



## Achievement Motivation

### THE IMPORTANCE OF ACHIEVEMENT MOTIVATION

Human life can be described as a continuous work at tasks. Individuals may or may not be successful in facing these tasks. The psychology of achievement motivation is engaged to run research projects aiming at a better understanding of individual performance and the nature of human resources as well as at the development of assessment and intervention techniques to increase achievement motivation. Tasks in industrial settings and in service organizations become more and more complex and underlie dynamic changes arising from changing market demands. To keep individuals highly achievement motivated while doing their jobs, tasks have to be designed with high motivating potentials.

From a motivational perspective the action process is divided into two parts. The first part describes the development of achievement motivation as a consequence of a fit between the achievement motive and the achievement-oriented motivating potentials of the situation. *Achievement motivation* initiating action arises through interaction of achievement-oriented *motivating potentials* of the task in its situational context and the strength of the *achievement motive* on the side of the performing person. Personal goals controlling actions result directly from the strength of this achievement motivation (Figure 1). The second

part of the motivation process responsible for the translation of motivation into action is often called the volitional phase in the control of behaviour (Heckhausen, 1989); during this phase, goal-oriented action turns into outcomes controlled by the degree of *goal commitment*. Goal commitment affects the way persons choose to reach their goals and the selection of strategies they pursue (Brandstädter & Renner, 1990). Examples for such strategies are to pursue a goal persistently even in case of hindrances or to adapt flexibly to changing aspects of the situation. The translation process works better when more specific and concrete goals are set, the higher the goal commitment the more effective the chosen strategies of goal pursuit (Vroom, 1964; Locke & Latham, 1990; Kleinbeck, in press).

A goal-oriented course of action immune to disturbances is especially supported by specific and concrete goals (goal characteristics; Figure 1).

Because of the many single concepts subsumed under the label of achievement motivation, it is necessary to develop as many measurement tools as possible to differentiate between the concepts. Outside current research projects, measures of achievement motivation are principally used in industrial setting, in service organizations and in educational fields. Here achievement motivation measurement is used to investigate the motivating

## 2 Achievement Motivation

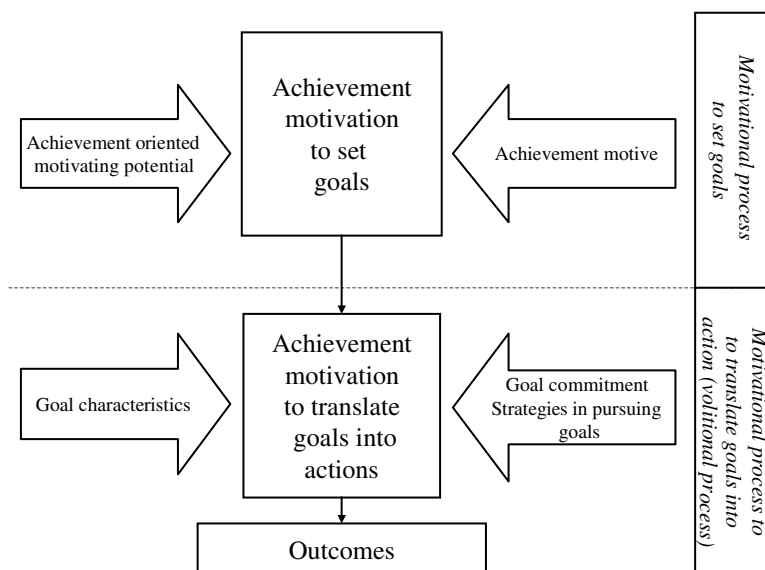


Figure 1. Components of achievement motivation.

potentials of work tasks and work contexts to make full use of individual resources.

### INSTRUMENTS TO ASSESS ACHIEVEMENT MOTIVATION

The theory of achievement motivation describes performance as multidimensional and as influenced by many different factors. The main personal factor is the achievement motive; the main task-specific factor is the motivating potential of the situation. For diagnostic information about mode and strength of the achievement motive there are three different sources (see Schneider & Schmalz, 2000: 50–56):

- 1 Self-judgement
- 2 Judgement by others
- 3 Behavioural indices

Assessing the strength of the achievement motive different strategies are used according to these sources: operant procedures (e.g. the Thematic Apperception Test – TAT) and respondent procedures (e.g. questionnaires), and the grid technique that according to Schmalz (1999) lies in its methodological background between the first two types of measurement. Due to this fact, one can differentiate implicit and explicit components of the achievement motive. Using the material of the

TAT with pictorial presentations of situations it becomes possible to penetrate implicitly into the achievement motive system, because this kind of measurement allows one to approach materials of memory relevant for the motive system. Filling out questionnaires requires ego involvement, self-insight and self-reflection, and also explicit memory, because the answers to the questions can only be given with the help of conscious reflection to earlier experiences (Graf & Schacter, 1985: 501).

