smart. valid. preferred.

Online Diagnostics

Best practice:
Basics, processes and decision criteria

cut-e GmbH
Neuer Wall 40
20354 Hamburg
Phone: +49-40-3250.3890
Fax: +49-40-3250.3891
E-Mail: info@cut-e.com

www.cut-e.com

Last update: January 2010

Content

Online diagnostics – basics, processes and decision criteria

By choosing appropriate instruments for personnel selection it is often hard to distinguish between 'good' and 'bad' instruments.

This guideline provides help to answer the questions:

- How to identify a good diagnostic instrument?
- Which specific criteria apply to online assessments?
- What options for the interpretation of results are there?

But mainly the purpose of this white paper is to provide a guide through the jargon jungle of general and psychometric criteria.

Application of online diagnostics

As technology and the Internet have spread, so have the uses of online psychometric assessment. But in which situations is online assessment particularly useful or where does it outclass the classic (paper-and-pencil) diagnostics?

<u>Online assessment can be considered particularly effective when:</u>

- Implementing consistent standards in local or dispersed selection processes.
- Speed is of the essence, for example, if vying for a select few candidates due to a limited and tight market.
- You need a reliable way to sift out unsuitable candidates from a large group of candidates by a multi-level, sequential selection process to save time and resources and focus on real potentials.

Local selection with central control

Online diagnostics are often used within situations where people of a standard profile e.g. of sales people, service staff, field technicians or consultants need to be selected locally at different sites.

With the help of online diagnostics, it is relatively easy to implement standards for the selection process, to manage the selection process centrally, but to execute it locally. This simplifies organisational structures, saves resources and accelerates the decision process.

Small market of candidates

Online diagnostics can be very helpful whenever there are a limited number of appropriate candidates per vacancy, as is often the case in the engineering and IT sector. In order to obligate interesting candidates, a fast response and efficient selection process are necessary. Online diagnostics can help to identify good candidates very quickly.

Sequential approach: many candidates to choose from

Online diagnostics are of benefit whenever a sequential approach is required. As shown in fig. 1: Sequential online selection process, online diagnostics play an especially important and resource saving role when screening applicants.

1. A selection process normally begins with defining the job specification (1).
2. On this basis, HR and Marketing implement recruitment communication campaigns (2) e.g. advertisements to generate a more or less large pool of candidates.
3. This pool is then sifted according to 'hard' criteria (3), also called 'gross negative disqualifiers'. Those include defined CV or qualification criteria e.g. driving license, degree level.
4. Following this pre-selection, online diagnostics are applied (4). Candidates remaining in the pool after the pre-selection are asked to complete one or more test procedures. The pool is further reduced after based on the scores of these tests.
5. Candidates who passed the online assessment are invited to participate in an assessment centre or an interview (5). If necessary, tests can be repeated in this stage.
6. Last step is to make offer of employment (6).

In a sequential approach, online diagnostics work as a filter to identify suitable candidates from a pre-selected candidate pool. The main goal is to increase the base rate (see page 14 et seqq. for a definition).

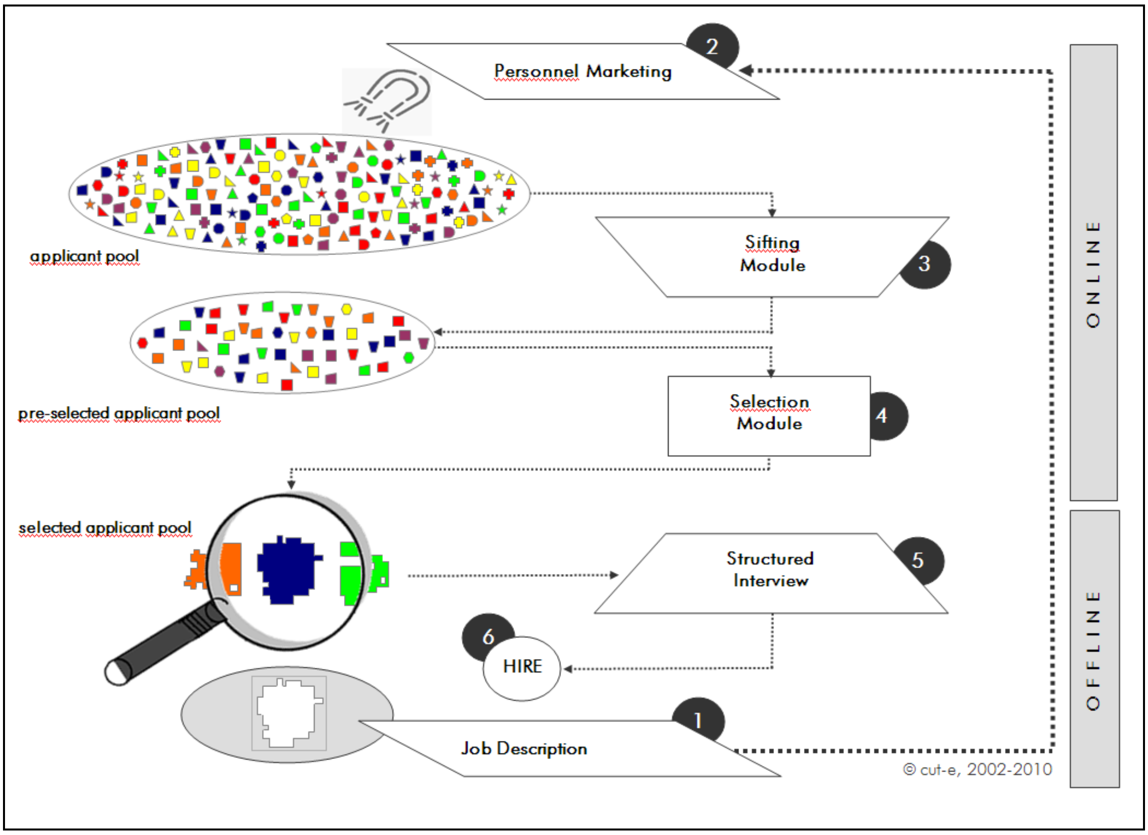

Fig. 1: Sequential online selection process

Administration modes of online diagnostics

With the administration of online diagnostics there are several alternatives: open and controlled as well as supervised / direct administration of instruments.

- The **open and controlled administration** takes place without an administrator, normally from the candidate’s own PC. In these modes, the identity of the candidate cannot be verified. Therefore, they are especially useful in terms of pre-selection or where there will be a later face-to-face selection phase. A common procedure in these modes is to explain to the candidate that the tests will be repeated in a controlled environment, face to face, at a later stage.
- A **supervised administration** relates more to the classic test procedure. The identity of the participant is known and the instruments are completed on the company's computers. In this case, the Internet only works as a medium of distribution.

Open administration

An open administration implies that the candidates register themselves in the system and complete the online assessment. Normally, they receive a password and a URL to start the test from the administrator. With these, the candidates can log in to the particular test system and complete one or more tests.

The key advantage of this form of administration is the low work load for the administrator; for this reason, open administration is especially suitable for large processes with many candidates.

Controlled administration

In the case of a controlled administration, the candidate’s name and email address are registered in the system by the administrator in advance. Then, each candidate receives an individual password (often sent by email) which they can use to log in to the system in order to complete the instruments.

This mode of administration allows the administrator to control who uses the system to complete the instruments. At the same time it creates a higher administrative workload.

