study

Kim T., a junior consultant in a management consultancy specializing in HR, is getting to know a new client today. The client is looking for a new employee for data processing. Finding suitable employees is not easy. The last person hired for the vacant position had to be dismissed during the probationary period because their performance was hardly what was expected, which had not been apparent from the previous interview and the analysis of the application documents.

The client is now curious about including the performance of the applicants in a concentration test in the personnel selection process. He asks: “What added value should this process offer me?” In her answers, Kim refers to the requirements analysis she carried out beforehand: “A good ability to concentrate is very important in this workplace. The future employee will work with complex data sets. The better they can concentrate, the more likely it is that they will notice mistakes and that they will make fewer mistakes themselves. At the same time, it is planned that the new employee will be accommodated in an open-plan office. Closed rooms are reserved for meetings with clients and other meetings. This means that the future employee will be constantly exposed to stimuli that can distract them, such as a corresponding background noise from conversations and telephone calls as well as the activities and movements of the colleagues present. So, you need an employee who can also concentrate under these circumstances and work well with them.”

In performance tests, people are expected to solve tasks or problems, reproduce knowledge, and demonstrate ability, perseverance, or the ability to concentrate (Katz & Brown, 2019). Performance tests capture the best possible performance, also known as maximum performance. It is always possible to portray one’s own performance as worse than it actually is (“faking bad”). For example, mistakes can be made on purpose, or tasks can be processed more slowly. However, it is not possible to falsify one’s own performance upwards, to present oneself better than one is (“faking good”). Therefore, the use of performance tests makes sense in situations in which the test participants have the motivation to present themselves particularly well, e.g., in an application process (Schmidt-Atzert & Amelang, 2012).

In the case of performance tests, a distinction is made between speed and power tests (Schmidt-Atzert & Amelang, 2012). In the case of speed tests, a limited time is specified within which tasks must be processed. The difficulty of these tasks is usually easy to moderate, and the tasks would be easily solvable for most people without a time limit. The difficulty here is the time component. The different processing speeds of the subjects can then be compared. Theoretically, there is no time limit in power tests. The questions or tasks continuously increase in difficulty and are not easy to solve for most people. In most cases, the level of difficulty up to which the tasks are mastered is recorded. This can then be used to compare subjects. For economic reasons, power tests usually also have a time limit, which is generous enough so that there is no time pressure in the sense of a speed test. It is also possible to combine both methods, e.g., via power tests with speed components, in which tasks of increasing difficulty are to be processed under time specifications.

The use of aptitude testing has historically varied to some extent depending on whether it was determined that an aptitude was something that was stable over time or whether it was adjustable (Stemler & Sternberg, 2013). Silzer and Church (2009) offered the compromise position that some aptitude components reflect foundational dimensions and others reflect growth dimensions. In other words, while attributes like receptivity to feedback, risk-taking, and achievement orientation can be cultivated and increased, foundational dimensions are more constant features like strategic thinking and interpersonal skills (Katz & Brown, 2019). Consequently, depending on the conceptual perspective of people performing aptitude exams, the assessment procedures may vary to some extent. This is especially true in some contexts, including testing for employment, in the military, and for college admissions, where there has been a development in the way aptitude testing has been conceived and used over time (Katz & Brown, 2019).

Salgado (2017) points out that “general mental ability (GMA) and specific cognitive tests have been recognized as the most powerful predictors of overall job performance, task performance, academic performance and training proficiency” (p. 115). According to Webb et al. (2002), certain math abilities displayed in early adolescence predicted later job satisfaction and achievement in math and science-related occupations. It’s interesting to note that these early abilities were a better predictor of employment choice than actual college majors. Therefore, general skills predict success regardless of the industry, while more specialized academic skills are a strong indicator of career success.

Various test procedures are presented in this unit. They are used in clinical, educational, and/or occupational assessments.

4.1 Wisconsin Card Sorting Test (WCST)

The Wisconsin Card Sorting Test (WCST) was developed by D. A. Grant and E. A. Berg, and its professional manual was written by Heaton et al. (1993). It is a neuropsychological test that was developed to assess components of executive functions (abstract thinking ability, cognitive flexibility) in patients with frontal brain lesions. It was originally designed as a test for “abstract reasoning ability and the ability to shift cognitive strategies in response to changing environmental contingencies” (Heaton et al., 1993, p. 1).

The WCST is a card-matching task. The test contains four stimulus cards and two decks of 64 sorting cards. The cards with geometric figures should be sorted according to a rule that the respondent should recognize from the test administrator’s feedback. The task’s cards differ in three ways: color (red, blue, yellow, and green), shape (circles, triangles, stars, and crosses), and number (one, two, three, four). Without any direct guidance from the administrator, participants “match” the response card to one of the four stimulus cards for each “trial.” The participant determines the sorting rule through a process of trial and error. The sorting rule is the aspect on which the card needs to be appropriately matched (Miles et al., 2021). The sorting rule changes over the course of the process and should be deduced again. The test person has to recognize sorting criteria, develop and test solution hypotheses, evaluate positive and negative feedback, and carry out a change in the solution behavior against a dominant action tendency. This card sorting technique is designed to capture the inability to maintain a concept and impaired readjustment, lack of learning from feedback, and tendencies to perseverate. The development of a problem-solving strategy under changing stimulus conditions can be examined.

Although the number of trials required to complete the task is now capped at 128 (Heaton et al., 1993), there are shorter and maybe more useful versions of the WCST that have been developed and are frequently utilized in clinical settings (e.g., the WCST-64 card version; Greve, 2001).

The "perseverated-to" principle is the crucial WCST score factor (Heaton et al., 1993). Miles et al. (2021) define this principle as "the incorrect sorting dimension which a participant is repeatedly responding to (e.g., form, when the correct sorting rule is colour). The perseverated-to principle applies to only one dimension at a time (i.e., colour or form or number)" (p. 2086). According to Strauss et al. (2006), scores on the WCST "can range from 0 for the subject who never gets the idea at all to 6" (p. 528–529).

The methods used by Grant and Berg (1948) and Heaton et al. (1993) to score perseverative replies and perseverative errors differ. According to Grant and Berg (1948), perseverative responses were those that met the prior category's sorting rule and appeared after a rule change. The new WCST manual's more up-to-date definition, however, clarifies that perseverative reactions are those that adhere to the perseverated-to principle (Heaton et al., 1993). As a result, a perseverative answer (or perseverative error) can take place at any point during the task, including just before a rule change and in the first category (Heaton et al., 1993). Problematically, both scoring techniques are still applied in current research, which makes it challenging to compare findings between studies.

The chance of human mistake and misinterpretation of the scoring instructions stated by Heaton et al. (1993) is reduced by using a computerized version of the WCST (Heaton & PAR Staff (2008) program).

Despite widespread agreement that perseverative responses and/or perseverative errors are signs of cognitive flexibility, there is no concrete evidence to support the claim that these variables measure this construct. Instead, there is a widespread understanding that a pattern of consistently incorrect responses denotes rigidity and a lack of flexibility (Miles et al., 2021).

Where it was argued that executive functions may show a high overlap to intelligence, Faber et al. (2022) were able to showcase a significant difference to intellectual ability with the use of discriminant validity analysis. They concluded that "executive abilities, although non-unitary, can be reasonably well distinguished from intellectual ability" (Faber et al., 2022, p. 1).

The WCST was created to screen patients with neurological impairments, and its use in hiring and personnel assessment is, therefore, limited. Due to the significant discrepancies between the patient and candidate target populations, Hommel et al. (2022) started to close the current research-practitioner gap by developing and analyzing a new **gamefied assessments (GA)** tool based on the pattern of the WCST.

GA
The term GA describes different types of assessment tools which share the feature to include game design elements.

4.2 Test of Everyday Attention (TEA)

The test of everyday attention (TEA) is an instrument for recording attention-demanding processes that are relevant to everyday life (Robertson et al., 1996). Three aspects of attention are assessed with eight subtests: 1) selective attention, 2) sustained attention, and 3) attentional switching. The neuro-anatomical model of attention proposed by Posner and Petersen (2012) forms the foundation of the TEA. They suggest that attention is divided into at least three separate systems, each with a unique neuro-anatomical basis: 1) a selection system that selects important processes or stimuli and inhibits unimportant ones; 2) a vigilance system that maintains alertness in the absence of external cues; and 3) an orientation system that engages and disengages attention in space, such as to focus and divert attention.

