WELCOME!

Assessment and Measurement in Work and Organizational Psychology (HPP8370; 4.0 ECTS)

Methodology of assessment


The origins of WOP

WOP has at least two distinct roots (Arnold, Cooper, Robertson, 2010)

Fitting the man to the job tradition manifests itself in employee selection, training and vocational guidance.

Fitting the job to the man tradition focuses instead on the design of tasks, equipment and working conditions which suit a person's physical and psychological characteristics.
Foundations of psychological measurement

Whatever exists at all exists in some amount
(Thorndike, 1918)

Anything that exists in amount can be measured
(McCall, 1939)

WORK

Stimulus

Perceived stimulus

Reaction

Psychological processes

Stimulus

Stimulus

Perceived stimulus

Reaction

Action, behaviour

Psychological processes

Perception

Attention

Memory

Emotions

Thinking

Stimulus

Stimulus

Perceived stimulus

Reaction

Action, behaviour

Psychological processes

Perception

Attention

Memory

Emotions

Thinking

Stimulus

Stimulus

Perceived stimulus

Reaction

Action, behaviour

Psychological processes

Perception

Attention

Memory

Emotions

Thinking

Historical perspective

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1219</td>
<td>First formal oral examinations in law held at University of Bologna</td>
</tr>
<tr>
<td>1636</td>
<td>Oral examinations for degree certification used at Oxford University</td>
</tr>
<tr>
<td>1879</td>
<td>Wilhelm Wundt – first psychological laboratory University of Leipzig</td>
</tr>
<tr>
<td>1888</td>
<td>James M. Cattell opens testing laboratory at University of Pennsylvania</td>
</tr>
<tr>
<td>1896</td>
<td>Emil Kraepelin proposes new classification of mental disorders</td>
</tr>
<tr>
<td>1896</td>
<td>Hermann Ebbinghaus develops first completion test</td>
</tr>
<tr>
<td>1904</td>
<td>Charles Spearman describes two-factor theory of mental abilities</td>
</tr>
<tr>
<td>1905</td>
<td>E.L. Thorndike’s textbook “Introduction to the theory of mental and social measurement”</td>
</tr>
<tr>
<td>1905</td>
<td>Binet-Simon Intelligence Scale has published</td>
</tr>
</tbody>
</table>
### Historical perspective

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1908-1914</td>
<td>E.L. Thorndike develops standardized tests of arithmetic, handwriting, language, spelling</td>
</tr>
<tr>
<td>1914</td>
<td>Arthur Otis develops first group test of intelligence (based on Terman’s Stanford Revision of the Binet-Simon Scales)</td>
</tr>
<tr>
<td>1916</td>
<td>Stanford-Binet Intelligence Scale</td>
</tr>
<tr>
<td>1917</td>
<td>Army Alpha &amp; Army Beta</td>
</tr>
<tr>
<td>1920</td>
<td>Robert Woodworth’s first standardized personality inventory</td>
</tr>
<tr>
<td>1923</td>
<td>Stanford-Binet Intelligence Scale</td>
</tr>
<tr>
<td>1926</td>
<td>Scholastic Aptitude Test</td>
</tr>
<tr>
<td>1935</td>
<td>IBM test-scoring machine</td>
</tr>
<tr>
<td>1939</td>
<td>Wechsler-Bellevue Intelligence Scale</td>
</tr>
<tr>
<td>1942</td>
<td>Minnesota Multiphasic Personality Inventory (MMPI)</td>
</tr>
<tr>
<td>1949</td>
<td>Wechsler Intelligence Scale for children</td>
</tr>
<tr>
<td>1970</td>
<td>Increasing use of computers in designing, administering, scoring, analyzing, interpreting tests</td>
</tr>
<tr>
<td>1981</td>
<td>Wechsler Adult Intelligence Scale (WAIS)</td>
</tr>
<tr>
<td>1989</td>
<td>MMPI-2</td>
</tr>
<tr>
<td>1990</td>
<td>Wechsler Intelligence Scale for Children 3</td>
</tr>
<tr>
<td>1992</td>
<td>The Big Five Personality Test (Costa &amp; McCrae)</td>
</tr>
<tr>
<td>1994</td>
<td>Occupational Stress Indicator (OSI-1) Cary L. Cooper</td>
</tr>
<tr>
<td>1996</td>
<td>Occupational Stress Indicator (OSI-2)</td>
</tr>
<tr>
<td>1997</td>
<td>WHODQOL 100 and WHODQOL-Brief</td>
</tr>
<tr>
<td>1999</td>
<td>Job-related Affective Well-being Scale (JAWS) Paul T. Van Katwyk, Suzy Fox, Paul E. Spector, E. Kevin Kelloway</td>
</tr>
<tr>
<td>2000</td>
<td>Work-Family Conflict Scale (Carlson, Kacmar, Williams)</td>
</tr>
<tr>
<td>2002</td>
<td>Counterproductive Work Behavior Checklist (CWB-C) Paul E. Spector</td>
</tr>
<tr>
<td>2003</td>
<td>Factual Autonomy Scale (FAS) Paul E. Spector and Suzy Fox</td>
</tr>
<tr>
<td>2008</td>
<td>Violence Prevention Climate Survey Stacey R. Kessler, Paul E. Spector, and Chu-Hsiang Chang</td>
</tr>
<tr>
<td>2011</td>
<td>Organizational Citizenship Behavior Checklist (OCB-C) Suzy Fox and Paul E Spector</td>
</tr>
</tbody>
</table>
Most frequently used tests in counseling
(Watkins, Campbell, McGregor, 1988)