Schmalz & Sokolowski (2000) discuss the quality of the different techniques to measure the achievement motive and conclude that all available instruments work reliably. TAT and the grid technique have comparable and widely diversified validity ranges that are related to respondent and operant behaviour. Questionnaires used to diagnose motives seem to be specialized to predict respondent behaviour and conscious experiences (Spangler, 1992).

To measure the achievement-oriented motivating potentials of tasks, Hackman & Oldham (1965) developed and presented an instrument, the Job Diagnostic Survey (JDS), that has well proven its validity (Fried & Ferris, 1987). The JDS measures the motivating potentials of tasks in work situations and also of tasks that students are confronted with in learning situations (Schmidt & Kleinbeck, 1999). Measuring the achievement motive and the

motivating potentials of tasks allows one to determine the strength of achievement motivation.

Rheinberg, Vollmeyer & Burns (in press) present an instrument to measure achievement motivation as a comprehensive construct. With 18 items, four components of the current state of achievement motivation are measured: (1) fear of failure; (2) probability of success; (3) interest; (4) challenge. In its German and English version, the instrument shows satisfying consistencies and according to the first validation data, the measured components of current achievement motivation correlate positively with learning behaviour and performance. Schuler & Prochaska (2001) define achievement motivation as a general behavioural orientation. The instrument they developed – the Hohenheim Test of Achievement Motivation (HTML) – allows measuring achievement motivation with 17 scales in a highly differentiated way. The results of the HTML measures correlate significantly with neuroticism and conscience in the five-factor model of personality (Costa & McCrae, 1980). Measures in HTML are positively related to success at school, university and work so that one can expect a successful application in personality research and in educational and occupational testing.

To measure goal characteristics (e.g. goal specificity and goal difficulty) that influence the

achievement-oriented process of translating goals into action, Locke & Latham (1984, 1990) present a questionnaire that has been used mainly in research settings. Other questionnaires try to measure clarity of tasks and goals (Sawyer, 1992), clarity of methods (Breaugh & Colihan, 1994; Schmidt & Hollmann, 1998) and also clarity of performance judgements (Breaugh & Colihan, 1994; Kleinbeck & Fuhrmann, 2000). These components of achievement motivation measured by the mentioned questionnaires affect the motivation to translate goals into action and as a consequence performance outcome.

Recently researchers began to measure goal commitment (Hollenbeck et al., 1989). They invested considerable effort because goal setting is no homogeneous construct. As Tubbs (1993) could show there are three different components of goal commitment: the first component has to do with processes of weighing and evaluating the potential goals. During these processes, one calculates mainly values and expectancies that affect the strength of motivational tendencies for specific goals. The second component contains the result of these evaluative processes focussing on calculations of values and expectancies and leading to setting a personal goal. This component is also related to the decision to attain this

Table 1. Instruments for measuring components of achievement motivation

Instruments	Author	Concepts measured	Method used
TAT	Murray, 1943; McClelland et al. 1953	Achievement motive and other motives	Content analysis of stories (operant)
OMT	Kuhl, 2000	Achievement motive	Content analysis of written stories (operant)
MARPS Grid-technique	Mehrabian, 1968 Schmalt & Sokolowski, 1999	Achievement motive Achievement motive	Questionnaire (respondent) Judgement of fit between pictures and motive- related statements
Questionnaire for current motivational states	Rheinberg et al., 2000	Current motivation for learning and performance	Questionnaire (respondent)
AVEM	Schaarschmidt & Fischer, 1996	Current work motivation	Questionnaire (respondent)
HLMT JDS	Schuler & Prochaska, 2001 Hackman & Oldham, 1965	Achievement motivation Motivating potential of tasks	Questionnaire (respondent) Questionnaire (self-judgement and judgement by others)
Fragebogen für Zielcharakteristika	Locke & Latham, 1990	Goal specificity and others	Questionnaire (respondent)
Goal commitment Strategies of goal pursuit	Hollenbeck et al., 1987 Brandtstädter & Renner, 1990	Goal commitment Strategies of goal pursuit	Questionnaire (respondent) Questionnaire (respondent)

particular goal. The third component of goal commitment is characterized by maintaining the set goal and by staying persistent even when faced with hindrances. Future research will show whether it will be possible to develop differentiated measurement procedures on the basis of these considerations.