The TEA evaluation can be used with anyone, from young, healthy individuals to those suffering from Alzheimer's disease. With the inclusion of everyday materials in authentic contexts, the TEA becomes more relevant to the examinee. Different attentional patterns can be identified. 154 UK controls, four age groups, and two levels of educational attainment were used to standardize the TEA. It is sensitive enough to detect typical aging effects in a population of healthy people. There are three parallel versions available, showing high test-retest reliability and significant correlations with existing measures of attention (Robertson et al., 1996). The TEA is translated in multiple languages and offers a version for children ages six to 16 (test of everyday attention for children; TAE-CH; Manly et al., 1999).

It would seem reasonable to suggest that the TEA is relevant to the assessment of clients who are employed in any field where attentional demands of various kinds are likely to play a significant role in their job performance even though the test does not provide data on occupational area of use. McAnespie (2001) points out that "two possible applications could include aptitude testing for candidate selection where occupations demand a certain level of attentional capacity to function at an optimal level, and as part of a battery of assessments used in order to map a client's cognitive profile to a task analysed work rehabilitation programme" (p. 54). While the test would be a helpful supplementary, it must be acknowledged that it will be most useful when employed as part of a larger assessment context that will allow the examiner to take into account the shortcomings of a test not designed with reference to personnel evaluation.

4.3 Wonderlic Cognitive Ability Test

The Wonderlic cognitive ability test, also known as the Wonderlic personnel test, was created in 1939 by Eldon F. Wonderlic (Wonderlic, 1992; 2007). It is a test used in aptitude measurement to determine cognitive ability and problem-solving. The task is to answer 50 multiple-choice questions within 12 minutes. A score of 20 is considered to be indicative of average intelligence. The score is computed as the number of accurate answers provided in the specified time. There are various test formats available. Kazmier and Browne's (1959) investigation, however, demonstrates that, where the test does show high test-retest reliability (Dodrill, 1983), neither of these types can be viewed as directly equal, and therefore, its general reliability is questionable.

According to Matthews and Lassiter (2007), the Wonderlic's strongest correlations were found with general intellectual functioning, which is what it is meant to assess, but at the same time, the Wonderlic test scores did not distinctly demonstrate convergent or divergent validity evidence across these two broad domains of cognitive ability (Matthews & Lassiter, 2007). As a result, they concluded that the Wonderlic was not a successful measure of fluid and crystallized intelligence. The validity of the test was similarly criticized by Hicks et al. (2015). They found that Wonderlic was a significant predictor of working memory capacity for subjects with low fluid intelligence but failed to discriminate as well among subjects with high fluid intelligence. Their study also revealed that Wonderlic has no direct relationship to fluid intelligence once its commonality to working memory capacity is controlled for. These results imply that the Wonderlic is less informative when administered to people with higher-than-average cognitive capacity, implying greater measurement error and reduced practical utility (Hicks et al., 2015). Therefore, certain personnel assessment and aptitude testing (especially for highly qualified jobs) might benefit from using measurements based on recognized constructs with a stronger theoretical foundation, such as fluid intelligence or working memory capacity.

Interestingly, the Wonderlic test has been used as one measure within the NFL Scouting Combine when drafting athletes but is scheduled to be deducted from the process. Research found that, contrary to popular believe, there is no substantial association between a quarterback's Wonderlic score and passer rating, nor between a quarterback's Wonderlic score and compensation (McDonald, 2005). Similar findings were made by Lyons et al. (2009), who discovered that Wonderlic scores did not successfully and significantly predict future NFL success, draft position, or the number of games started for any position. They claimed that despite general mental ability being a highly robust predictor of work performance for the majority of careers, this cannot be assumed regarding an athletic career with other predictors (physical performances) to be of main concern. The study also discovered that, for a few positions, there was a negative correlation between Wonderlic test scores and future NFL performance, noting that the higher a player's Wonderlic test score, the worse their performance in the NFL will be (Lyons et al., 2009).

SUMMARY

Performance tests record the maximum performance of the tested person, e.g., concentration and attention. The ability to concentrate is an important skill in working life and enables people to work as precisely and error-free as possible without being impaired by external stimuli. It is possible for the participants in a performance test to present themselves as worse than they actually are (“faking bad”). On the other hand, it is not possible to present one’s own performance more positively than it actually is (“faking good”). Performance tests are, therefore, suitable for aptitude assessment, and there is a large selection of different tests available.

A neuropsychological test called the WCST was created to evaluate aspects of executive functions (cognitive flexibility and abstract thinking) in people with frontal brain lesions but can be used in aptitude diagnostics testing for these requirements. Another instrument for recording attention-demanding processes is the TEA, assessing selective attention, sustained attention, and attentional switching. A prominent test in personnel selection in the US is the Wonderlic cognitive ability test, which focuses on assessing general intellectual functioning in a short 12-minute frame. The use of these performance tests should always be considered within a framework of clear psychological questions and theoretical approaches and best administered in a multimodal diagnostic.

UNIT 5

INTELLIGENCE TESTING

STUDY GOALS

On completion of this unit, you will be able to ...

- recognize what value intelligence testing has in professional aptitude assessment and its connection to job performance.
- understand how the Stanford-Binet and Wechsler intelligence scales are structured and which theoretical models they are based on.
- describe the quality criteria and application of the Stanford-Binet and Wechsler intelligence scales.

5. INTELLIGENCE TESTING

Case Study

A customer of Kim T., a junior consultant specializing in HR, reports that the performance of the work-study program students at his company fell short of expectations in the last year and the year before. In two cases, the work-study program was discontinued entirely, resulting in financial losses for the company. He would like to see an improvement in personnel selection so that future participants can perform as well as possible in their training and studies and can then be employed profitably in the company. An assessment center is already being used for personnel selection. Kim suggests extending this to include the use of an intelligence test.

General intelligence is considered a valid predictor for characteristics of professional success such as income and job satisfaction. The performance shown on an intelligence test is useful in predicting later career success, particularly in younger people and the more complex the job to be performed (Barrick et al., 2001). Intelligence tests are performance tests; they have established themselves as a separate sub-area in personnel diagnostics. From a psychological point of view, one question in particular is difficult to answer: What actually is intelligence?

FLUID AND CRYSTALLINE INTELLIGENCE
According to Cattell's two factors of intelligence (1987) fluid intelligence encompasses the culture-independent ability to reason and solve problems while crystalline intelligence describes abilities that depend on knowledge and learned experiences (e.g., general knowledge, vocabulary).

5.1 Stanford-Binet Intelligence Scale

Intelligence is not a fixed construct that everyone agrees on – there are various definitions of intelligence and corresponding subsets and abilities that are assigned to it (Freeman & Chen, 2019). The tasks that are set can also be very different. It is possible to get an above-average result on one intelligence test and an average result on another. When using intelligence tests, it must, therefore, always be considered which theoretical model of intelligence and its facets best meets the requirements in order to select the appropriate procedure based on the requirements analysis.

Spearman (1904) made a distinction between general ability and specialized ability, denoted by the letters g and s, respectively. According to Carroll (1993), specific skills like verbal, mathematical, and figurative thinking, for instance, are all positively connected. This positive accumulation results from the fact that it is a general ability of people that is engaged in certain areas during their development. Due to different "investments" of their general cognitive capacity, people of similar intelligence can have different standings in certain abilities (see **investment theory of intelligence**, Cattell, 1987). Because of developmental and educational experiences, distinct interests and preferences, personality traits, and other patterns of unique abilities and talents are created (Ones et al., 2010).

Investment theory of intelligence
This describes how different aspects of intelligence develop in relation to each other. According to Cattell's theory of intelligence (1987), cognitive abilities are divided into fluid (innate) and crystalline (acquired) parts. According to investment theory, fluid intelligence can be "invested" in the acquisition of crystalline intelligence.