Supervised / direct administration

In a supervised / direct administration, the candidate is invited to the company. The administrator registers the candidate and starts the test system whilst present. Then the candidate completes the tasks directly and under the supervision of the administrator.

This approach allows the administrator a high amount of control. On the other hand, it necessitates the computers, the administrator’s time and travel time and cost.

The supervised / direct administration is more appropriate for sequential approaches in a later stage or for smaller selection processes.

General criteria for diagnostics

Diagnostic methods must meet certain criteria to ensure that the instrument measures what it should measure and that the data collection is accurate enough to make secured statements.

Guidelines and criteria sources

DIN 33430 establishes consistent standards for diagnostic instruments.

It contains all important criteria and requirements concerning the instruments and the process. Similar guidelines exist from the International Test Commission (ITC: www.intestcom.org) and from the European Federation of Psychologists' Associations (EFPA: www.efpa.be).

A detailed overview about the use of DIN 33430, with regard to the evaluation of test procedure, can be found in the book 'DIN SCREEN' (Kersting, 2007).

Excellent diagnostics should meet several criteria which are outlined on the following pages. As these are general criteria for best practice, all references are presented independent of instrument or provider.

Objectivity

Objectivity refers to the extent to which the diagnostic results are independent of the investigator.

Aspects of objectivity

There are three aspects of objectivity which correspond to the different stages of examination:

1. Standardised administration: standardised instruction and testing situation
2. Standardised scoring: small decision range in allocation of results; patterns or algorithms (in case of computer-based instruments) for scoring
3. Standardised interpretation: the same conclusions are always drawn from a particular set of results; ideally, the score is compared with the results of a norm group

A test can only be interpreted unambiguously if objectivity in all three stages is assured.

Reliability

Reliability, sometimes also termed accuracy, is the extent to which a diagnostic instrument shows consistent results in repeated application. In other words, does it measure the same thing again and again? A reliable instrument is robust against interfering effects and, at the same time, sensitive to underlying differences in characteristic values.

There are different possibilities to define reliability:

- Re-test reliability
- Alternate form reliability
- Internal consistency

- **Re-test reliability**
  This means that the results of a first application are correlated with the results of a second application that has been carried out at a later time.

  A totally reliable test would provide a correlation coefficient of 1.00 as the persons would reach the same results at the two different measurement times.

- **Alternate form reliability and internal consistency**
  Two more possibilities to calculate the reliability of a test are the determination of the alternate form reliability as well as the calculation of internal consistency (Cronbach's Alpha).

  To determine the alternate form reliability two parallel forms of the same test are carried out. The results of these two test forms are then correlated.

  The consistency analysis (calculation of internal consistency) refers to test results that only derive from a single application and uses characteristic values of the single items of the test, namely item difficulty and discriminatory power.

Although there are no completely reliable instruments, realistically, a reliability value under 0.7 should not be accepted. In the case of questionnaires, the ideal value lies between 0.75 and 0.85, regarding ability tests between 0.8 and 0.9.

Excursus:

Standard error of measurement, expectancy range and confidence interval

With any measurement of characteristics or capabilities it must be assumed that measurement errors occur. Measurement errors are described as the difference between the observable value in the testing situation and the true value of a person that would have been measured if the measurement had been perfect. As the measurement of a characteristic or capability by diagnostic instruments is never at a perfect 100% level, it must be assumed that the observed score differs from the true score. Thus, the standard error of measurement is the part of the statistical spread (standard deviation) that is to account for the deficient reliability of the test. The higher the reliability of a test the lower the standard error of measurement. With a reliability of 1.00 the standard error of measurement would be 0, with a reliability of 0 it would be 1.00.

As each score provides errors, the individual score of a candidate must not be equated with the true score of a candidate. Thus, there is a certain domain of uncertainty that depends on the reliability and the standard error of measurement. The observed scores spread around the true score. Determination of expectancy range depends on the relative security with which a statement should be made. If the standard error of measurement is known, the expectancy range can be determined in which the true score lies with a e.g. 95% probability. A 95% probability contains a 5% probability of error. Hence, a high reliability is important to minimise the expectancy range. If the reliability is low it would only be possible to work with wide expectancy ranges in diagnostics and the diagnostic judgements would not be very accurate. Those adjustments would also provide a high probability of error. Normally, the true score is not known. Thus, a confidence interval can be calculated around the test score in which the true score probably lies. The confidence interval is calculated as follows: observed score +/- probability of error * standard error of measurement.

Validity

Validity is a statement of how far the test measures what it is designed to measure. To what degree of certainty can conclusions be drawn from the testing behaviour to the behaviour outside the test situation?

An example: a valid intelligence test provides a measure for the intelligence of a person and therefore allows a prediction of performance in situations in which intelligence is important.

<u>There are different types of validity which are explained below:</u>

- **Criterion-related validity**
  To estimate the criterion related validity of a test, the scores are compared with the results in another measure, a criterion that is theoretically connected to the test. The criterion represents the measurement objective or, in a pragmatic sense, represents the test purpose.

  The calculated correlation between test result and criterion represents the criterion-oriented validity.
- **Concurrent validity**
  If test and criterion values are collected in a close temporal distance (whereas it is assumed that both characteristics are existent in the persons at this time) the concurrent validity is determined. With this validity coefficient conclusions can be drawn from test result to isochronous criteria. These coefficients are for instance needed to analyse the factors of actual performance problems.
- **Predictive validity**
  Predictive validity can be ascertained depending on the temporal relationship between the collection of measurement and criterion values concurrent. If the criterion characteristic is collected later than the test score the predictive validity is addressed. With this, validity predictions from the test results are possible. Although both values are calculated in the same mathematical manner, their different diagnostic meaning has to be considered.
- **Construct validity**
  The construct validity refers to the extent to which a test, that measures a certain construct, is related to the results of other tests: external ratings, behaviour measurements or experimental results, which are classified as valid indicators of the construct, should be measured and compared.

  The construct validation does not end in a validity coefficient; it rather raises an overall picture of validity. For construct validation all statements that can prove the postulated relations are consulted, even those methods of content and criterion- related validation. In principle, a test should correlate higher with a test that measures approximately the same thing than with a test that measures something else.
- **Face validity**
  If a test or the content of items is purchased in a way that it apparently covers the interesting characteristic or the measured capability, then face validity is meant.
- **Internal and external validity**
  Furthermore, internal and external validity can be identified.

  An instrument has <u>internal validity</u> if its' results are definitely interpretable. Internal validity decreases with the increasing number of uncontrolled interfering variables.

  Moreover, an instrument is <u>externally valid</u> if its' results can be generalised across different situations and persons. That means that the external validity is heavily dependent on the naturalness of the testing situation and the representativeness of the tested sample.

A perfectly valid instrument would provide a validity coefficient of 1.00. As with reliability, this value is utopian in reality. Nevertheless, to give a secured statement about the performance of a candidate, good diagnostic instruments should have a value of at least 0.3.

Fig. 2 shows the cost-value ratio of different diagnostic instruments.