With respect to goal commitment in goal-oriented action, people seem to be able to use stable dispositions. They either persist tenaciously in pursuing their goals or they adjust flexibly to new or other goals. Brandtstädter & Renner (1990) described two scales to measure 'tenacious goal pursuit' and 'flexible goal adjustment'. Their results show relations between these different strategies and age. Older people adapt more often flexibly instead of pursuing their goals tenaciously against hindrances. Table 1 summarizes the instruments for measuring components of achievement motivation.

#### FUTURE PERSPECTIVES

The current state of research can be described as presenting a set of different measurement approaches for the central components of achievement motivation. Future tasks for research and applications mainly in work and educational settings will be to determine the range of validity for the different measures more exactly. This can help to decide under what circumstances specific instruments can be used profitably. Although there are now some reliable and valid instruments to measure single components of achievement motivation, it would be helpful to have new instruments and procedures to relate them to each other.

#### CONCLUSIONS

A high achievement motivation in people guarantees success and wealth in human societies. To produce adequate conditions for the development of a high achievement motivation it is necessary to understand how achievement motivation is formed and how it can be translated into successful action. In accordance with the importance of this kind of motivation, a series of instruments have been designed to measure the different components of achievement motivation

reliably, validly and practically. The existing instruments can be used in research and practical settings.

#### References

- Brandtstädter, J. & Renner, G. (1990). Tenacious goal pursuit and flexible goal adjustment: explications and age-related analysis of assimilative and accommodative strategies of coping. *Psychology and Aging, 5*, 58–67.
- Breaugh, J.A. & Colihan, J.P. (1994). Measuring facets of job ambiguity: construct validity evidence. *Journal of Applied Psychology, 79*, 191–202.
- Costa, P.T. & McCrae, R.R. (1989). *The NEO PI/FFI Manual Supplement*. Odessa, FL: Psychological Assessment resources.
- Fried, Y. & Ferris G. (1987). The validity of the job characteristics model: a review and meta analysis. *Personnel Psychology, 40*, 287–322.
- Graf, P. & Schacter, D.L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 501–518.
- Hackman, J.R. & Oldham, G.R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60*, 159–170.
- Heckhausen, H. (1989). *Motivation und Handeln*. Berlin: Springer.
- Hollenbeck, J.R., Klein, H.J., O'Leary, A.M. & Wright, P.M. (1989). Investigation of the construct validity of a self-report measure of goal commitment. *Journal of Applied Psychology, 74*, 951–956.
- Kleinbeck, U. (in press). Das Management von Arbeitsgruppen. In Schuler, H. (Ed.), *Lehrbuch der Personalpsychologie*.
- Kleinbeck, U. & Fuhrmann, H. (2000). Effects of a psychologically based management system on work motivation and productivity. *Applied Psychology: An International Review, 49*, 596–610.
- Kleinbeck, U. & Schmidt, K.-H. (in press). Gruppenleistung und Leistungsförderung. In Schuler, H. (Ed.), *Organisationspsychologie, Enzyklopädie der Psychologie*. Göttingen: Hogrefe.
- Kuhl, J. & Scheffer, D. (2000). Auswertungsmanual für den Operanten Multi-Motiv-Test (OMT). Osnabrück, unveröff. Manuskript.
- Locke, E.A. (1990). *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice Hall.
- Locke, E.A. & Latham, G.P. (1984). *Goal-Setting: a Motivational Technique that Works*. Englewood Cliffs, NJ: Prentice Hall.
- McClelland, D.C., Atkinson, J.W., Clark, R.A. & Lowell, E.L. (1953). *The Achievement Motive*. New York: Appleton-Century-Crofts.
- Mehrabian, A. (1968). Male and female scales of the tendency to achieve. *Educational and Psychological Measurement, 28*, 493–502.
- Murray, H.A. (1943). *Thematic Apperception Test Manual*. Cambridge: Harvard University Press.