Measurement of Intelligence

Previously, correct or incorrect answers accounted for the majority of an intelligence test's score. In order to calculate an estimated mental age equivalent, these scores were added up to a total (Freeman & Chen, 2019). Under the present point-scale system that Yerkes (1915) invented, answers to questions might be graded according to their degree of correctness (e.g., 0 = incorrect, 1 = correct, 2 = ideal), as well as their speed. By the 1930s, the majority of IQ tests were mainly aimed at providing a broad indication of intellectual capacity by assessing either verbal or performance abilities.

Because it demonstrates the precision of measurement and, consequently, the trust we have that individuals' scores accurately reflect their standing on a construct of interest, reliability plays a significant role in workplace assessment generally. The inaccuracy around a person's score, which is directly related to test score dependability, is particularly important for employee selection purposes. The smaller the measurement error, the more reliable the test is. As a result, high test reliability enables us to discriminate more clearly between candidates who earned comparable but different test scores and to be more confident in the candidates' observed rank order on the predictor construct (Ones et al., 2010).

According to Freeman & Chen (2019), the basic communalities of all modern intelligence tests are as follows (p. 69):

- a standardized measure to classify individuals on general cognitive ability
- general cognitive ability measured across a variety of interrelated domains and methods
- general cognitive ability measured through relatively novel or unique tasks
- point-system scoring at least for some subtests
- scores that reflect an individual's ability relative to same-age peers

Where most intelligence tests heavily rely on verbal presentation, there are exceptions. Some IQ tests do not discriminate against someone based on their language and culture (Moore, 2017). The administration of these non-verbal tests, like Raven's Progressive Matrices (Raven, 1938; 2000), is simple and often requires less time, effort, and organizational resources than standard intelligence tests like the Wechsler Intelligence Scale.

Another concept which is not thoroughly covered in IQ tests is creativity. Where creativity appears to have a great impact on certain job performances, it is often overlooked. Creativity is not represented on any IQ test, with a few extremely minor exceptions (Kaufman et al., 2011). Current IQ tests fail to account for ideas like creativity. If measuring intelligence as a whole is what employers desire, then a simple IQ test will not be sufficient (Kaufman & Kaufman, 2015).

Intelligence and Job Performance

Previous studies have discovered a connection between intelligence and professional success (Zagorsky, 2007). Findings from Brown and Reynolds' (1975) investigation into the relationship between general aptitude and earnings showed a significant connection between intelligence and annual income.

According to Gottfredson (2002), cognitive intelligence accounts for 25 percent of the variance in work performance and is a strong predictor for success. Schmidt and Hunter (2004) also noted the same result, indicating a positive correlation between general intelligence and job performance that varied from 0.31 to 0.73. They argue that intelligence is a great indicator of professional performance. Kuncel et al. (2010) showed that IQ predicted job performance significantly better than talent, personality traits, and disposition, which are relevant and established factors on their own. Kuncel and Hezlett (2019) summarize these findings by stating "the vast body of accumulated knowledge about these [IQ] tests is clear: They are among the strongest and most consistent predictors of performance across academic and work settings" (p. 344).

More than just job performance, success or earning, Murtza at al. (2021) confirmed the previous findings, adding the awareness that IQ also predicts job satisfaction, which is highly important considering skills shortage in recent years and employee retention.

Emotional Intelligence

Emotional intelligence
This term was introduced by John D. Mayer and Peter Salovey in 1990. It describes the ability to perceive, understand, and influence one's own feelings and those of others and is, therefore, connecting emotion and cognition.

Emotional intelligence (EI) is a research field that has generated much discussion but seems to have an impact in occupational psychology research. EI as a psychological construct is linked to a number of opposing ideas (Herpertz et al., 2016). There are two prominent ideas regarding EI in scientific research: (1) the ability model by Salovey and Meyer (1990) and (2) the mixed and trait model by Joseph and Newman (2010). Where Salovey and Meyer (1990) strictly divided ability EI and personality, Joseph and Newman (2010) approach EI in a hierarchical order of abilities which influence the development of each other, creating a cascading model. The structure of the cascade model of EI is supported by current studies as well as meta-analytic data (Joseph & Newman, 2010; Shao et al., 2015).

In the context of personnel selection and success in the workplace, the construct of EI has gotten some attention, and studies were able to show that the effectiveness of group activities within an assessment center was predicted by applicants' capacity for emotional regulation (Herperetz, 2016). Farh and colleagues (2012) were able to show that, when working in job contexts with high managerial work demands, employees with higher overall EI and emotional perception ability demonstrate higher teamwork effectiveness (and subsequent job performance). This is because these situations contain important emotion-based cues that activate employees' emotional capabilities.

The different perspectives and definitions of EI have also led to a variety in the measurement of the construct. Objective performance tests, subjective self-report measures, and neurocognitive imaging methods are used, without reaching a gold standard of the assessment method (Hogeveen et al., 2016).

The Stanford-Binet Intelligence Scale

At the start of the 19th century, the education minister in France asked for a workable system for identifying which students were intellectually impaired and may be excluded from normal schooling (see Freeman & Chen, 2019). In 1905, Binet, Henri, and Simon developed an intelligence test in response to this demand, moving away from simple assessments of sensory processing toward more intricate assessments of mental functions including language, learning and memory, judgment, and problem-solving (see Freeman & Chen, 2019). The 1905 Binet-Simon Scale, which served as the model for later intelligence tests, assessed intelligence by testing verbal skills on a variety of relatively novel items (such as comprehension questions that assessed a child's understanding of an abstract question, digit span tests that involved repeating numbers, and vocabulary tests that involved defining concrete terms). An individual's cognitive ability measured by the Binet-Simon Scale scores was calculated in relation to their chronological age. The Stanford-Binet Intelligence Scale was created in 1916 as a result of the scale's improvement through the application of more contemporary psychometric techniques following its translation into English and greater standardization using a population of US schoolchildren (Freeman & Chen, 2019).

The **intelligence quotient** (IQ), which was initially defined as the product of chronological age and mental age $(\mathrm{IQ} = (\text{mental age}/\text{chronological age}) \times 100)$, was introduced by the Stanford-Binet. The main focus in research of intelligence was placed on g, the general factor of intelligence established by Spearman (1904), yet the Stanford-Binet and Binet-Simon remained largely concerned with verbal skills as an assessment of intelligence (Freeman & Chen, 2019).

Intelligence quotient
Today, the IQ is determined by comparison with a large sample in such a way that an IQ of 100 corresponds to the average of the sample (M = 100) and two thirds of the people achieve an IQ between 85 and 115 (SD = 15). The following guidelines apply to the interpretation of IQ:

- IQ less than 70: intellectual disability
- IQ 70 to 84: below average intelligence
- IQ 85 to 115: average intelligence
- IQ 116 to 130: above average intelligence
- IQ greater than 130: gifted

The Stanford Binet Intelligence Scales, fifth edition (SB5; Roid, 2003), which was released in 2003, is the most current change of the instrument in the past century. The SB5 is a frequently used psychometric tool, and like earlier iterations, it is administered in clinical, neuropsychological, and psychoeducational situations. The SB5 can be taken by individuals from 2 to 85 years of age and is translated into many languages, allowing an international application with an implementation time between 20 (for the screening) and 120 (full test) minutes (Grob et al., 2019). The full-scale IQ score on the SB5 has a hierarchical structure, with the five factors of cognitive ability – fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory – being divided into two domains (verbal and nonverbal IQ; Roid, 2003). The intelligence values of the SB5 correlate highly with those of other, established intelligence test procedures. The reliability of the SB5 is between .97 and .99 (Grob et al., 2019). Roid (20003) showed that factorial validity can be assumed for total IQ, nonverbal IQ, and verbal IQ; construct validity in terms of convergent validity for a multitude of other intelligence tests (e.g., WISC-IV) and in terms of discriminant validity for different motor skills tests. In addition, criterion validity for school performance, differential validity for above-average intelligence and intellectual disability, foreign language skills, developmental disorders of school skills, and attention deficit (hyperactivity) disorder are shown. The procedure comes with extensive and child-friendly designed materials (Grob et al., 2019).