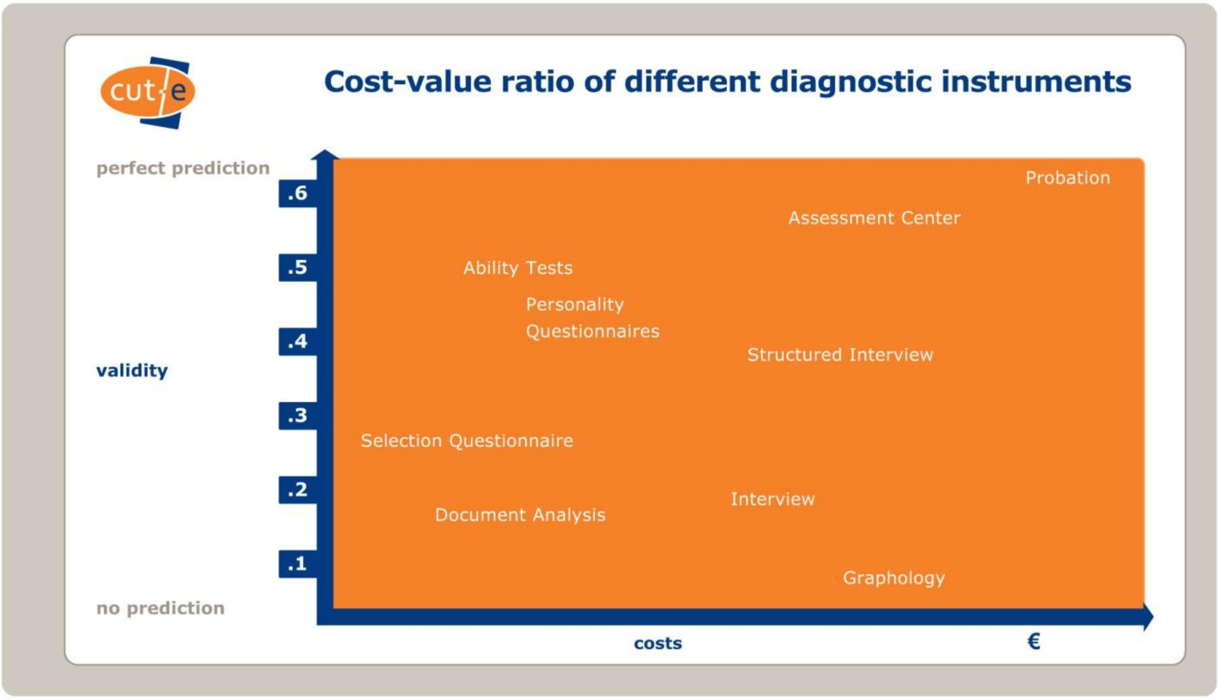

Fig. 2

The most suitable diagnostic instruments provide a good cost-value ratio, so instruments which are low cost and high validity: ability tests and personality questionnaires lead the field.

Probation periods and assessment centres have the highest validity but also the highest costs. Graphology on the other hand, is expensive and provides barely any practical benefit as the validity values are close to zero.

<u>Excursus</u>: Standard error of mean and confidence interval

<u>Standard error of mean</u>

The standard error of mean is a measure of the faultiness of estimation from the test result to the value of the validity criterion. Hence, the standard error of mean specifies the deviation of sample mean from the 'true' mean. The higher the validity of a diagnostic instrument, the lower the standard error of mean and the more accurate the measurement of the test.

<u>Confidence interval</u>

The preciseness of prediction by a test is characterised by a confidence interval in whose borders the 'true' score lies. The confidence interval decreases the more the confidence coefficient decreases (95% or 99%). The higher the standard error of mean, the wider the confidence interval. If the standard error of mean was 0 (this would represent a perfect linear relationship between test results and criterion) the confidence interval would be 0 so that precise predictions would be possible.

Scaling

A test fulfils the criterion ‘scaling’ if the test scores that result from the scoring algorithms adequately represent the empirical relations of characteristics.

In ability tests, this requires that the more effective test candidate receives a better test score than the less effective one; that means that the relation of performance is reflected in the test scores. The practicability of this criterion is dependent on the measurement criterion.

Test economy

Test economy mainly includes the time and costs needed to account for test material, scoring, interpretation, rental fees, etc. The less time needed for preparation, execution and interpretation and the less costs incurred for preparation (purchasing test material, license fees, rental fees), execution and interpretation, the more economic a test is. To justify the use of a test, the information benefit should always be larger than the arising costs.

Standardisation and norms

Diagnostic instruments are beneficial and meaningful only if they are standardised. If an instrument is applied to all persons of a selected and exactly described sample in the same manner and under comparable conditions, this process is called standardisation (adjustment). This data collection of a representative sample under constant hold conditions enables the calculation of norms.

Norms are statistically comparable values which allow comparison of a person’s specific individual test score with the test results of other people of a defined (norm) group to definitely classify and interpret the individual score. The defined group with which the individual scores are compared should share important attributes with the tested individuals (e.g. age, stratum, education, and job). Any good diagnostic instrument should have solid and manifold norm groups available.

The following is critical for the quality of standardisation:

- The choice of the sample on which the test is standardised and adjusted. The norm sample has to be representative for the scope of the test (and for the population).
- Securing the test objectivity during adjustment. Normally, a training of test administrators is necessary.
- The age of the test norms is critical for the correct interpretation of norm or percent values. Norms which were collected more than 10 years ago should be considered out of date.

The single steps for standardisation of tests are the following:

1. Execution of the test at a norm sample which has to be representative for the scope (population) of the test.
2. Calculation of raw scores and distribution of score within the norm sample.
3. Choice of a framework, e.g. IQ-scale.
4. Preparation of exchange tables where a norm score can be read off for each test score. The mean of the norm sample will be equated with the mean of the scale, the deviation of norm sample with the deviation of scale values and so on.

Standardisation techniques

Standardisation techniques which are consulted to relativise a test result normally refer to the distance between the individual test score and the mean of the specific norm sample and express the resulting difference in units of standard deviation of the distribution.

Norms which are based on standard values, including for instance IQ-scores, Z-scores, T-scores, centile scores, stanine scores and percentage scores are especially widely accepted and used.

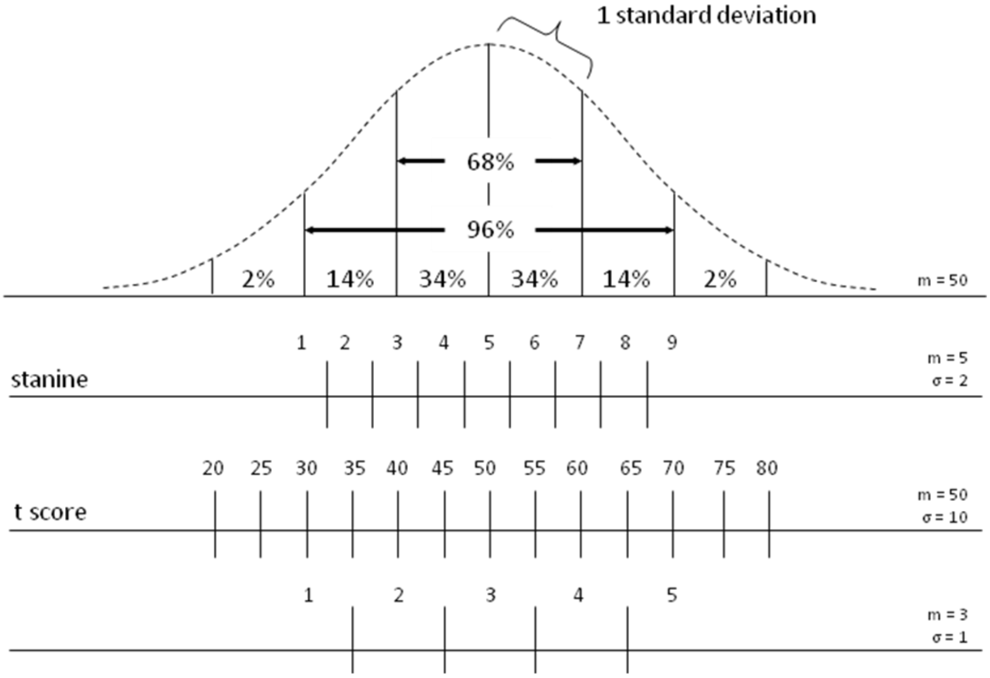

Fig. 3: Comparative illustration of conventional test norm scales

In comparison to linearly transformed scores (Z, T, IQ-scores) the percentage norm is understandable and easy to interpret without having deeper knowledge about the characteristics of the scale (as mean and standard deviation).