- Minnesota Multiphasic Personality Inventory (MMPI)
- Wechsler Adult Intelligence Scale (WAIS)
- Strong–Campbell Interest Inventory
- Wechsler Intelligence Scale for children
- House-Tree-Person Test
- Draw-a-Person Test
- Thematic Apperception Test
- Rorschach Inkblot Test
- Sixteen Personality Factor Questionnaire, Cattell 16PF // Big Five
- Wide Range Achievement Test
- Edwards Personal Preference Schedule
- Kuder Occupational Interest Survey
- Wechsler Memory Scale
- Differential Aptitude Test

Testing as profession

<table>
<thead>
<tr>
<th>Test</th>
<th>Qualification</th>
</tr>
</thead>
<tbody>
<tr>
<td>A level</td>
<td>May be purchased by schools and other organizations and by individuals certified or licensed to administer those tests</td>
</tr>
<tr>
<td>B level</td>
<td>Requires a master’s degree in psychology or education, equivalent training relevant to assessment, or membership in professional association that requires appropriate training in assessment of its members</td>
</tr>
<tr>
<td>C level</td>
<td>Requires a PhD in psychology or education, the equivalent in training in assessment, or verification of licensure or certification requiring appropriate training and experience in psychological assessment</td>
</tr>
</tbody>
</table>

Ethical Codes

American Psychological Association (APA) and American Personnel and Guidance Association (APGA) have ethical codes cover of test administration, standardization, reliability, validity.

Ethical Principles of Psychologists and Code of Conduct (APA, 1992)

Ethics and Values in Industrial – Organizational Psychology (LeRowitz, 2003)

Ethical Codes of member Associations. Meta-Code of Ethics (EFPA, 2005)
Standards for Educational and Psychological Testing (APA, 1985, 1999)

Test Construction, Evaluation, and Documentation
1. Validity
2. Reliability and Errors of Measurement
3. Test Development and Revision
4. Scales, Norms, and Score Comparability
5. Test Administration, Scoring, and Reporting
6. Supporting Documentation for Tests

Fairness in Testing
7. Fairness in Testing and Test Use
8. The Rights and Responsibilities of Test Takers
9. Testing Individuals of Diverse Linguistic Backgrounds
10. Testing Individuals with Disabilities

Testing Applications
11. The Responsibilities of Test Users
12. Psychological Testing and Assessment
13. Educational Testing and Assessment
14. Testing in Employment and Credentialing
15. Testing in Program Evaluation and Public Policy

Assessment in WOP-P

Personnel selection:
China was the first country in the world that implemented a nationwide standardized test, which was called the imperial examination. The main purpose of this examination was to select for able candidates for specific governmental positions (in 605 AD)
Assessment in WOP

Now for:
• Personnel selection
• Job/work/task analysis
• Personnel classification and placement
• Performance testing
• Vocational interests
• Training and re-training
• Career planning and developing
• Appraisal of employees
• Consumer behavior

Test design and construction
Planning a test
✓ What is an aim of testing or why we will test at all?
✓ For what we need the results to use?
✓ Which test(s) are most useful for our purposes?
✓ What are the specific topics on which employees are to be tested?
✓ What kind of questions should be constructed?
✓ What kind of response scales should be used?
✓ What items and test formats or layouts should be used?
✓ When, where and how should test be given?
✓ How should the completed test papers be scored and evaluated?
✓ How should the competed tests be hold or in what level of security?