- Rheinberg, F., Vollmeyer, R. & Burns, B.D. (in press). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen.
- Sawyer, J. E. (1992). Goal and process clarity: specification of multiple constructs of role ambiguity and a structural equation model of their antecedents and consequences. *Journal of Applied Psychology*, 77, 130–142.
- Schaarschmidt, U. & Fischer, A. (1996) AVEM – Arbeitsbezogenes Verhaltens- und Erlebensmuster (Manual). Frankfurt am Main: Swets Testservices.
- Schmalt, H.-D. (1999). Assessing the achievement motive using the grid technique. *Journal of Research in Personality*, 33, 109–130.
- Schmalt, H.-D. & Schneider, K. (2000). *Motivation*. Stuttgart: Kohlhammer.
- Schmalt, H.-D. & Sokolowski, K. (2000). Zum gegenwärtigen Stand der Motivdiagnostik. *Diagnostica*, 46, 115–123.
- Schmidt, K.-H. & Hollmann, S. (1998). Eine deutschsprachige Skala zur Messung verschiedener Ambiguitätsfacetten bei der Arbeit. *Diagnostica*, 44, 21–29.
- Schmidt, K.-H. & Kleinbeck, U. (1999). Job Diagnostic Survey (JDS – deutsche Fassung). In: Dunckel, H. (Ed.), *Handbuch psychologischer Arbeitsanalyseverfahren* (pp. 205–230). Zürich: vdf.
- Schmidt, K.-H. & Kleinbeck, U. (in press). Leistung und Leistungsförderung. In Schuler, H. (Ed.), *Organisationspsychologie, Enzyklopädie der Psychologie*. Göttingen: Hogrefe.
- Schuler, H. & Prochaska, M. (in press). Entwicklung und Konstruktvalidierung eines berufsbezogenen Leistungsmotivationstests.
- Spangler, W.D. (1992). Validity of questionnaire and TAT measures of need for achievement: two meta-analyses. *Psychological Bulletin*, 112, 140–154.
- Tubbs, M.E. (1993). Commitment as a moderator of the goal-performance relation: a case of clearer construct definition. *Journal of Applied Psychology*, 78, 86–97.
- Vroom, V.H. (1964). *Work and Motivation*. New York: Wiley.

Uwe Kleinbeck

### Related Entries

APPLIED FIELDS: ORGANIZATIONS, APPLIED FIELDS: WORK AND INDUSTRY, PERSONNEL SELECTION, LEADERSHIP IN ORGANIZATIONAL SETTINGS, LEADERSHIP PERSONALITY, MOTIVATION

# Achievement Testing

## INTRODUCTION

Achievement testing plays a central role in education, particularly given the current context of high-stakes educational reform seen in countries like the United States. This entry provides a brief overview of achievement testing beginning with a description of its role in education. Different types of achievement tests, commonly used derived scores, recent advances such as performance assessments, and future directions are described.

## ACHIEVEMENT TESTING AND ITS ROLE

Achievement tests are designed to measure the knowledge and skills that individuals learn in a relatively well-defined area through formal or informal educational experiences. Thus,

achievement tests include tests designed by teachers for use in the classroom and standardized tests developed by school districts, states, national and international organizations, and commercial test publishers.

Achievement tests have been used for: (a) summative purposes such as measuring student achievement, assigning grades, grade promotion and evaluation of competency, comparing student achievement across states and nations, and evaluating the effectiveness of teachers, programmes, districts, and states in accountability programmes; (b) formative purposes such as identifying student strengths and weaknesses, motivating students, teachers, and administrators to seek higher levels of performance, and informing educational policy; and (c) placement and diagnostic purposes such as selecting and placing students, and diagnosing learning disabilities, giftedness, and other special needs.

## 6 Achievement Testing

The most controversial uses of achievement testing have been in high-stakes accountability programmes and minimum competency testing (MCT). Accountability practices vary and may include financial rewards for improved performance to providing remediation for students who perform poorly to sanctions such as public hearings, staff dismissals, and dissolution of districts. Two negative consequences that have been associated with high-stakes accountability include a pattern of inflated achievement results as highlighted by Cannell's (1988) finding that all states were reporting that their students were scoring above the national norm (Lake Wobegon effect), and the narrowing of instruction or 'teaching to the test' so that student scores compare favourably to norms.

MCT programmes were implemented in response to concerns about high levels of illiteracy and innumeracy and subsequent poor 'work force readiness' among high school graduates. In addition to course completion requirements, such programmes require students to pass tests of minimal basic skills (usually in reading, writing, and arithmetic) to graduate from high school. Legal cases such *Debra P. vs. Turlington* raised questions about what constitutes minimum competency, whether the skills assessed are reflected in school curriculum, and whether students have been given adequate opportunity to learn the skills required.

### STANDARDIZED ACHIEVEMENT TESTS

Standardized tests may be classified using the overlapping categories of purpose, breadth, administration, item format, and interpretation.

#### Purpose

Screening tests tend to be relatively brief with only one subtest covering each subject area. These tests are useful in determining if more expensive comprehensive testing is warranted. Screening tests include the *Wechsler Individual Achievement Test (WIAT) – Screener*, *Wide Range Achievement Test – 3*, and *Basic Achievement Skills Individual Screener (BASIS)*. Comprehensive or diagnostic tests typically include more than one subtest

per subject area so each can be explored in depth. Examples of these tests include the *WIAT – Comprehensive Test*, *Woodcock–Johnson Complete Battery III*, *Gates–McKillop–Horowitz Reading Diagnostic Test*, *Comprehensive Tests of Basic Skills*, and *Terra Nova*.