A percent score of 50 means that the test result is better than the results of 50% of the compared sample. At the same time it means that it is worse than the results of 50% of the compared sample.

Percentage norms have the drawback that the difference between different percent scores cannot always be interpreted equally. That means that the difference between the 55th and the 60th percent rank is not as large as the difference between the 70th and the 75th percent rank although, seen objectively, 5 percent points lie between the two scores in both cases.

The interpretation of linearly transformed scales is always similar. Generally, scores which spread within 2/3 standard deviations around the mean can be interpreted as average.

Scores which lie more than 2/3 standard deviations below or above mean can be interpreted as below or above average. Scores which lie more than 4/3 standard deviations below or above mean can be interpreted as largely below or above average. Considering the T-scale this would mean that scores between 44 and 56 can be interpreted as average, scores between 38 and 44 and between 56 and 62 as below or above average and scores below 38 and above 62 as largely below or above average.

Usefulness of diagnostic instruments

An essential criterion that should be considered when selecting diagnostic instruments for personnel selection or assessment processes is the usefulness of the respective instrument. A test is useful if the characteristic that it measures possesses practical relevance and if the decisions (procedures) that are based on it are expected to provide more benefit than damage.

The aim of diagnostics is to preferably hire as many suitable und to refuse as many unsuitable applicants as possible. The proportion of suitable candidates appointed as well as the proportion of unsuitable candidates rejected should be preferably high; the aim is to minimise the proportion of unsuitable candidates appointed and suitable candidates rejected.

The success ratio of a diagnostic instrument describes the relationship between suitable candidates appointed (SA) and the total of suitable and unsuitable candidates appointed (SA + UA). The more the number of unsuitably appointed candidate strives against 0, the more the success ratio strives against 1 (100%).

In assessing the usefulness of a diagnostic instrument the following four factors play an important role:

1. Validity
2. Selection ratio
3. Base rate
4. Costs of a wrong decision.

The selection ratio is the rate of individuals who are chosen as appointed out of the applicant population. It presents the relation between the number of appointed applicants to the number of all applicants.

Furthermore, the base rate provides information about the proportion of individuals who would be suitable without the usage of selection strategies. It is calculated as the relation between the number of suitable applicants and the number of all applicants.

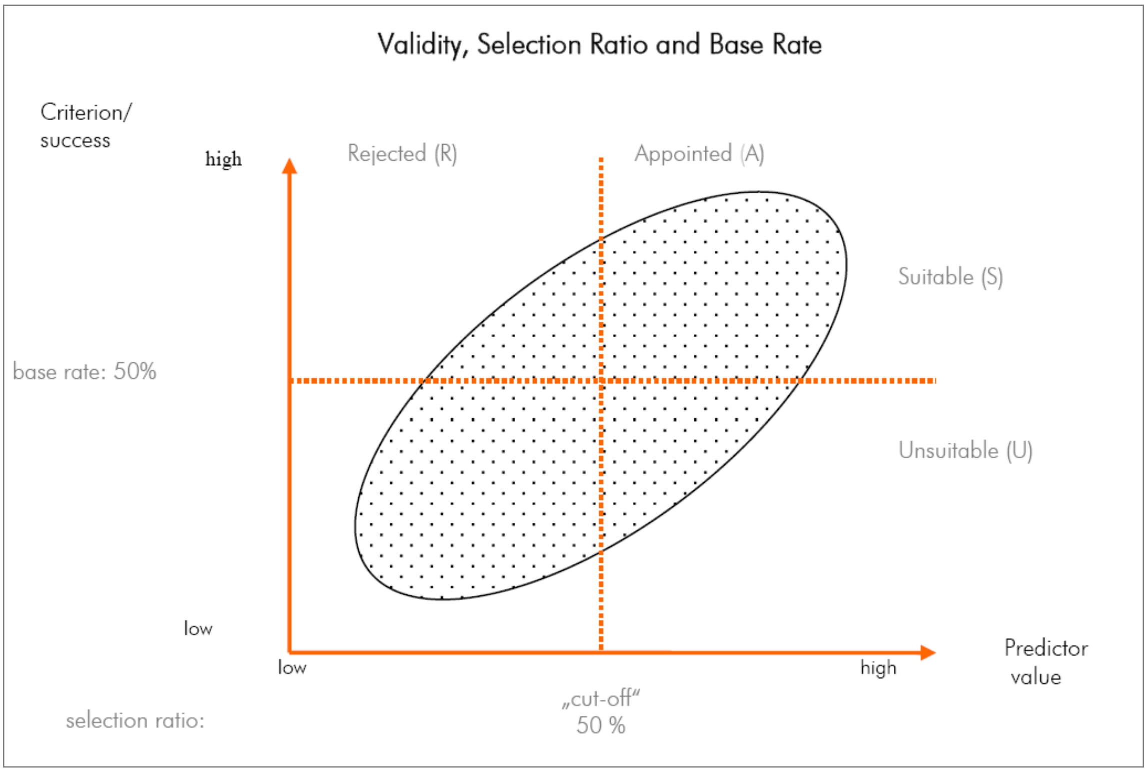

Fig. 4: Validity, Selection Rate und Base Rate

Effect of validity

Assuming selection ratio and base rate are constant and two instruments with different validities are compared, the instrument with the higher validity will have the higher success ratio. This is illustrated in fig. 5: Effect of validity.

The higher the validity, the narrower the scatter plot and the higher the proportion of suitable candidates appointed to all candidates appointed.