Test design and construction
Preparing test items
1. Essay items
   Directions: Write a half-page answer to each question

2. Objective items
   A. Short answer
      Directions: Write the appropriate word in each blank
      Example: Test is the best tool for …..

   B. True-false
      Directions: Circle T if the statement is true or F if it is false.
      Example: T ✗ F 1. Test is useful only in education
Test design and construction

Preparing test items
C. Matching
Directions: Write the letter corresponding to the correct name in the appropriate marginal dash

<table>
<thead>
<tr>
<th>Inkblot test</th>
<th>A. Binet</th>
<th>B. Rorschach</th>
<th>C. Newton</th>
</tr>
</thead>
</table>

Preparing test items
D. Multiple choices
Directions: Write the letter of the correct option in the marginal dash opposite the item

<table>
<thead>
<tr>
<th>Occupational stress is</th>
<th>A. emotion</th>
<th>B. pressure</th>
<th>C. satisfaction</th>
</tr>
</thead>
</table>

Test design and construction

Preparing test scales
The format of a typical five-level Likert scale:
1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

Six-point scale is used and this is a forced choice scale:
1. Very definitely is not a source of pressure
2. Definitely is not a source of pressure
3. Generally is not a source of pressure
4. Generally is a source of pressure
5. Definitely is a source of pressure
6. Very definitely is a source of pressure
Test design and construction

Assembling a test
- Is the length of the test appropriate for time limits?
- How should the items be grouped or arranged on the pages of the test booklet?
- Are answers to be marked in the test booklet or is a special answer sheet to be used?
- What socio-demographic information is necessary?
- Is the testing anonymous? If yes, whether and how anonymity is guaranteed?

Test design and construction

Oral, written and E-testing

Oral testing is defined as an evaluation situation in which responses to questions are given orally by examinees.

Test administration, scoring, and items analysis

Standards for Educational and Psychological Testing (APA, 1985)

Chapter 15, Test Administration, Scoring, and Reporting

Standard 15.1.
In typical applications, test administrators should follow carefully the standardized procedures for administration and scoring specified by the test publisher. Specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment should be strictly observed. Exemptions should be made only on the basis of carefully considered professional judgment, primarily in clinical applications.
Test administration, scoring, and items analysis

Standard 15.2

The testing environment should be one of reasonable comfort and with minimal distractions. Testing material should be readable and understandable. In computerized testing, items displayed on a screen should be legible and free from glare, and the terminal should be properly positioned.

Comment:

Testing sessions should be monitored where appropriate both to assist the test taker when a need arises and to maintain proper administrative procedures. Among the conditions that should be avoided in testing situations are:

- Noise,
- Disruption in the testing area,
- Extremes of temperature,
- Inadequate work space,
- Unlegible material, and so forth.

In the context of computer-administered tests, the novelty of the presentation may have an unknown effect on the test administration.

Test administration, scoring, and items analysis

International Test Commission’s (ITC) International Guidelines on Test Use (Bartram, 2001; ITC, 2001)

EFPA-EAHOP Working Group modified and expanded the ITC Guidelines into a more detailed format.

The format adopted is as follows:

<table>
<thead>
<tr>
<th>Unit 1</th>
<th>Task responsibility for ethical test use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard 1.1</td>
<td>Act in a professional and ethical manner</td>
</tr>
<tr>
<td>Standard 1.2</td>
<td>Ensure you have the competence to use tests</td>
</tr>
<tr>
<td>Standard 1.3</td>
<td>Take responsibility for the results of tests</td>
</tr>
<tr>
<td>Standard 1.4</td>
<td>Ensure that test materials are kept securely</td>
</tr>
<tr>
<td>Standard 1.5</td>
<td>Ensure that test results are treated confidentially</td>
</tr>
</tbody>
</table>

Occupational Context:

Testing of people for

- Recruitment and selection
- Organizational or team fit
- Identification of development needs
- Career advice and guidance
- Other organizational purposes

Testing of groups for

- Team development
- Competence audit
- Other organizational purposes

Testing methods

- Psychological tests of ability (maximum performance measures)
- Psychological tests of personality, motivation etc. (typical performance measures)