#### Breadth

Single-subject tests include a number of subtests ranging from lower to higher skill levels to assess different aspects of a subject area. Single-subject tests include the *Woodcock Reading Mastery Tests – Revised* and *KeyMath – Revised*. Multiple-subject tests assess at least the three commonly taught subject areas of reading, mathematics, and written language. Such tests include the *Iowa Tests of Basic Skills*, *California Achievement Test*, *SRA Achievement Series*, *Stanford Achievement Test Series*, and *Tests of Achievement and Proficiency*.

#### Administration

Group administered achievement tests are usually multiple-subject tests that contain comparable subtests for students in different grades. These tests usually are administered within the classroom and are used throughout school districts or states. Examples include the *Iowa Tests of Basic Skills*, *Metropolitan Achievement Test 8*, *Iowa Tests of Educational Development*, *Gray Oral Reading Test – 3*, and *Sequential Tests of Educational Progress – III*. Individually administered achievement tests may include single- or multiple-subject tests and typically are administered in clinical and educational settings. Such tests include the *Kaufman Test of Educational Achievement*, *Wide Range Achievement Test – III*, *Gates–MacGinitie Reading Test*, and *Peabody Individual Achievement Test – Revised*.

#### Item Format

Fixed-response items include multiple-choice, true-false, matching, and stem completion items. A key advantage of fixed-response items is that considerable material can be covered in a relatively short period of time. Criticisms of these items are that they emphasize recall of facts over higher order thinking and problem-solving, they are susceptible to guessing and testwiseness, and they discourage creative thinking. They also tend to be difficult

items to prepare. Nonetheless, multiple-choice items are the most common item format used in standardized achievement tests.

Constructed items include short answer and essay responses. The advantages of constructed items are that they require students to construct a response rather than simply recognize the correct answer, they assess students' ability to organize, connect ideas, and problem-solve, they reduce the impact of guessing, and preparation of questions is relatively quick and easy. Disadvantages of constructed items are that relatively few questions can be asked and thus adequate coverage of the subject area may not occur, they are susceptible to bluffing, and scoring is time-consuming, requires considerable subjective judgement, and is less reliable than scoring of fixed-response items.

### Interpretation

When achievement test results are interpreted with reference to a normative group, the test is referred to as a norm-referenced test (NRT). Students' NRT scores usually are expressed in age- or grade-equivalent scores, standard scores, or percentiles. NRTs are designed to discriminate among students' performance; they do not provide information on the amount of information learned. Most of the tests discussed already are NRTs. When test results are interpreted in terms of whether each student has mastered specific knowledge and skills without reference to other students or a normative group, the test is said to be criterion-referenced (CRT). Students' CRT scores are usually expressed as percent correct or by descriptors such as mastery/non-mastery. Most CRTs are developed by schools or states. Examples are the *Basic Skills Assessment Program*, *Kentucky Instructional Results Information System*, and *Louisiana Educational Assessment Program* (LEAP 21). Some NRT tests also provide CRT interpretations such as *BASIS*.

### DERIVED SCORES ASSOCIATED WITH ACHIEVEMENT TESTS

Raw scores obtained on achievement tests typically are converted to derived scores, so we can make comparisons among test scores. Commonly found derived scores include age or grade equivalent scores, standard scores, and percentile scores. Age or grade equivalent scores

(‘developmental scores’) reflect average performance at different age and grade levels. These scores are often (a) misinterpreted when individual performance is compared to the wrong reference group, and (b) inappropriately used as standards of performance when teachers and parents expect all students in a particular age- or grade to achieve age- or grade-equivalent scores.

Standard scores provide an indication of a student's relative performance on a test in terms of how far his/her score is from the mean in standard deviation units. Common types of standard scores are z-scores, T-scores, deviation IQ scores, and stanines. Standard scores are the most highly recommended derived scores.

Percentiles (percentile ranks) indicate the point in a distribution at or below which the scores of a given percentage of students fall and should not be confused with percentages or percent correct. Percentiles are the most easily interpreted derived scores. However, percentiles do not represent equal intervals across the distribution, which means that they magnify small differences near the mean and minimize large differences in the upper and lower ends of the distribution.