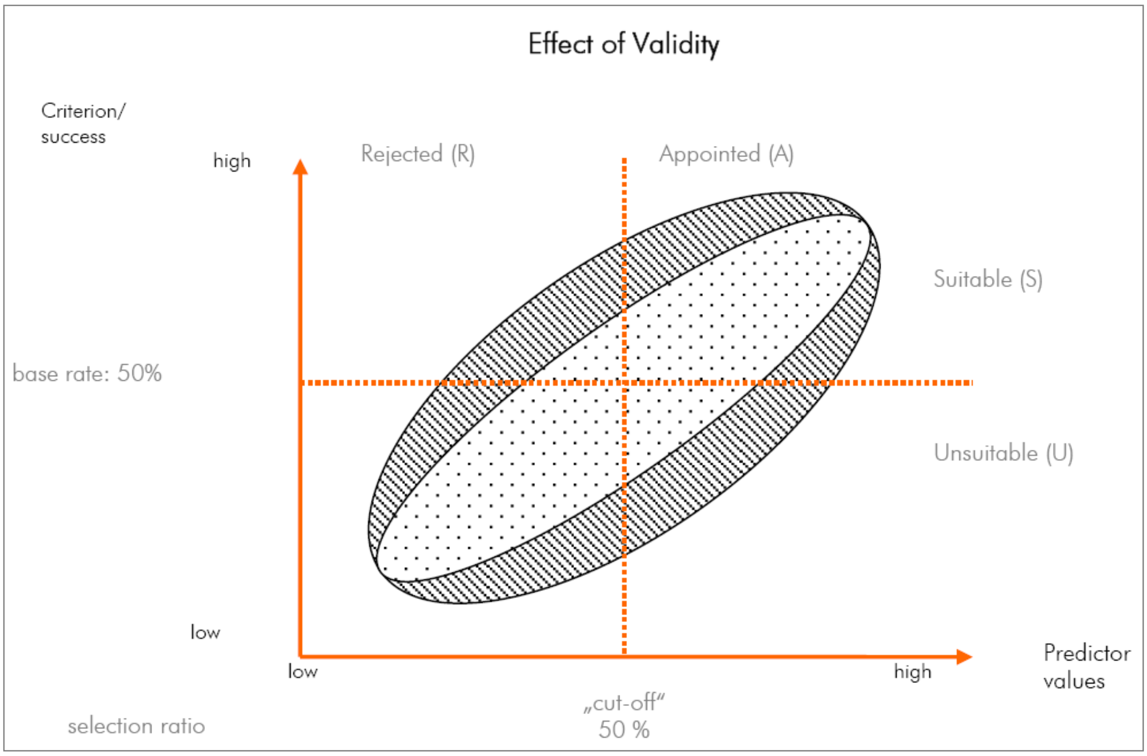

Fig.5: Effect of Validity

- Effect of base rate

If validity and selection ratio are held constant, and the base rate changes, this has an additional impact on the success ratio.

With increasing base rate (that means there are more suitable applicants in the population), the relation between suitable candidates appointed and all appointed candidates clearly improves. The comparison of fig. 6: Effect of base rate, with the previously presented original figure (fig. 4) shows this effect.

However, a high base rate may raise the question as to the usefulness of a diagnostic instrument as the probability to select a suitable applicant is relatively high even with a random selection.

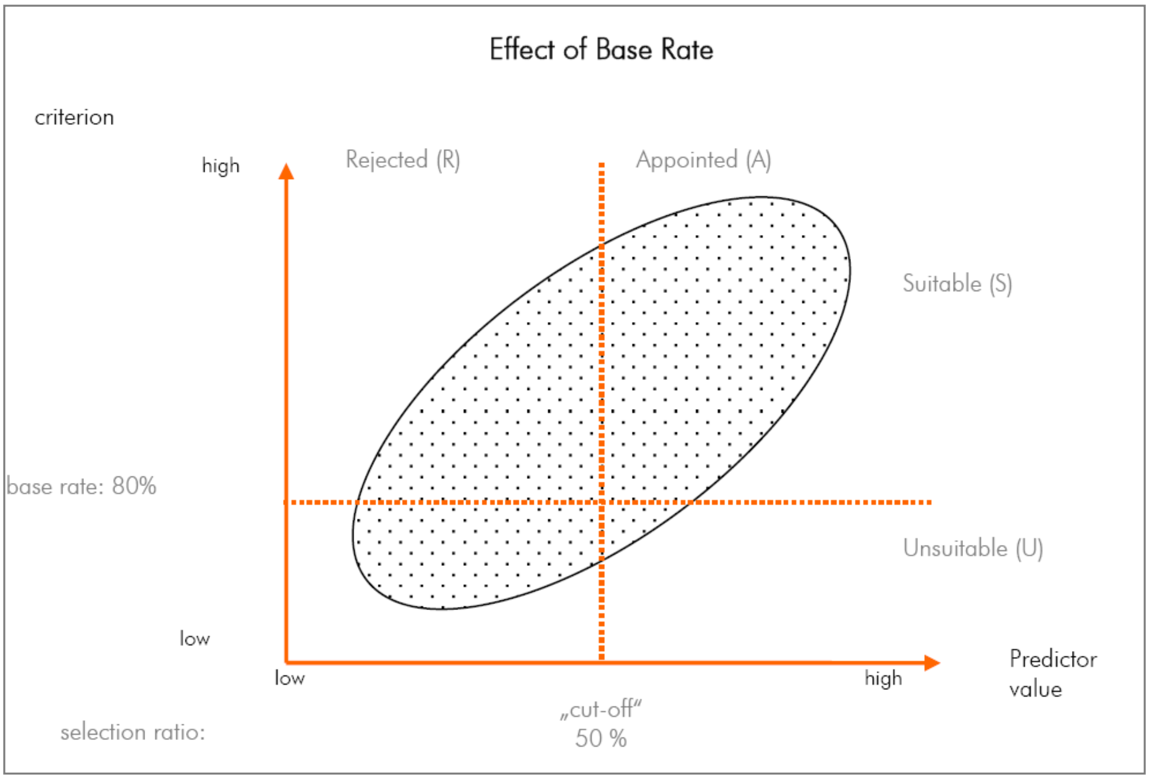

Fig .6: Effect of Base Rate

Effect of selection ratio

If validity and base rate are constant and the selection ratio changes, the success ratio also changes. If the selection ratio is minimised, the success ratio improves as well (see also fig. 7: Effect of selection ratio). In diagnostics, a low selection ratio is often chosen as normally only a small proportion of the applicants will be employed.

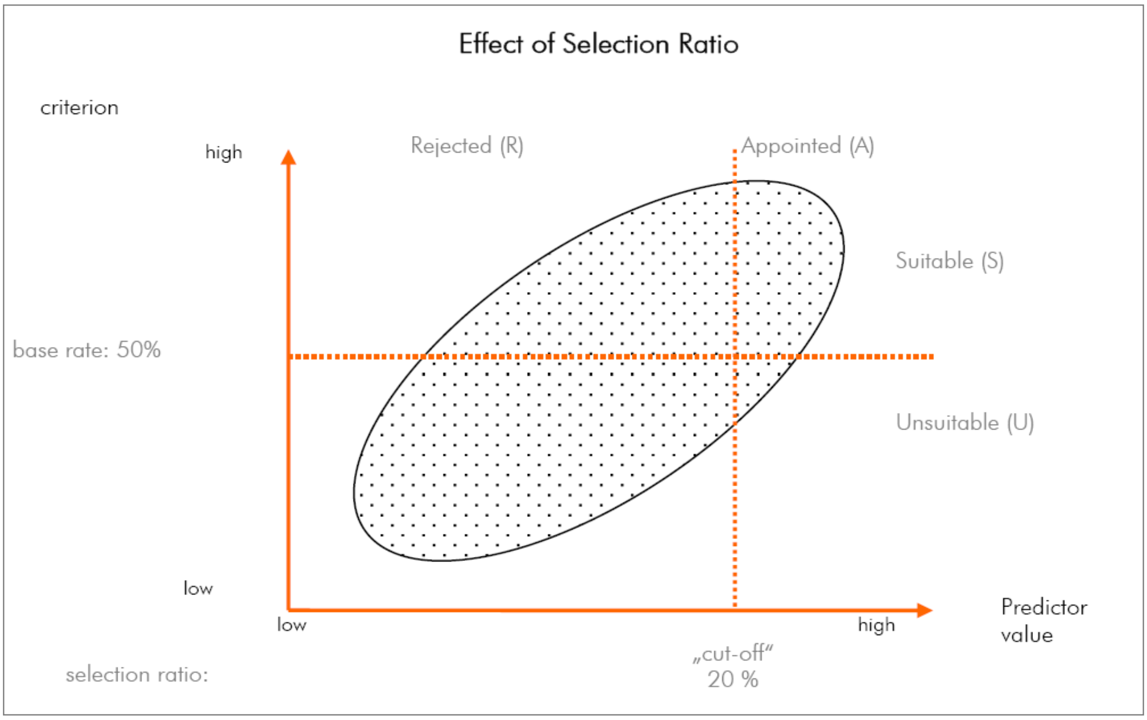

Fig. 7: Effect of Selection Ratio

The practical benefit of diagnostic instruments is particularly high if validity is high, base rate is low (that means that the job requirements are high and the number of unsuitable candidates is low) and if additionally the selection ratio is low.