SB5 in the Workplace

As mentioned above, numerous empirical studies have discovered strong relationships between intelligence and professional success, earnings, and even job satisfaction (Gottfredson, 2002; Murtza et al., 2020; Zagorsky, 2007), making intelligence an important and rational construct to be evaluated in, for example, personnel selection. A closer look reveals that these studies typically employ a shortened form of screening in order to determine intelligence. The cost of administering an extremely thorough intelligence test does not appear to be justifiable given the research questions and the wealth of data gathered. Fortunately, the SB5 offers a screening which can be completed in about 15 to 20 minutes. Kell and Lang (2017) argue that researching and using particular cognitive abilities is useful in real-world situations, such as personnel selection, more so than simply assessing general g. In this line of thought, the specific application of a nonverbal or verbal IQ measure (which the SB5 offers) might be of use to certain workplace environments and certain job profiles.

5.2 Wechsler Adult Intelligence Scale (WAIS-IV)

While serving as the head psychologist at Bellevue Psychiatric Hospital in New York, Wechsler created the first edition of his adult IQ test. He then published it under the name Wechsler-Bellevue Intelligence Scale in 1939 (Wechsler, 1939). The Wechsler-Bellevue scale was distinct from the Binet scales in a number of significant aspects aside from the fact that it was created solely for use with adults. Wechsler's scale utilized 11 subtests that were arranged according to topic and produced verbal and nonverbal (or performance) IQ scores as well as scores for each of the subtests (Holdnack, 2010).

The original Yerkes (1915) point-scale system, which gave credit for each correct answer and combined the correct answers to provide a raw score for each subtest so it could subsequently be translated to a standard score, replaced the classification of items based on mental age. These unprocessed results may also be added together to get verbal and performance IQ standard scores based on the "deviation IQ," where 100 represented the population's mean intellect, and increments of 15 represented the standard deviation. The first revision of the scale was published in 1955 as the Wechsler Adult Intelligence Scale and followed by subsequent revisions in 1981 (WAIS-R), 1997 (WAIS-III), and 2008 (WAIS-IV; see Holdnack, 2010).

The current WAIS-IV (Wechsler, 2008) has changed significantly compared to its predecessors. Newly developed sub-tests (e.g., shape scales, visual puzzles) make it possible to record facets of intelligence that current research has shown to be significant (Wechsler, 2008). The main battery consists of 10 subtests focusing on four specific domains of intelligence: verbal comprehension, perceptual reasoning, working memory, and processing speed. In addition, the division into verbal and performance parts was abandoned and replaced by four index values. Also, an overall IQ can be determined using the Wechsler Intelligence Scale (Wechsler, 2008). With the four indices, detailed statements can be made in the areas of language comprehension, perceptual logical thinking, working memory, and processing speed.

The intelligence test procedure distinguishes between verbal and action intelligence. This division enables a differentiated assessment of a person's level of intelligence (Petermann, 2008). Further analysis can be done at the subtest level. In this way, a targeted statement about a person's strengths and weaknesses can be made with the profile analysis. In addition, process analyses provide valuable information for well-founded interventions.

The reliabilities of the subtests are between $\mathrm{r} = .76$ and $\mathrm{r} = .91$ and at index level between $\mathrm{r} = .87$ and $\mathrm{r} = .94$. For the overall value, the reliability is $\mathrm{r} = .97$ (see Petermann, 2008).

Content validity can be assumed, the internal structure considered proven. Clinical validation studies with highly skilled and intellectually disabled people have been conducted and confirm the validity of the test (Petermann, 2008). The evaluation is based on representative and comprehensive standardization samples. The test was also translated into multiple languages, and the assessment time is 90 and 115 minutes.

WAIS-IV in the Workplace

A similar argumentation as mentioned in the paragraph about the use of the SB5 in personnel selection is reasonable. Many empirical studies, as described above, have found significant correlations between professional success, earnings, and even job satisfaction and intelligence (Gottfredson, 2002; Murtza et al., 2020; Zagorsky, 2007), making intelligence an important and reasonable construct to assess in, for example, personnel selection. A closer look at these studies reveals that the WAIS was never used or carried out in its entirety. Due to the research questions and a wealth of variables collected, it does not seem economical to carry out an extremely comprehensive intelligence test such as the WAIS. After all, the implementation time is between 90–115 minutes. Since such tests are difficult to carry out in groups, the application in the context of personnel selection seems rather limited.

However, the WAIS also allows you to use its individual subscales and, thus, for example, only use those scales specifically assessing language comprehension, perceptual logical thinking, working memory, or processing speed, depending on the requirement profile of the job.

SUMMARY

Intelligence is summarized as cognitive or mental performances in the context of problem-solving. The term encompasses the totality of differently developed cognitive abilities to solve a logical, linguistic, mathematical, or meaning-oriented problem. Intelligence is considered a valid predictor of academic and professional success. Conducting an intelligence test or specific sub-tests is, therefore, particularly suitable in the area of personnel selection. Possible methods here are the Stanford-Binet Intelligence Scale (SB5) and the Wechlser Adult Intelligence Scale (WAIS-IV). Both intelligence tests are valid, reliable methods of measurement. Since intelligence tests are often associated with an increased amount of execution time, their use should be verified, and it may be advisable to use short versions or single scales, if appropriate.

UNIT 6

PERSONALITY TESTING

STUDY GOALS

On completion of this unit, you will be able to ...

- understand and define personality and its impact on work performance.
- describe which personality traits are referred to as the Big Five and why they are often considered important in a professional context.
- apply and evaluate the different personality tests called 16 personality factor questionnaire (16PF); neuroticism, extraversion, openness to experience five factor inventory (NEO-FFI); and occupational personality questionnaire (OPQ).

6. PERSONALITY TESTING

Case Study

Kim T. sits in the conference room of her HR consultancy across from a client who leads a small team. Low employee turnover is just as important to the client as team members fitting together, so that her team can act effectively as a cohesive unit. Now, there is a position to be filled.

“We don’t just want to choose a high-performing colleague but also someone who fits into our team, who cares about the same values, and whose personality traits can complement our team. Assessing personality traits can be useful in this context,” explains Kim. “In addition, there are also personality traits that are said to be related to professional success. It makes sense that concentration and intelligence tests can predict professional success.” “But personality traits?” asks the client. “Are there personality traits that can predict career success?”

6.1 Definition and Models

“Personality refers to relatively enduring patterns of thoughts, ideas, emotions, and behaviors that are generally consistent over situations and time and that distinguish individuals from each other” (Barrick & Mount, 2012, p. 226).

Trait Models

A trait is a personality attribute that is a generally constant quality that leads individuals to behave in specific ways. Approaches to personality based on traits suggest that behavior is determined by these relatively stable features that serve as the core units of one’s personality. These models of understanding personality suggest that, regardless of the scenario, an individual’s traits would lead them to respond in an expected and consistent manner. This indicates that traits should be stable across settings and time (Laible, 2020).

According to the trait theory of personality, individuals have a number of basic qualities, and the degree and intensity of those traits account for personality variations. These ideas are frequently developed with the use of psychometric tests that measure personality. These approaches consider trait scores to be continuous quantitative variables. A person is given a numerical score to indicate how much of a quality they possess.

Therefore, the attribute approach to personality focuses on distinctions between people. Trait theories are concerned with the components of personality, not with how personality develops.

Numerous theories have emerged as a result of research into personality psychology. Here are four of the most prominent trait theories:

1. The five-factor model of personality (FFM) assesses extraversion, neuroticism, conscientiousness, agreeableness, and openness to experiences (McCrae & Costa, 1989).
2. Cardinal traits, core traits, and secondary traits are the three fundamental categories into which trait theorist Gordon Allport classified the functions and hierarchical structure of personality traits (Allport, 1937).
3. The psychologist Hans Eysenck (1967) developed his assessment of human personality based on three primary characteristics: introversion vs. extraversion, neuroticism vs. emotional stability, and degree of psychoticism.
4. In order to ascertain how an individual's personality type emerges from a collection of sixteen distinct variables, Raymond Cattell employed factor analysis (Cattell & Cattell, 1995).