## RECENT ADVANCES IN ACHIEVEMENT TESTING

### Computer Adaptive Testing

Computer adaptive testing (CAT) attempts to match the difficulty of test items to the knowledge and skill level of the student being assessed by tailoring the test so that a pre-selected sequence of items is administered based on whether or not the response to the previous item is correct. The advantages of CAT over traditional achievement tests include reduced testing time, the need for fewer items at a given level of measurement error, minimized frustration for students who perform poorly, and more precise estimates of achievement across the entire distribution.

### Large-Scale Assessments

Large-scale assessments are conducted by the district, state, or nation(s) to examine the educational achievement of groups. The best-known large-scale assessments today are the *National Assessment of Educational Progress*

(NAEP) in the United States and the surveys conducted by the International Association for the Evaluation of Educational Achievement (IEA). The purpose of NAEP, which was first introduced in 1969, was 'to improve the effectiveness of our Nation's schools by making objective information about student performance in selected learning areas available to policymakers at the national, State and local levels' (Public Law 100-297, Section 3401). The IEA has conducted numerous international achievement surveys since its first cross-national survey in 1959 and is best known for the longitudinal *Third International Mathematics and Science Study* (TIMSS) first conducted in 1995.

### Performance Assessment

Increasing attention has been paid to performance assessments (also known as authentic or alternative assessments), which consist of students' constructed responses to 'real world' (authentic) tasks and problems and the cognitive skills and processes involved in the construction of those responses. Examples of performance assessments include portfolios of students' work over time, poetry, science experiments, conversations in a foreign language, and open-ended mathematics problems. The students' work is judged using an agreed-upon set of criteria. The advantages of these assessments are that they are meant to measure processes involved in the acquisition of knowledge and skills in ways that make the link between learning and instruction clearer. Disadvantages are that fewer tasks can be included given time constraints, creating agreed-upon criteria for scoring is time-consuming, and judgement of students' work is highly subjective, all of which make performance assessment expensive and open to bias.

### Standards-Based Assessments

Standards-based reform describes efforts to improve education for all students through the setting of high standards. Its beginnings rest with the 1983 National Commission on Excellence in Education report, *A Nation at Risk: The Imperative for Educational Reform*, and it has culminated in the passing of 'Goals 2000: Educate America Act' by the U.S. Congress in 1994.

Standards-based approaches include content and performance standards, assessments that are

aligned with these standards, and accountability measures. Content standards define what a student should know and be able to do and thus drive curriculum. Performance standards define how much a student should know and be able to do, and thus set the benchmarks or expected levels of achievement to be used for accountability. Standards-based assessments (also known as standards-referenced testing) are based on content and performance standards, involve multiple measures of student performance, and apply to all students. A critical aspect of such assessments is to produce and use 'better tests' such as performance assessments. Accountability measures focus on strengthening standards-based reform initiatives by rewarding teachers and schools whose students meet performance standards and sanctioning those who do not.

### CONCLUSIONS AND FUTURE PERSPECTIVES

Current and future advances in achievement testing appear to be focussed on the development, improvement, and evaluation of standards-based and performance assessments. Five areas for future development include: (1) best practices for developing, and methods for evaluating, performance assessment scoring rubrics, (2) comparisons of the various types of data to be used in accountability models such as mean scores, value-added data, and residual scores adjusted for socio-economic status, (3) longitudinal research examining the impact of performance and standards-based assessments on student achievement, instructional practices, and student learning, (4) comparisons of traditional standardized testing (including multiple-choice formats) and performance assessments as measures of student achievement, and (5) exploration of computer-based, and notably Internet, delivery and scoring of performance assessments for large scale assessment.

### References

- Anastasi, A. & Urbina, S. (1997). *Psychological Testing*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Cannell, J.J. (1988). Nationally normed elementary achievement testing in America's public schools: how all 50 states are above the national average. *Educational Measurement: Issues and Practice*, 7, 5-9.



Hambleton, R.K. & Zaal, J.N. (1991). *Advances in Educational and Psychological Testing: Theory and Applications*. Boston, MA: Kluwer Academic Publishers.

National Commission on Excellence in Education (1983, April). *A Nation at Risk: The Imperative for Educational Reform* (Doc. 20402). Washington, DC: U.S. Government Printing Office.

Anita M. Hubley

## A ADAPTIVE AND TAILORED TESTING (INCLUDING IRT AND NON-IRT APPLICATION)

### INTRODUCTION

In computerized adaptive testing (CAT), a computer administers the items and gathers the examinee's responses, but its most distinctive feature is that the items finally administered depend on the examinee's ability. The test then adapts to the examinee's performance on the items. The idea of adaptive measurement can be traced back to Binet, but it never became a reality until the appearance of the item response theory (IRT) and the development of the computer. However, adaptive testing is also possible without IRT, as will be seen later. The first ideas on CAT appeared in the early 1970s (Lord, 1970). CAT has spent in the laboratory the greater part of the elapsed time since then, because the main concern of the researchers has been to obtain the most efficient, precise and possible strategies for item selection. They have become operational only in the last decade. Computerized adaptive tests were administered to more than a million people in 1999 (Wainer, 2000). Its main applications are to the areas of personnel selection, educational assessment, certification and licensure. Due to its practical applications, new concerns such as test security, profitability and social impact have arisen.