Besides these three factors another criterion plays a role in selecting the appropriate diagnostic instruments: the cost of a wrong decision and consequently the costs of a wrong hire to the vacancy.

Cost of a wrong hire

If the cost of a wrong hire is extremely high, appropriate diagnostic instruments to select applicants for this position should be used in any case, as the costs for using such instruments are usually much lower than the costs of a wrong hire.

If the costs of a wrong hire are low, the cost-value ratio of a diagnostic instrument has to be weighed up.

Reasonableness

A test is reasonable if it, as measured by its benefit, does not stress the tested person too much with regards to temporal, psychical and physical aspects.

Unfakeablility

If a diagnostic instrument is constructed in a way that the tested person cannot control and distort the concrete values of test scores by targeted testing behaviour, the test can be seen as unfakeable.

High face validity can make a test more fakeable as it is easier to identify the principle of measurement and the tested person may try to influence the test results. Personality questionnaires are more susceptible to such distortions than ability tests.

Transparency and fairness

Furthermore, another essential factor that plays a role in selecting assessment instruments is the treatment of the candidate before, during and after the test.

- Transparency
  A good diagnostic instrument should provide an appropriate degree of transparency. Before the test, understandable instructions explaining the test and the administration process (time limits etc.) should be provided. Practice items explaining the principle of the test or questionnaire should also be available. Additionally, in case of online administrations, a trained test administrator should explain the system.
- The importance of feedback
  Best practice dictates that candidates should receive appropriate feedback about the results of each instrument. The minimum is a short written feedback report. The best is personal feedback provided by a trained person combined with a detailed feedback report.
- Fairness
  Furthermore, it is important in terms of fairness that the resulting test scores do not cause a systematic discrimination against any group based on ethnic, socio-cultural, gender, or other reasons.

  Under accessibility principles, all persons should have the same possibilities to undergo diagnostic instruments. That means, amongst other things, that people with disabilities should also be given the opportunity, according to their abilities, to undergo a diagnostic instrument, possibly by means of input assistance.

Excursus:

Adverse impact

Within the last years, a specific aspect of fairness has come into the focus of the public and of the test practice: equal treatment of applicants.

This states that all applicants - independent of disabilities, gender, marital status, religion, race, skin colour, and nationality or ethnic/national backgrounds - are treated equally and fairly. In the USA, the demand for equal treatment goes to such lengths that enterprises have to fulfil specific quotas in the recruitment of applicants.

Increasingly diagnostic instruments have to substantiate a claim that they do not discriminate against any group of persons.

Criteria for computer-based diagnostics

The exponential increase in the number of diagnostic instruments administered via computer or Internet brings with it new challenges. Companies developing and distributing online tests need to deal with these adequately. In addition to the general criteria outlined previously, there are specific criteria which need to be regarded if test administrations are computer-based.

Good computer-based diagnostic instruments consider both the general and the computer-based specific criteria. Good orientation is again provided in the guidelines of the International Test Commission (ITC). The following criteria can be extracted from the 'International Guidelines on Computer-Based and Internet-Delivered Testing'.

Technology

Technology means ensuring that the technical aspects of computer-based / Internet testing are considered, especially in relation to the hardware and software required to offer and run the testing. The following aspects can be summarised under the criterion technology:

- Giving due regard to technological issues (hardware: processor, graphics card, monitor, etc.; software) in computer-based and Internet testing, at test provider and at test user sites.
- Taking account of the robustness of the computer-based / Internet test. That means that the test should be relatively independent of internet connections etc. and should run stable.
- Considering human factor issues in the presentation of material via computer or the Internet. That means that the test should be as user-friendly as possible.
- Considering reasonable adjustments to the technical features of the test for candidates with disabilities. The test, or its' technical features, should be adaptable and input assistance devices should be possible.
- Providing help, information, and practice items within the computer-based / Internet test.

Quality

Quality means assuring the quality of testing and test materials and ensuring high standards throughout the testing process. The following aspects can be summarised under the criterion quality:

- Ensuring the knowledge, competence and appropriate use of computer-based / Internet testing by provider and user.
- Considering the psychometric qualities of the computer-based / Internet test (see first chapter for detailed information).
- Scoring and analysing computer-based / Internet testing results accurately (e.g. by defined scoring algorithms).
- Interpreting results appropriately.
- Providing appropriate feedback.
- Considering equality of access for all groups.

Control

Control means to control the delivery of tests, candidate and test taker authentication and prior test practice. The following aspects can be summarised under the criterion control:

- Detailing the level of control over the test conditions needed.
- Detailing the appropriate control over the performance of the testing, if needed (open or supervised administration).
- Giving due consideration to controlling prior practice / time of tutorial.
- Giving due consideration to controlling item exposure.
- Controlling test taker's authenticity.
- Ensuring the prevention of cheating (copying of answers, assistance).

Security

Security means to ensure the protection of the testing materials, privacy and data in order to guarantee confidentiality. The following aspects can be summarised under the criterion security:

- Taking account of the security of the editor's and authors' intellectual property.
- Ensuring security of test materials.
- Ensuring the security of test takers' data transferred over the Internet.
- Maintaining the confidentiality and security of test takers' results, e.g. by permitting access only to authorised individuals.

Specific criteria for online diagnostics

Beyond the criteria that apply to computer-based diagnostics there are some aspects that concern especially Internet administered diagnostics. Basically, Internet-based instruments should be self-explainable, forgery-proof, hardware independent, plug-in ready and accessible.

This means in detail:

- **Self-explainable**
  As Internet-based instruments are mostly administered openly, that means uncontrolled, it is very important that they are self-explainable. This means that the test provides interactive examples on which candidates can familiarise themselves with the tasks of the respective instrument and on which they get feedback about the way they dealt with the examples. The explanations for the examples should be as detailed as possible as there is not normally possible to contact the test administrator should questions arise.

- **Forgery-proof**
  As previously mentioned, the administration of Internet-based instruments is often uncontrolled. Therefore, it is important to ensure that the test is forgery-proof. That means that preferably there should be no model answers available which could support the candidate in completing the tasks speedily and correctly. In Internet-based tests this can easily be ensured by generating the test dynamically at runtime. With this approach it is almost impossible that exactly the same test is generated twice.

- **Hardware independent**
  The problem of many Internet-based instruments is that they are often optimised for a certain computer system (certain conditions: screen resolution, proportions) and that they are presented blurred or grainy on other computers. This can be avoided by vectorised item material which optimally adapts to monitor conditions and thus mitigates any effects due to the hardware and Internet connection used.