Personality in the Workplace

There have been numerous research studies on the influence of personality in the workplace, including success and leadership. Most of this research is based on the FFM and its relationship between the **Big Five** personality traits. Research has consistently found that individuals high in conscientiousness tend to perform better in their jobs as they are responsible, dependable, and hard-working (Huo & Jiang, 2021). Studies have shown that individuals high in agreeableness tend to perform better in jobs that require cooperation and interpersonal skills but may not perform as well in competitive or individualistic environments (Bradley et al., 2013). Individuals high in openness tend to perform well in jobs that require creative thinking and problem-solving skills. Research has shown that individuals high in neuroticism tend to have lower job satisfaction and may experience higher levels of stress and burnout (Bianchi, 2018). Extroverted individuals tend to perform well in jobs that require social skills and assertiveness but may not perform as well in jobs that require solitude and concentration. According to some studies, extraversion is associated with several positive leadership qualities, such as confidence, sociability, and the ability to communicate effectively. These traits can be beneficial for leaders, as they allow them to form strong relationships with their followers and effectively communicate their vision and goals. That being said, introverted leaders can also be successful (Bono & Judge, 2004).

Big Five
This is a widely accepted trait model of personality that identifies five broad dimensions of personality: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. These dimensions are thought to capture the most important and enduring traits that shape human personality and behavior.

The relationship between the Big Five personality traits and job performance is complex and can vary depending on the specific job and context. Personality is just one of many factors that can influence job performance, and individual performance can also be affected by skills, abilities, motivation, and work experiences.

Response distortions

According to some, faking is not a huge issue because people fill out personality inventories both instinctively and purposefully with the intention of making an impression and building a reputation (Hogan et al., 2007). This viewpoint contends that the personality inventory results reflect the impression the person chooses to project, which is typical of how most individuals behave in day-to-day interactions with others. It should be emphasized that response distortion can also occur with various non-cognitive selection methods, including interviews, assessment centers, biographical information, situational judgment tests, letters of reference, application blanks, and resumes. This does not lessen the possible implications on people's personality test results, but it does highlight how common the issue is for many selection criteria. Yet, the reality remains that response distortion has the potential to be a significant issue for personality tests (Barrick & Mount, 2012).

Faking has typically been considered to be unimportant in simulations. The desire to lie may be just as strong during simulations as it is during personality tests, but applicants frequently lack the capacity to lie because of the cognitive demands of the exercises or because of their own low skill and restricted behavioral repertoire. Due to the heightened degrees of response fidelity and involvement in assessment center tasks, it may be more difficult to succeed in faking for contestants (Lievens & De Soete, 2012).

6.2 Cattell's 16PF

The 16PF Questionnaire is a personality assessment tool that was developed by psychologist Raymond Cattell in the 1940s and 1950s through multiple factor analysis (Cattell, 1946). The 16PF has gone through four revisions since it was first published in 1949 (in 1956, 1962, 1968, and 1993; Cattell & Cattell, 1995). The 16PF Questionnaire is designed to measure an individual's personality traits. The questionnaire is based on a comprehensive theory of personality that posits the existence of 16 primary personality traits that underlie human behavior. The questionnaire measures 16 primary personality factors and five global factors that describe a person's overall personality (Cattell & Schuerger, 2003).

The 16 primary personality factors are:

1. Warmth: friendliness, affection, and openness to others
2. Reasoning: logical thinking and problem-solving abilities
3. Emotional stability: calmness, self-control, and the ability to handle stress
4. Dominance: assertiveness, leadership, and ambition
5. Liveliness: sociability, enthusiasm, and energy
6. Rule-consciousness: conscientiousness, responsibility, and adherence to rules and standards
7. Social boldness: self-confidence and assertiveness in social situations
8. Sensitivity: emotional sensitivity and empathy towards others
9. Vigilance: alertness, caution, and suspicion towards potential threats
10. Abstractedness: imagination, creativity, and an interest in abstract ideas
11. Privateness: reserve, introspection, and a need for privacy
12. Apprehension: anxiety, worry, and self-doubt
13. Openness to change: openness to new experiences, ideas, and perspectives
14. Self-reliance: independence, autonomy, and self-sufficiency
15. Perfectionism: attention to detail, thoroughness, and a drive for excellence
16. Tension: stress, nervousness, and unease.

The five global factors are:

1. Extraversion: how outgoing and sociable a person is
2. Anxiety: how prone a person is to worry or anxiety

3. Tough-mindedness: how rational and tough-minded a person is
4. Independence: how independent and self-reliant a person is
5. Self-Control: how self-controlled and disciplined a person is

The 16PF questionnaire consists of 185 items that are designed to assess an individual’s personality across each of the 16 primary personality traits. The items are answered on a five-point Likert scale ranging from “strongly disagree” to “strongly agree,” The questionnaire takes approximately 35–45 minutes to complete and can be administered in both paper-and-pencil and computerized formats (Cattell & Cattell, 1995).

Research has shown that the 16PF questionnaire is a reliable and valid measure of personality. Studies have demonstrated that the questionnaire has high test–retest reliability and internal consistency, meaning that the scores are consistent over time and across different items. Additionally, the questionnaire has been found to have good convergent and discriminant validity, indicating that it measures what it is intended to measure and distinguishes between different aspects of personality (Cattell & Cattell, 1995).

The intricacy of Cattell’s “all-inclusive” psychometric approach has proved to be challenging, acting as an ongoing source of irritation for some psychological researchers and practitioners alike despite his considerable publishing and research output (Shye & Gorsuch, 2008).

The 16PF commonly uses clinical terminology that, applied in occupational psychology, practitioners and test subjects can find challenging to understand. Using the 16PF, which may provide additional information, such as the finer distinctions in the emotional arena, necessitates a greater comprehension of psychological concepts in occupational psychologists (Swinburne, 1985). When it comes to hiring for positions needing a wider variety of attributes than, for example, those assessed by the Occupational Personality Questionnaire (OPQ), the 16PF is probably more helpful, especially when distinct emotional qualities are crucial. The ability to analyze 16PF data will continue to depend on good knowledge regarding psychological terminology and concepts as well as the workplace and the capacity to make connections between the two (Swinburne, 1985).

6.3 NEO-FFI as a Measurement of the BIG 5

In the 1980s and 1990s, a model was developed that divides an individual's personality into five traits.

Five-Factor Model of Personality

The five-factor model of personality (FFM) is a collection of five major trait dimensions or domains, often known as the Big Five: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. The Big Five/FFM was devised to express as much variance in people's personalities as possible with a limited number of attributes (McCrae & Costa, 1989).

The five-factor model was established in the 1980s and 1990s, primarily based on the lexical hypothesis, which is based on the idea that the essential aspects of human personality had become inscribed in language over time. The aim of the personality psychologist is to extract the core features of personality from the hundreds of adjectives available in language that identify individuals based on their behavioral tendencies (Allport & Odbert, 1936). A variety of measures, including self-report questionnaires, may be used to assess the five factors (Costa & McCrae, 1992).

LEXICAL HYPOTHESIS

The sedimentation hypothesis, also lexical hypothesis or lexical approach, refers to the assumption in psychology that all important personality traits are colloquially represented by adjectives of the respective language.

In order to arrive at the most parsimonious taxonomy possible of basic trait dimensions, lexical studies usually use four sequential steps: (1) extraction of personality-descriptive words from the dictionary of a language; (2) cleaning up the list (e.g., excluding very rare and obsolete words and synonyms); (3) obtaining self-report and/or third-party reports on the words; (4) reduction of the data to a few dimensions using factor analysis.

Neuroticism

This characteristic typically refers to emotional instability and shows up as behavior that may come across as inflexible, irritated, and impatient. People with high neuroticism levels are frequently worried. They are more apprehensive and frequently uneasy. They spend more time overthinking and ruminating over their concerns. Neuroticism may lead to an individual's inability to cope with typical pressures in their daily lives. Lower scores of this dimension are associated with being less preoccupied with concerns. Less neurotic individuals can manage their stress and handle issues in relation to their importance without overacting. As a result, they tend to be less concerned with trivial issues (Widiger & Oltmanns, 2017).