Administration modes

- Individual vs. Group
- Paper-based, equipment-based and multi-media testing
- Online (internet/intranet) vs offline computer-based testing
- Open, Controlled, Supervised or Managed administration
Knowledge required in Unit I

<table>
<thead>
<tr>
<th>Theories, models and principles of:</th>
<th>Facts, data and information about:</th>
<th>Methods, techniques and procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethics and principles of good practice in testing including:</td>
<td>EFPA Meta-code of ethics</td>
<td>Assessment needs analysis</td>
</tr>
<tr>
<td>Informed consent</td>
<td>Psychological tests</td>
<td>Testing of people for:</td>
</tr>
<tr>
<td>Privacy and confidentiality</td>
<td>Good practice guides:</td>
<td>Recruitment and selection</td>
</tr>
<tr>
<td>Ethics and principles of good practice in client relationships</td>
<td>• ITC Guidelines on Test Use</td>
<td>Organizational or team fit</td>
</tr>
<tr>
<td>Principles of good practice and conduct regarding relationships with paying clients</td>
<td>• ITC Guidelines on Computer-based testing and testing on the Internet/intranet</td>
<td>Identification of development needs</td>
</tr>
<tr>
<td>Professional ethics</td>
<td>• ITC Guidelines on test adaptation</td>
<td>Career advice and guidance</td>
</tr>
<tr>
<td></td>
<td>Local and national good practice, guidelines relating to testing and test use</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Law and Codes of Practice relating to protection of personal data and intellectual property rights in relation to individual testing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Law and Codes of Practice relating to equal opportunities, direct and indirect discrimination, employment law</td>
<td></td>
</tr>
</tbody>
</table>

Physical conditions

- Seating
- Lighting
- Ventilation
- Temperature
- Noise level
- Other physical conditions

Scientific requirements for test

- Reliability
- Validity
- Standardization
- Norms
- Opportunity to repeat the test
Classical measurement theory

Classical test theory is a body of related psychometric theory that predicts outcomes of psychological testing such as the difficulty of items or the ability of test-takers. The aim of classical test theory is to understand and improve the reliability of psychological tests.

Classical Test Theory assumes that each observed score (X) contains a True component (T) and an Error component (E).


Reliability

Reliability is used to describe the overall consistency of a measure.

Reliability - within the limits of measurement failure (error) to provide similar results if the conditions are not changed.

A measure is said to have a high reliability if it produces similar results under consistent conditions.

The goal of reliability theory is to estimate errors in measurement and to suggest ways of improving tests so that errors are minimized.

Reliability

There are several general classes of reliability estimates:

Inter-rater reliability assesses the degree to which test scores are consistent when measurements are taken by different people using the same methods.

Test-retest reliability assesses the degree to which test scores are consistent from one test administration to the next.

Measurements are gathered from a single rater who uses the same methods or instruments and the same testing conditions.
Reliability

Test-retest reliability directly assesses the degree to which test scores are consistent from one test administration to the next. It involves:

- Administering a test to a group of individuals
- Re-administering the same test to the same group at some later time
- Correlating the first set of scores with the second

The correlation between scores on the first test and the scores on the retest is used to estimate the reliability of the test using the Pearson correlation coefficient (R).

Reliability

Internal consistency reliability assesses the consistency of results across items within a test. Internal consistency assesses the consistency of results across items within a test.

1. The most common internal consistency measure is Cronbach's α (alpha), which is usually interpreted as the mean of all possible split-half coefficients. Cronbach's α (alpha) is a coefficient indicating the internal consistency of measures and is commonly used for the estimation of the reliability of the psychometric tests.

The acceptable internal consistency assessed with the widely accepted Cronbach's α (alpha) standard of 0.70 (Nunnally, 1978).

Reliability

2. Parallel-forms method

The key to this method is the development of alternate test forms that are equivalent in terms of content, response processes and statistical characteristics. For example, alternate forms exist for several tests of general intelligence, and these tests are generally seen equivalent.