- **Plug-in ready**
  Regarding Internet-based instruments it is important that seamless integration into existing application management, assessment or HR systems (e.g. SAP, Peoplesoft, Oracle) is possible. This guarantees that all participant data stays within the enterprise and therefore ensures the security of the data.

- **Accessible**
  Each Internet-based instrument should be accessible according to the regulations for users with disabilities. This contains for instance that other input assistance devices besides the mouse (e.g. touch screens, touch pads, keyboard, specific input assistances for motor limited persons) as well as reading assistance devices (e.g. computer internal loupe, magnifier) and other assistance can be used to complete the instruments.

Options for interpretations of results

Points scores

Points scores alone hardly have any explanatory power as a result of a diagnostic instrument. To be comparable and interpretable they have at least to be put in relation to existing norms.

Dichotomous scales, typologies

If the results are presented on dichotomous scales (scales with only two gradings, e.g. typologies: introverted vs. extroverted), as is often the case in personality inventories, this holds some risks. Only few people can definitely be classified in the one or the other category.

The statistical property that someone can definitely be classified in the one (type A) or the other (type B) is 1/2. This only applies if only one dichotomous scale is used. If two dichotomous scales are used for presenting the results, the probability for a definite type classification is 1/4. If three dichotomous scales are used, the probability decreases to 1/8, if four scales are used, it is only 1/16. This means that only 1 out of 16 people can definitely be classified into one type (with four feature characteristics). At the same time this means that 15 out of 16 persons cannot be definitely classified into that typology.

Attempting to assign these 15 people to a definitive type regardless, should be seen very critically, both statistically and theoretically.

Multiple graded scales (e.g. fivefold, sevenfold, ninefold graded scales) and profiles

Concerning the analysis of test data it is assumed that the results are mainly normally distributed.

Most people who complete a test or questionnaire will, according to this, receive median values. Standardised graded scales promise a better interpretation possibility of results, as especially at scales with an odd number of grades the middle is also interpretable.

A profile that is presented by means of graded scales mostly allows a precise and short and at the same time vivid presentation of results.

References

Bartram, D. (2000). Internet Recruitment and Selection: Kissing Frogs to find Princes. International Journal of Selection and Assessment, 8 (4), 261-274.

Bortz, J. (2005). Statistik für Human- und Sozialwissenschaftler (6. Aufl.). Heidelberg: Springer.

ITC (2005). International Guidelines on Computer-Based and Internet-Delivered Testing.

Kersting, M. (2006). DIN SCREEN - Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen. Lengerich: Pabst Science Publisher.

Naglieri, J. A.; Drasgow, F.; Schmit, M.; Handler, L.; Prifitera, A.; Margolis, A.; Velasquez, R. (2004). Psychological Testing on the Internet: New Problems, Old Issues. American Psychologist, 59 (3), 150-162.

Zimbardo, P. G. (1995). Psychologie (6. Aufl.). Berlin, Heidelberg: Springer.

Glossary

Alternate form reliability

Two instruments are parallel if both measure the same characteristic to the same extent. If a parallel form of a test exists, the reliability of the test can be calculated as the correlation between both forms. Thus, the reliability of both tests is the same. To find out the reliability coefficient both test forms are administered in a temporal distance on the same persons and the results of both forms are correlated. The temporal distance is necessary to minimise recall and learning effects as well as other systematic errors.

Base rate

In a sample of applicants the base rate describes the relation of suitable applicants to all applicants. A base rate of 50% therefore means that 50% of all applicants can be seen as suitable.

Coefficients

In mathematics, a coefficient (lat.: coefficere = cause) is a factor which belongs to a certain object such as a variable or a basis vector.

Confidence interval

In statistics, confidence intervals help to estimate the position of a parameter (e.g. of the mean) with a certain probability. It describes an uncertainty range for the estimation of such a parameter. The result of such an estimate depends on the sample. A 95% confidence interval contains the searched score with a probability of 95%.

Construct

A construct is an issue within a scientific theory that cannot be observed immediately. Constructs are of notional and theoretical origin. This does not mean that the certain issue does not ‘exist’; it just means that it has to be forged by other issues that are easy to observe (so-called indicators). Intelligence, for instance, is such a construct.

Construct validity

Construct validity describes the extent to which the measurement procedure captures the theoretical construct that it should measure.

Content validity

Content validity describes the extent to which the measured results immediately represent what should be measured by the test. This is done by analysing the questions and tasks by experts. How far an instrument is content valid is typically decided by means of content consideration of items but not in a mathematical manner. Accordingly, no coefficient representing the content validity of an instrument exists.

Correlation

A correlation is a relation between two or more statistical variables.

Criterion-related validity

The criterion-related validity describes the relation between test results and one or more external criteria. To determine criterion-related validity the test that should be validated, e.g. an intelligence test, is correlated with a criterion value, e.g. success in school.

Dichotomous

Dichotomous means 'cut in two' (Greek), and consequently is the division in two structures or terms. In diagnostics, dichotomous scales are often used. This means that the results of a diagnostic instrument can be presented on a two-split scale. Thus, the tested person can receive a test result that can either be assigned to the one (type A) or the other (type B) end of the scale. Dichotomous scales are often used in personality inventories, e.g. when the task is to distinguish between introverts and extroverts.

Expectancy range

The expectancy range describes an uncertainty range in which the observed scores spread around the true score. It depends on the reliability and the standard error of measurement and decreases with constant security (often 95%) the higher the reliability and the lower the standard error of measurement is.

External validity

External validity means the generalisability of test results. If the test results of an instrument are generalisable across the situation and the tested persons, it is externally valid. Accordingly, external validity depends to a large extent on the naturalness of the testing situation and the representativity of the examined sample.

Face validity

Face validity is the characteristic trait of a diagnostic instrument with which a subjective assumption about the objective of the measurement of the diagnostic instrument can be induced in the tested person. An instrument possesses face validity if the tested person can recognise a direct relationship between the instrument used and the diagnostic question. A test of concentration, for instance, has high face validity if it is used for the selection of data typists.

Generalisability

Generalisability means that the results of a single examination or sample can be assigned to other examinations or samples and that the results are therefore generalisable.

Interfering variables

Interfering variables are effect variables that potentially cause changes in the results of an examination. In diagnostics, the following factors are often supposed to affect the test result: test administrator, testing situation, someone's constitution (physical, psychical).

Internal consistency

The internal consistency is a measure of the reliability of an instrument. To calculate the internal consistency, each test item is considered as an own test part. These test parts are then correlated and provide information about the reliability of a diagnostic instrument.

Internal validity

Internal validity means the degree to which the results of an instrument can be definitely interpreted, that means whether the test result can be attributed only to the measurement procedure or whether there possibly are interfering factors (interfering variables) that could have influenced or distorted the result. The internal validity decreases with the increasing number of uncontrolled interfering variables and increases with the decreasing number of interfering variables.

Norming

Norming (adjustment) means the creation of a numeral frame of reference (normally in tabular form) with whose help individual test scores (e.g. number of correct solutions in a test) can be compared to test scores of a reference population. These tables of comparative values, that means the test scores of the norm sample, are characterised as norms.

There are different norm scales that are customary in practice: the IQ-scale with a mean mx = 100 and a standard deviation sx = 15, the T-score scale (mx=50, sx=10) and the C-score scale (mx=5, sx=2). The Z-score standard scale (mx=0, sx=1) is hardly used in practice, it is used more for statistical transformations.

Objectivity

Objectivity of a diagnostic instrument means that a test result is independent of the testing situation and the test administrator.