Openness to experience

This personality trait is characterized by a curiosity to explore new things. These individuals are typically more receptive to new ideas and beliefs, particularly those that disrupt their previous ways of thinking. Individuals who possess low degrees of openness, or those who are closed off to experience, are cautious of uncertainty and the unfamiliar. They are more skeptical of beliefs and ideas that disrupt their social order (Ali, 2019). Individuals who perform well on verbal/crystallized IQ tests have been proven to be more receptive to new experiences. One reason for this is that people who are more open position themselves in situations where they are more likely to learn new knowledge (for example, during a visit to a museum) than those who stay in the same, familiar environment (Schretlen et al., 2010).

Agreeableness

Individuals who show high levels of agreeableness are generally sociable and cooperative. These individuals are often liked by others and hold more trust for others. They tend to be altruistic and eager to assist people in need. Because of their aptitude in collaborating well with others, they frequently function effectively as members of a team. Arguments, disagreement with others, and other types of confrontation are all avoided by agreeable individuals. They aim to calm and appease others by functioning as the group's mediator or peacemaker in attempts to avoid conflicts (Ali, 2019). Individuals who score lower on this dimension of personality care less about appeasing others and creating new connections. These people are skeptical of other people's motives and are motivated to behave in their own self-interest, with little concern for altering their behaviors to suit the interests of others.

Conscientiousness

Conscientious individuals are more cognizant of their actions and the results of their behavior than unconscientious individuals are. They assume responsibility for others and generally make sure to follow through on their commitments to others. Higher levels of conscientiousness are also associated with more goal-oriented conduct. They have the drive and ambition to establish and accomplish what they set out to achieve (Ali, 2019). Less motivated activity is associated with lower conscientiousness levels. Punctuality and cleanliness are less important to those with lower levels of conscientiousness. People who lack conscientiousness frequently act more impulsively. Rather than considering the effects of their decisions, they will make decisions on the spur of the moment. According to research, conscientiousness may be influenced by genetics as well as environmental variables (Ali, 2019).

Extraversion

Extraversion is characterized by outgoing, confident social behavior. These types of individuals often try to gain attention because they enjoy being the center of attention. Extraverts are affable and approachable and enjoy interacting with others. They flourish in social settings (McCrae & Costa, 1987). Introverts tend to avoid demanding social engagements since they may feel exhausted in large groups, such as at parties. Smaller social groupings, ideally with known faces, appeal to introverts.

NEO-FFI

The NEO-FFI (Costa & McCrae, 1989) was developed as a short version of the NEO-PI-R or the NEO Personality Inventory-Revised (Costa & McCrae, 1992) and it serves as the standard questionnaire measure of the Five Factor Model (FFM) openness to new experiences, consciousness, levels of extraversion, levels of agreeableness, and levels of neuroticism.

The NEO-FFI is a condensed version of the NEO-PI, which includes 60 items (12 on each scale), rather than the original 240. Following the assessment of the original 240 questions, these can be used to calculate scores for 30 facets (six subfactors for each of the five domains) and five domains (N, E, O, A, and C). In contrast, only the five domain scores from the NEO FFI's 60 questions may be used. Following the method of factor analysis, the top items were used to design it.

The NEO-FFI did not provide the best short form of the instrument because it was built on elements from the first iteration of the NEO-PI. Its psychometric qualities, particularly its item factor structure, have drawn criticism (Becker, 2006; Egan et al., 2000). Hence, 14 items from the NEO-PI-R item pool were recommended as substitutes for the NEO-FFI's original items by McCrae and Costa in 2004. In comparison to the NEO-FFI, the Revised NEO-FFI (NEO-FFI-R) displayed somewhat superior psychometric characteristics and better readability. McCrae et al. (2005) created the NEO-PI-3, replacing 37 NEO-PI-R items, to enhance the psychometrics and readability of the complete NEO-PI-R. The NEO-PI-3 scales were nearly interchangeable with the NEO-PI-R scales and could be utilized by both adults and adolescents as young as 12 (Costa, McCrae, & Martin, 2008; McCrae, Martin, & Costa, 2005). The abbreviated version of the NEO-PI-3, the NEO-FFI-3, consists of 59 NEO-FFI-R items plus the substitution ("I have no sympathy for beggars") for one of the discarded NEO-FFI-R items ("I'm hard-headed and tough-minded in my attitudes") during the development of the NEO-PI-3 (McCrae & Costa, 2007).

This condensed form has been investigated in many nations (Aluja et al., 2005). According to Holden's (1992) findings (Holden & Fekken, 1994), two distinct samples of Canadian university students had alpha reliability indices ranging from 0.76 to 0.87 and 0.73 to 0.87. It is assumed that the five domains are largely complementary to one another. The NEO inventories are made up of descriptive items that are scored on a 5-point Likert-type scale (1 being strongly disagreed with and 5 being highly agreed), such as "I am not a worrier" and "I truly love chatting to people."

The NEO-FFI should be able to be completed in around 15 minutes. There are no time limits for either edition, which is available online or in print. Instead of assigning a single general numerical value to the test results, each component is given a score on a scale. The NEO-FFI´s outcomes for the assessment could turn out to be like this for a single assessment taker: a low score on the neuroticism scale, a medium score on extraversion, agreeableness and conscientiousness, and a high score on the openness scale (Costa & McCrae, 2014).

6.4 Occupational Personality Questionnaire (OPQ)

The Occupational Personality Questionnaire (OPQ) is a psychometric assessment tool that measures various personality traits relevant to the workplace (Saville et al., 1994). Developed in the 1980s by Peter Saville and his colleagues, the OPQ has become one of the most widely used personality assessments in the world.

The OPQ consists of a series of multiple-choice questions designed to measure 32 personality traits grouped into 10 categories (Saville et al., 1994). These categories include energy and drive, assertiveness, influence, social adaptability, team working, attention to detail, decision making, emotional resilience, self-confidence, and managerial potential. The OPQ takes approximately 30 to 40 minutes to complete and can be administered online or in paper form.

Where the test has not been founded on the Big Five dimensions, it is possible to map the specific traits measured by the OPQ and OPQ32 onto the Big Five dimensions (Visser & Du Toit, 2004). For example, the energy and drive category of the OPQ includes traits such as initiative, achievement orientation, and competitiveness, which are related to the Big Five dimension of extraversion. Similarly, the emotional resilience category of the OPQ includes traits such as stress tolerance, emotional stability, and optimism, which are related to the Big Five dimension of neuroticism.

The OPQ has been subjected to extensive reliability and validity testing and has been found to be a reliable and valid measure of personality traits in the workplace (Matthews & Stanton, 1994). The OPQ has also been shown to have predictive validity in a variety of settings, including job performance, job satisfaction, and career success (Tett et al., 1991).

Overall, the OPQ is a useful tool for employers and HR professionals to assess the personality traits of potential employees and to help identify individuals who may be a good fit for particular job roles (Saville et al., 1996). However, it should be noted that no assessment tool is perfect, and the OPQ should always be used in conjunction with other selection methods, such as interviews and job performance tests.

The OPQ32 is a shortened version of the full OPQ assessment, designed to provide a quicker and more efficient way of assessing personality traits in the workplace (Saville & Holdsworth, 1999). The OPQ32 consists of 104 multiple-choice questions that measure 32 personality traits grouped into the same 10 categories as the full OPQ. It was developed to meet the growing demand for a more streamlined personality assessment tool that could be easily administered online. Like the full OPQ, the OPQ32 has been validated and shown to be a reliable and valid measure of personality traits in the workplace (Visser & Du Toit, 2004).

The OPQ32 can be used for a variety of purposes, including recruitment and selection, career development, and team building (Burke, 2008). Overall, the OPQ32 is a useful tool for employers and HR professionals looking for a quick and efficient way of assessing personality traits in the workplace (Swinburne, 1985). However, it should be noted that the OPQ32 is not a substitute for other selection methods, such as interviews and job performance tests, and should always be used in conjunction with other selection tools.

SUMMARY

Personality tests are used to describe and record personality traits. In practice, the Big Five have established themselves as the main characteristics of personality description and form the internationally most used model.