### BASIC PRINCIPLES

The basic principles of a CAT are well established. Its aim is to apply to each examinee only the items that best serve to assess his/her level of ability. Its main advantage is that more

### Related Entries

APPLIED FIELDS: EDUCATION, ACHIEVEMENT TESTING, CRITERION-REFERENCES TESTING: MODELS AND PROCEDURES, NON-REFERENCES TESTING: METHODS AND PROCEDURES, ITEM RESPONSE THEORY

efficient measurements are obtained. It needs fewer items (sometimes, less than half) than conventional tests to achieve the same level of precision as a full-length test. The elements that make up a CAT are: an item pool with known properties, a heuristic to choose the items, a method to evaluate ability and a criterion to end the application. Though they are all important, the efficiency of a CAT mostly depends on two closely related complementary processes: the statistical method of estimating ability and the criterion for item selection. This explains the great amount of procedures known and why they are two of the most studied aspects of CAT.

### ITEM BANK

A CAT chooses items from a database (item bank) containing the available items and various information about each item, such as its stem, correct and incorrect options, item parameter estimates under an appropriate IRT model, classical item difficulty and discrimination indices, information on the specific domain the item measures, the proportion of times the item has been administered, etc. The bank has to be calibrated and its unidimensionality and acceptable fitting to an IRT model should be checked and accepted. Item banking and IRT are specific entries in this encyclopedia, and further details can be found there.

A CAT does not need a specific item format. A CAT may be developed both for dichotomous and polytomous items, and for multiple-choice or

open-ended items. Items may be visual, auditory and also multimedia items. It is also possible to consider a testlet (cluster of related items) instead of single items as the analysis unit.

An important question to pay attention to is bank size. Well-known high-stake CATs, such as CAT-ASVAB (Sands et al., 1997), have more than one thousand items, but CATs for other uses ordinarily have smaller banks (even below 150 or 200 items). The number of items also depends on the restrictions the item-selection algorithm has implemented and the IRT model in use.

An item bank should also consider the ability prospective examinees have and the intended test use. It should contain discriminative items for the entire range of ability. The information function of the item bank should match these requirements.

Banks should be updated, both in the information on each item, as this information changes after each item administration, and also in the items themselves, because as the CAT is increasingly administered, new items should be added and old ones removed. Online calibration deals with effective procedures to carry-out bank updating.

## ADAPTIVE HEURISTICS

A CAT needs four components in order to measure an examinee: (a) a procedure to select the first item to administer; (b) a method to estimate the examinee's ability and precision after each administered item; (c) an algorithm for selecting the remaining items; and (d) a criterion to end the test administration.

There are some alternatives available for selecting the first item. If we know the examinee's grade on other variables, and his/her course, this information may be used to predict the examinee's ability by linear regression. The first item is then selected to match the predicted ability. If no information on the examinee is available, the first item will then match a random ability selected from the central values of the ability distribution.

After each item is administered the examinee gives his/her response. The CAT needs to obtain the ability estimation from the observed responses to the set of administered items. The examinee's

test score will be his/her ability estimate when the test is over. The most widely used methods of estimation are based on the principle of maximum likelihood or Bayesian procedures. These methods have good properties when the number of items is high. Nevertheless, CATs are far from this ideal situation because they use very few items. This circumstance gives place to biased estimations, especially in the early stages of the test. So, a problem with CAT is to find a method that provides accurate estimations, which are unbiased and computationally efficient. Wang and Vispoel (1998) and Cheng and Liou (2000) have compared the characteristics (standard error, bias, efficiency, etc.) of the IRT estimation methods to determine when they are advisable in CAT. Since one of the problems that has been paid the most attention is the bias of the estimations, several strategies have been proposed for its control. Other questions, like the initial estimation and the effects of non-model fitting responses, have generated interest from researchers.