Parameter

In statistics, a parameter is one out of (mostly) several scalar values that, together with the distribution category, determines the exact form of a probability distribution. It is often used to describe attributes of a frequency or probability distribution. Examples for parameters are mean and median.

Reliability

The reliability is a measurement of the accuracy and dependence of a diagnostic measurement. Concerning tests, the reliability describes how accurate and dependable the test measures what it should measure. Normally, the accuracy of differentiation between persons is examined. This means that it is measured if a test effectively assigns two different scores to two persons with different values of an attribute or a characteristic.

Representativity

In statistics, representativity describes a basic attribute of statistical examinations: a statistical examination is representative if it is based on a random sample and if it allows conclusions about the universe. The term representativity refers to an examination and not to a sample.

Re-test reliability

To estimate the reliability based on the re-test method, the same test is administered to the same persons in a temporal distance (2 weeks to 1 year, usually 6 to 8 weeks) twice. The correlation of test results can be interpreted as a measure of reliability (time stability) of the test.

Scaling

Scaling is a term based in mathematics that describes a transformation of values. In diagnostics it means the transformation of a test score onto a scale that adequately represents the empirical relation of the characteristics (the relation of good and bad results).

Selection ratio

The selection ratio describes the relation of appointed applicants to the whole pool of applicants. Accordingly, a selection rate of 50% means that 50% of all applicants are (or should be) selected and hired.

Split-half reliability

Calculating split-half reliability, a test is split in two halves. The tasks (items) can be assigned either randomly to the one or the other test half, or the first tasks are assigned to the first and the last tasks are assigned to the second half, or the odd tasks are assigned to the first half and the even tasks are assigned to the second half.

The results, received in the two halves, are then correlated (calculation of relations). As the whole long test is more reliable than the two halves, the split-half reliability of the whole test is calculated on the basis of this correlation by means of the Spearman-Brown formula (correction formula).

Standardisation

Standardisation means, in a literal sense, the unification of measures, types, procedures or other things. In statistics, standardisation means the transformation of differently scaled numerical values into a uniform range of values to be able to compare differently distributed values.

Standardised administration

Standardised administration means that the test results should be independent of the test administrator. In paper-and-pencil tests this is assured by a scoring pattern, in computer based tests by clearly defined (and programmed) rules for the calculation of results. Concerning questions with open answer formats, previously defined categorical systems may help with the interpretation of results.

Standardised interpretation

The test raw score alone is, in most cases, not very meaningful. In most tests, norm tables are available which contain comparative values. The raw scores are, by means of these norm tables, transformed into values that allow the direct comparison with certain norm groups or populations. This guarantees that the test scores are interpreted similarly. Tests providing good norm tables therefore have a well standardised interpretation.

Standardised scoring

Standardised scoring means that the test or questionnaire is administered under certain requirements which have been previously defined. This ensures that the test result is affected as little as possible by external circumstances. This contains e.g. a written instruction at the beginning of the test as well as exactly defined requirements for the test administration (e.g. point in time) and for the arrangement of the testing situation. Standardisation of test material, test instruction and test environment should guarantee the standardised scoring.

Standard error of mean

The standard error of mean is the standard deviation of estimation errors, thus the deviation of the sample mean from the ‘true' mean of the population.

Standard error of measurement

The standard error of measurement is the part of the statistical spread that goes to the account of its reliability. If the reliability has a value of 1, the standard error of measurement would be 0; if the reliability is 0, the standard error of measurement would be 1.

Transformation

A transformation (lat.: the conversion) generally describes the change of figure or form or structure into another figure, form or structure without the loss of matter. In mathematics, transformation means a kind of mapping. In this connection, test raw scores, for instance, are often transformed onto a standardised scale (e.g. T-scale); that means that the raw scores are converted into T-scores and therefore mapped onto a T-scale.

Validity

The validity provides information about the extent to which a test measures what it should measure. Thereby, it can be estimated which purpose a test can fulfil: whether predictions are possible, whether the theoretical construct or the desired characteristic is actually captured etc. The validity provides the degree of accuracy with which the test actually measures the personality characteristic or behaviour pattern that it should measure or pretends to measure.

Only a test that provides a high validity can be interpreted usefully. Therefore, during construction of diagnostic instruments, optimisation of validity is one of the most important (and at the same time one of the most demanding) objectives.

Checklists

General criteria for diagnostic instruments

This checklist can be used for the evaluation of important criteria that play a role in the selection of the appropriate instrument.

Criterion	Sub criterion	Evaluation*	Comment
Objectivity	Standardised administration		
	Standardised scoring		
	Standardised interpretation		
Reliability	Re-test reliability		
	Internal consistency		
	Alternate form reliability		
Validity	Construct validity		
	Criterion-related validity		
	Content validity		
Scaling			
Standardisation			
Norms			
Test economy	Expenditure of time		
	Costs		
Usefulness			
Reasonableness			
Unfakeability			
Transparency	Instructions		
	Feedback		
Fairness	Accessibility		

** Suggestions for evaluation: available / not available; good / bad; ensured / not ensured*

Criteria for computer- and internet-based diagnostic instruments

In computer and Internet-based tests the following criteria, in addition to the general criteria for diagnostics, should be checked:

Criterion	Sub criterion	Evaluation*	Comment
Technology	Compatibility of the test with user technology		
	Independence from Internet speed / robustness		
	Usability		
	Accessibility		
	Practice items / support		
Quality	Scoring algorithms		
	Specified interpretation of results		
	Feedback / report		
Control	Control of administration		
	Control of earlier test practice		
	Control of item exposure		
	Control of authenticity of test candidates		
	Elimination of cheating / distortion		
Security	Security of test materials / copy protection		
	Security of personal data		
	Security of test results		
Forgery-proof	Item generators to generate the instrument at runtime		
Hardware independent	Vectors for adaptation of task presentation		
Plug-in ready	Integration in existing systems		
Accessibility	Use of input assistance		

** Suggestions for evaluation: available / not available; good / bad; ensured / not ensured*

About cut-e

cut-e is a world leader in the design and implementation of online tests and questionnaires for use in recruitment, selection and development of people in the business world.

cut-e assesses over 2 million people per year in over 70 countries and 20 languages.

cut-e combines psychometrics, innovative technology and related consultancy services with an exceptional understanding of business issues to provide personnel and financial benefits for people, companies and organisations.

Among others, we offer the following solutions

- **Online assessment** for the pre-selection of candidates (e.g. trainees, apprentices, sales people)
- **Selection tools** for the computer-based employment testing of miscellaneous specialists and managers
- Online **self-assessment** for self-directed development strategies
- **Online 360° feedback systems**
- **Potential analysis**, **self-assessment** and **management audit**
- **Sales development assessment**
- **Check-up systems** for working knowledge and professional skills

The founders, Andreas Lohff and Dr. Achim Preuss, together have more than 40 years of experience in measurement of quality of human capital. Along with their team, they implement innovative and at the same time scientifically validated selection systems in many projects with top 100 companies worldwide.

With head quarters in Hamburg and a network of implementation partners in Europe, Asia and America, *cut-e* works for organisations such as Siemens, Commerzbank, Lufthansa, UBS, Telekom, PWC, Deloitte and the UN.

cut-e GmbH
Neuer Wall 40
20354 Hamburg
Tel: +49-40-3250.3890
Fax: +49-40-3250.3891
Email: info@cut-e.com

www.cut-e.com