With the parallel test model it is possible to develop two forms of a test that are equivalent in the sense that a person's true score on form A would be identical to their true score on form B. If both forms of the test were administered to a number of people, differences between scores on form A and form B may be due to errors in measurement only.
Reliability

Example: Parallel-forms method – test A and test B are both measurements for work satisfaction

1. Ants
2. Maali
3. Tiit
4. Tõnu
151. N

Reliability

3. Split-half method
This method treats the two halves of a measure as alternate forms. It provides a simple solution to the problem that the parallel-forms method faces: the difficulty in developing alternate forms. It involves:
- Administering a test to a group of individuals
- Splitting the test in half
- Correlating scores on one half of the test with scores on the other half of the test
The correlation between these two split halves is used in estimating the reliability of the test

Reliability

4. Test-retest reliability method: directly assesses the degree to which test scores are consistent from one test administration to the next.
It involves:
- Administering a test to a group of individuals
- Re-administering the same test to the same group at some later time
- Correlating the first set of scores with the second
The correlation between scores on the first test and the scores on the retest is used to estimate the reliability of the test using the Pearson coefficient (R)
Reliability
Example: Test-retest reliability method – test A is carried out twice in the same group

Monday

TEST

1. Ants
2. Maali
3. Tiit
4. Tõnu

Tuesday

TEST

151. N

Saturday

TEST

Validit
Validity means that test have to measure exactly this phenomenon, for which the test is designed (not anything else or another phenomenon)

In psychometrics, validity has a particular application known as test validity: "the degree to which evidence and theory support the interpretations of test scores" ("as entailed by proposed uses of tests") (APA, 1999)
Validation

Internal validity is an inductive estimate of the degree to which conclusions about causal relationships can be made (e.g., cause and effect), based on the measures used, the research setting, and the whole research design. Good experimental techniques, in which the effect of an independent variable on a dependent variable is studied under highly controlled conditions, usually allow for higher degrees of internal validity than, for example, single-case designs.

External validity concerns the extent to which the (internally valid) results of a study can be held to be true for other cases, for example to different people, places or times. In other words, it is about whether findings can be validly generalized. If the same research study was conducted in those other cases, would it get the same results?

Standardization

Standardization means that tests have to allow the use of different organizations, occupations, cultures etc.

A standardized test is a test that is administered and scored in a consistent, or "standard", manner. Standardized tests are designed in such a way that the questions, conditions for administering, scoring procedures, and interpretations are consistent and are administered and scored in a predetermined, standard manner (Popham, 1999).

Any test in which the same test is given in the same manner to all test takers is a standardized test.

Norms

A set must be representative (female/male, education, age etc.)

For example: Test results depending on age and education

SAMPLE

<table>
<thead>
<tr>
<th>EDUCATION</th>
<th>AGE (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25-30</td>
</tr>
<tr>
<td>Primary school</td>
<td>33,3</td>
</tr>
<tr>
<td>Secondary school</td>
<td>33,3</td>
</tr>
<tr>
<td>Higher education</td>
<td>33,3</td>
</tr>
</tbody>
</table>
Norms and Test Scores

Many major psychological measures are norm-based, meaning that the score for an individual is interpreted by comparing his/her score with the scores of a group of people who define the norms for the test.

1. The most familiar of these measures is the arithmetic average, more technically known as the mean (M).
2. Another measure of central tendency is the mode, or the most frequent score.
3. A third measure of central tendency is the median, or middle score of the distribution.

To organize and summarize data for normative purposes, begin by grouping the data into a frequency distribution. The information provided by frequency distributions can be presented graphically in the form of a bell-shaped normal distribution curve, as long as it approximates that symmetrical form.

A group of scores can be summarized by a measure of central tendency. For example: IQ

<table>
<thead>
<tr>
<th>Type of norms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>National</td>
<td>Derived from a standardization sample nationally representative of the population of interest, regional, local norms.</td>
</tr>
<tr>
<td>Age</td>
<td>Age-equivalent scores and grade (used to indicate the average test performance of test takers in a specific grade) norms.</td>
</tr>
<tr>
<td>Subgroup norms</td>
<td>Are created when narrowly defined groups are sampled.</td>
</tr>
<tr>
<td>Percentile norms</td>
<td>Consist of a table of percentages corresponding to particular raw score.</td>
</tr>
</tbody>
</table>
Testing

Test the development, reliability and validity evaluation, as well as testing itself is a task of specially trained psychologist, not the responsibility of other specialty professionals e.g. managers, teachers, doctors, etc.

Testing shall conform to both the company and the employee interests.

Recommendations for further reading


Thank You!

mare@pekonsult.ee