The characteristic of conscientiousness, in particular, is associated with professional success. The NEO Five Factor Inventory (NEO-FFI) is a personality test that measures these Big Five as a self-assessment scale. However, not every psychological test procedure for assessing personality is based on this theoretical background. Cattell’s 16 Personality Factor Questionnaire (16PF) is one such assessment with a hierarchical personality structure with both primary and secondary level traits. The Occupational Personality Questionnaire (OPQ) is an example of a personality measurement focusing on traits relevant for the workplace.

Personality questionnaires are often self-assessment procedures. This must be considered in the interpretation, since the motivation to present yourself well to a potential employer is very high, especially when selecting applicants. They are, therefore, usually more suitable for a global assessment of personality, such as those used in career advice or as a component in role playing and simulations as part of an assessment center.

UNIT 7

OCCUPATIONAL HEALTH

STUDY GOALS

On completion of this unit, you will be able to ...

- define mental health in the workplace.
- understand which screening options for occupational mental health exist.
- recognize the differences between burnout and depression.
- apply different assessment methods regarding occupational health, such as the mild behavioral impairment (MBI), the Well-Being Index, and the occupational depression inventory (ODI).

7. OCCUPATIONAL HEALTH

Case Study

Kim T.'s customer has noticed for some time that his employees' absenteeism has increased in recent years. Kim, a junior consultant at an HR consultancy, hears this all the time. Employees often complain about the high workload, among other things, due to the increased number of cases and the failure to fill positions. In order to support his employees, Kim's customer now wants to expand occupational health management in his company. He would like to know what exactly puts a strain on his employees and what he can do and offer to promote the long-term health of his employees. Kim first educates her client about the ethical and privacy concerns. After all, this is health data of his employees, which should only be recorded in justified cases and only after thorough information and consent of the employees. In addition, Kim refers to her limited expertise and, in this case, recommends cooperation with a clinical psychologist to collect and evaluate the data.

7.1 Definitions of Occupational Health, Mental Health, and Work-Life Balance

The World Health Organization (WHO, n.d.) defines occupational health as "an area of work in public health to promote and maintain highest degree of physical, mental and social well-being of workers in all occupations."

Mental Health

Securing physical health throughout different workplaces concerns occupational safety measures. In this context, we focus on mental health and social well-being. The World Health Organization defines mental health as follows (WHO, 2022):

> Mental health is a state of mental well-being that enables people to cope with the stresses of life, realize their abilities, learn well and work well, and contribute to their community. It is an integral component of health and well-being that underpins our individual and collective abilities to make decisions, build relationships and shape the world we live in. Mental health is a basic human right. And it is crucial to personal, community and socio-economic development. Mental health is more than the absence of mental disorders. It exists on a complex continuum, which is experienced differently from one person to the next, with varying degrees of difficulty and distress and potentially very different social and clinical outcomes.

Mental health problems are a global issue that affect working populations across the world, as noted by the Organization for Economic Cooperation and Development (OECD) in 2013. According to a recent OECD review, 5 percent of working populations in high-income countries suffer from severe mental health problems while an additional 15 percent are affected by moderate mental health issues (OECD, 2013). Workers with common mental health problems, such as depression, generalized anxiety, and simple phobia, as well as those with subclinical problems, like generalized distress, have reported the highest rates at work (Hilton et al., 2008; Sanderson & Andrews, 2006). The societal, familial, individual, and economic costs of mental health problems among working populations are substantial. Work-related mental health problems account for 3–4 percent of gross domestic product in Europe alone, and these costs are expected to rise in the future (OECD, 2013).

Studies in health care show that workplace relationships are crucial to job satisfaction and healthy team functioning, and they strongly affect overall well-being. Positive feedback, a sense of personal and interprofessional collaboration, and relationship-focused leadership are all factors that contribute to workplace well-being. Maintaining positive well-being supports a highly skilled and confident workforce, which in turn helps to meet organizational goals (Romppanen & Häggman-Laitila, 2017). Conversely, a lack of well-being can result in burn-out and fatigue, which can negatively impact not only workers themselves but also, for example, their patients when working in health care or students when working in education (Cleary et al., 2020). An overall sense of balance in well-being can be achieved by developing strategies and support systems that are tailored to both personal and professional perspectives (Barnett & Cooper, 2009; Lee & Miller, 2013).

A previously often held opinion that mental health problems arise exclusively outside of the workplace and are not the responsibility of employers cannot be supported by the data. There is mounting evidence that poor psychosocial working conditions, or "job stressors," can increase the risk of developing clinical and sub-clinical disorders such as depression, anxiety, burnout, and distress (Harvey et al., 2017; LaMontagne et al., 2007; LaMontagne et al., 2010). Employers should, therefore, acknowledge their responsibility for creating a healthy work environment that does not contribute to the development of mental health problems.

Work–Life Balance

Work–life balance is a term often tied to mental health regarding the workplace. Organizations today understand how important it is to address workers' work-life balance problems (Shockley et al., 2017). According to Amstad et al. (2011), this has a positive impact on employee well-being and the organization's ability to recruit and keep top talent. According to research (Twenge et al., 2010), a focus on work-life issues will continue to be essential since the newest generation of employees, known as Generation Y or Millennials, consider work-life balance as a fundamental work value. The phrase "work-life balance" is frequently used to refer to the management of several roles; nevertheless, academic research has paid little attention to the balance notion, concentrating instead on related but different constructs like conflict and enrichment in work-life situations (Shockley at al., 2017). There is disagreement about the concept of work-life balance itself, which contributes to the mismatch between popular vocabulary and study operationalization (Greenhaus & Allen, 2011). Context frequently plays a significant role in the link between work-life interactions and employee retention results, going beyond basic correlations. The understanding of the causes and consequences of employee turnover in the workplace is advanced by the examination of interaction effects for the relationships between work-to-family and family-to-work conflict and retention outcomes (such as turnover intentions). Several factors, including gender, national culture, support, and domain centrality, have been considered as moderators (for a review see Shockley at al., 2017).

7.2 Maslach Burnout Inventory (MBI)

Burnout
This is a well-documented occupational health problem that can lead to negative outcomes for both employees and organizations, including increased absenteeism, decreased job performance, and decreased job satisfaction.

The Maslach **Burnout** Inventory (MBI) was first published in 1981 by Christina Maslach and Susan Jackson (Maslach & Jackson, 1981). The third edition has been available since 1996. The aim of the Maslach Burnout Inventory was initially to record burnout symptoms in people in the helping professions, especially nursing staff, in order to scientifically research this area (Maslach et al., 1996; Maslach & Goldberg, 1998). Over time, the concept of burnout was also related to other work contexts, so that specific questionnaires were also created for teachers and for employees who do not work in the helping professions.

There are now various specified questionnaires that have been tailored to specific professional groups. Depending on the questionnaire, the number of items vary, but almost all of the MBI questionnaires use a seven-point Likert scale. The MBI created a method based on how frequently participants reported having certain sensations, with answers ranging from “never" to "every day” (Maslach & Leitner, 2021). The MBI “aligns with the World Health Organization’s 2019 definition of burnout as a legitimate occupational experience that organizations need to address” (Maslach & Leitner, 2021, p. 2). It is characterized by three subscales (Maslach et al., 1996):

1. Emotional exhaustion: This scale measures feelings of being emotionally overextended and exhausted by one’s work.
2. Depersonalization: This scale measures feelings of detachment or cynicism towards the people one works with.
3. Personal accomplishment: This scale measures feelings of competence and achievement in one’s work.

Each of these three burnout dimensions is evaluated independently by the MBI. Its structure was inspired by earlier exploratory work on burnout from the 1970s, which employed case studies, on-site workplace inspections, and interviews with employees in a range of health and social care professions. A sequence of these statements served as the MBI measure’s items since they all expressed recurring themes in the form of subjective sentiments or attitudes (such as “I feel emotionally drained from my work”).

The MBI has been found to have good validity and reliability across a range of settings and populations. Studies have shown that the MBI is related to a range of negative outcomes, such as decreased job satisfaction, increased turnover intention, and decreased quality of care (West et al., 2016).

In terms of reliability, the MBI has been found to have good internal consistency, test-retest reliability, and inter-rater reliability (Maslach et al. 1997; Wheeler et al., 2011). In contrast to other tests, the MBI does not give an overall value, since the individual dimensions must be considered separately (Maslach et al., 1996; Maslach & Leiter, 2021). The MBI is known worldwide and is the test used most frequently (e.g., Coker & Omoluabi, 2009).

MBI in the Workplace

The MBI can be used in a variety of occupational health settings, including healthcare, social work, education, and business. For example, healthcare organizations may use the MBI to assess burnout levels among physicians and nurses and to develop interventions to improve their well-being and job satisfaction (Shanafelt et al., 2012; Van Mol et al., 2015). Educational institutions may use the MBI to assess burnout levels among teachers and to develop programs to prevent burnout and promote teacher retention.

In addition to its use in assessing burnout levels, the MBI can also be used to evaluate the effectiveness of interventions aimed at reducing burnout (Schaufeli & Taris, 2014). For example, a study might use the MBI to assess burnout levels in a group of employees before and after an intervention, such as a stress management program or a workplace wellness initiative. This can help to determine whether the intervention was effective in reducing burnout and improving employee well-being.

7.3 Well-Being Index (WHO-5)

The WHO-Five Well Being Index (WHO-5) was developed at the Psychiatric Research Unit, Mental Health Centre North Zealand, Hillerød, Denmark in 1998 on behalf of the World Health Organization (WHO). It is a brief questionnaire to measure subjective well-being and mental health status. It consists of five items rated on a six-point Likert scale. Each question can be rated on a scale of 0 to 5, and the total value can be calculated by adding up the values of all the answers. A lower total value indicates lower levels of well-being. A total value of less than 13 can indicate a depression and should be followed up with a clinical interview by a clinical psychologist. Additionally, a percentage can be computed to track changes over time. The WHO-5 questionnaire is available in nearly 30 languages and can be used at no cost. It was developed through a thorough revision process and offers norm values for various populations (Sischka et al., 2020). The WHO-5 is a reliable and valid instrument for screening individuals for depression and monitoring changes in depressive symptoms over time (Topp et al., 2015).

The WHO-5 is widely used in research and clinical settings, and it has been applied in various populations, including adolescents, adults, and older adults, across different cultures and languages (Bech et al., 2013). The use of the WHO-5 in the workplace has also gained attention, particularly in occupational health settings, where it can serve as a tool for assessing and promoting employee well-being (Lara-Cabrera et al., 2020).

Table 6: WHO-Five Well-Being Index (WHO-5)

Please indicate for each of the five statements which is closest to how you have been feeling over the past two weeks. Notice that higher numbers mean greater well-being.		**All of the time**	**Most of the time**	**More than half of the time**	**Less than half of the time**	**Some of the time**	**At no time**
1	I have felt cheerful and in good spirits.	5 □	4 □	3 □	2 □	1 □	0 □
2	I have felt calm and relaxed.	5 □	4 □	3 □	2 □	1 □	0 □
3	I have felt active and vigorous.	5 □	4 □	3 □	2 □	1 □	0 □
4	I woke up feeling fresh and rested.	5 □	4 □	3 □	2 □	1 □	0 □
5	My daily life has been filled with things that interest me.	5 □	4 □	3 □	2 □	1 □	0 □
Total raw score on WHO-5 goes from 0 to 25. To obtain a percentage score ranging from 0 to 100, the raw score is multiplied by 4. A percentage score of 0 represents worst possible, whereas a score of 100 represents best possible quality of life.							
Total raw score		□□	× 4 =	□□			
		(0–25)		(0–100)			

Source: Created on behalf of IU (2023), based on WHO (1998).

In summary, the WHO-5 is a valid and reliable instrument for measuring subjective well-being and mental health status, and its use has been widely adopted in research and clinical settings. Its application in the workplace can provide valuable insights into employee general well-being and can inform interventions to promote a healthy work environment (Bolier et al., 2014).

7.4 Occupational Depression Inventory (ODI)

The Occupational Depression Inventory (ODI) was created by Bianchi and Schonfeld (2020) to assess job-related depressive symptoms and disorders. The authors developed the ODI in accordance with the fifth version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)'s nine main depression diagnostic criteria (American Psychological Association, 2013). As a result, the ODI contains questions to measure anhedonia, low mood, sleep abnormalities, fatigue/loss of energy, hunger abnormalities, feelings of worthlessness, cognitive impairment, psychomotor abnormalities, and suicidal thoughts. Participants are expected to rate any symptoms they have had in the last two weeks. Items are scored on a 4-point scale, with 0 representing "never or almost never" and 3 representing "nearly every day." Each ODI item includes causal attributions to respondents' work/job, as opposed to measuring depressive symptoms in a "cause-neutral" way (e.g., "My experience at work made me feel like a failure"; Bianchi & Schonfeld, 2020). A supplementary question about turnover intention is also included in the ODI: "If you have experienced at least some of the aforementioned issues, do these issues cause you to consider leaving your current job or position?" The choices for responses are "yes," "no," and "I don't know." This supplemental material is meant to aid diagnosticians in determining how the reported depressed symptoms will affect their line of work (Bianchi & Schonfeld, 2020). The authors claim that there are two possible applications for the ODI. Firstly, it can measure job-related depressive symptoms on a spectrum ranging from mild to severe. Secondly, the tool can generate tentative diagnoses of work-related depression (Schonfeld & Bianchi, 2021). A clear restriction needs to be pointed out: "Work-related depression" is not a clinically recognized disease. The ODI cannot be used to assess or diagnose a major depression episode. The ODI is presently offered in English, French, and Spanish.

DEPRESSION

The affected patient experiences a depressed mood and a decrease in drive and activity. The ability to be happy, interest, and concentration are reduced. There is marked tiredness after the slightest exertion. Sleep is disturbed, appetite diminished. Self-esteem and self-confidence are impaired. There are feelings of guilt or thoughts about one's own worthlessness. The depressed mood changes little from day to day, does not react to life circumstances, and can be accompanied by somatic symptoms (early awakening, psychomotor inhibition, agitation, loss of appetite, weight loss, and loss of libido). Depending on the number and severity of symptoms, a depressive episode can be classified as mild, moderate, or severe. 40 percent of women and 30 percent of men will experience at least one severe depressive episode in their lifetime (Andrews et al., 2005).

The ODI presents itself with very good parameters for validity and reliability and allows a high-quality measurement of distress and depressive symptomatology in the workplace (Bianchi et al., 2023). The ODI was also proven to be a good predictor of poor cognitive performance. When adjusting for age, sex, and pre-test transitory mood, the connection was still statistically significant. These findings are in line with the results of previous studies on clinical depression and neuropsychological performance (Bianchi & Schonfeld, 2022).

The development of a construct concentrating on depression closely related and founded in issues relating to the workplace puts the construct of burnout in question. A recent study supported the notion that symptoms of burnout are a component of a larger depressive syndrome rather than a distinct and separate entity (Sowden et al., 2022). This study, conducted in Australia, reinforces the generalizability of this finding and highlights the problematic overlap between burnout and depression. Given significant issues associated with the burnout concept, the authors suggest a paradigm shift towards occupational depression. This shift could lead to more accurate and valid assessments of the extent and frequency of job-related distress and ultimately enable more meaningful and productive conclusions regarding treatment, prevention, and public health decision-making (Sowden et al., 2022).

SUMMARY

Psychological stress is increasing in the world of work and leads to high costs for employers and health and pension insurance providers, lower job satisfaction, or early terminations. The health promotion of employees, therefore, requires more and more attention.

The Maslach Burnout Inventory (MBI), the WHO-5, and the ODI are self-assessment questionnaires for measuring mental health, each with a different focus (burnout, depression, mental well-being). They can be used for individual measurements as well as for course measurements over a longer period of time.

When carrying out psychological diagnostics, it is also important to adhere to the relevant legal framework. There are several pieces of legislation for diagnosticians depending on the country they work and operate in. This is especially crucial regarding the collection of medical data, and assessments should only be done in compliance with data protection regulations and ethical principles and should be carried out by a clinical psychologist.