Once some items have been administered and an ability estimate for the examinee obtained, a new item has to be selected from the unused items remaining in the item pool. Two common principles are used to guide item selection. Under the *maximum information* principle, the information provided for each unused item for the last ability estimate is computed. The item with the highest information value is selected and administered. In other words, the more helpful item in order to increase precision is selected. The maximum information principle faces some difficulties when the ability estimate is biased or inaccurate, which is often the case when the test is short. If the estimate separates appreciably from the final estimation, the more informative items for these provisional estimations will be less informative for the final estimation. As a result, some items will have been of little use in the test. Cheng and Liou (2000) have proposed the use of alternative information measures in order to circumvent this difficulty. Under the *maximum expected precision* criterion, the item selected will minimize the expected value of the variance of the Bayesian posterior distribution of ability. Several item-selection criteria based on this principle have been proposed (van der Linden, 1998).

Both the procedures share a problem derived from choosing in each test the best items in

the pool: some items are administered in most tests (even in more than 50% of the tests), risking test security and validity, whereas others are never shown (in one particular CAT, 80% of the items were never selected). To sort-out this difficulty, control exposure methods have been implemented. These methods trade-off precision with security. When a CAT has in use an exposure control method, such as the Sympson–Hetter method (Sympson & Hetter, 1985), precision of measurement is not as high, but the exposure rate of the most useful items is held under control, and the smaller exposure rates are obtained. Experimental CATs may have implemented one of the two pure item-selection procedures indicated above, but if a CAT has to give valid measures it needs to attend to other considerations in order to select items, such as the appropriate representation of the content or subject areas, the guarantee that the composition of the test is similar for all examinees, the control of the presence of items that should not appear together in the same test, etc. Item-selection rules should then consider not only the basic principles indicated above, but also item-control exposure and other restrictions. Linear programming techniques are often used to make item selection feasible when different restrictions have to be simultaneously considered.

The administration of items ends when either the test length or ability precision reach their preset values. In the second case, all the examinees will be measured with the same precision, but the number of items administered and testing time will differ. Sometimes the stopping criterion is mixed: the test stops after presenting a preset number of items if it does not reach the targeted precision.

## PSYCHOMETRIC PROPERTIES

As in a conventional test, reliability and validity studies have to be carried out in a CAT. Besides traditional reliability methods, such as test–retest, simulation may be used to obtain information on test functioning. Indices such as RSME, bias and efficiency can be easily computed. Concerning validity, the procedures in use for conventional tests are applicable to CATs. For further details on this, see Chapters 7 and 8 of Wainer et al. (2000).

## OTHER RESEARCH TOPICS

### New Types of CATs

Most of the CATs have been elaborated to measure intelligence, aptitudes or achievement, and they are based on IRT models for unidimensional dichotomous items. However, other alternatives have been considered in the past few years. The need to measure other constructs such as attitudes, whose items have the format of ordered categories, and the possibility of using the incorrect options of the multiple-choice items to improve the estimation of ability, started a new line of work interested in elaborating CATs based on diverse types of polytomous models. Also, the acknowledgement that more than one trait intervenes in almost all the tasks has led to the use of multidimensional CATs (Segall, 2000).

### CATs in Intelligent Tutoring Systems

The use of IRT in CATs imposes a few important constraints (Almond & Mislevy, 1999): (a) IRT has a simple way of representing knowledge and skills that intervene in complex tasks (unidimensionality); (b) it establishes strong assumptions that can be violated on some occasions (conditional independence); (c) it requires large samples to estimate the item parameters; and (d) it offers a score to express the level of ability, which does not exactly indicate what the subjects know or can do (diagnosis). All these aspects reduce their use in measuring domains that require multiple knowledge, skills and abilities, as in educational and job performance assessment. There is a tendency in education to integrate measurement, assessment, diagnosis, teaching and learning. This means that it is necessary to know in detail the knowledge and skills dominated by the students, the kinds of mistakes they make, the strategies they use, etc. to be able to adapt the contents and pedagogic strategies to them. To what extent can this be achieved by available CATs? Hardly at all, unfortunately.

This orientation in performance assessment is creating the need to introduce important changes in CATs, most of them coming from the literature on intelligent tutoring systems (ITS). In computerized teaching, since the ITS appeared,

there has been a growing interest in giving it more capacity of accommodation for the candidates, including a CAT in its module of assessment. Most of these systems do not use IRT, they are based instead on other methodologies, such as the rule-space methodology (Tatsuoka & Tatsuoka, 1997), the knowledge spaces (Hockemeyer & Albert, 1999), Bayesian networks, or graphical models (Almond & Mislevy, 1999), etc.

### Conditions of Application

Lastly, many practical problems emerge when CATs become operative instruments used in real life. One main concern is how to guarantee the security of the item pool against attempts at illegitimate appropriation of its contents as well as the complexity and high costs of the elaboration process, maintenance and renewal. A second topic of interest tries to make the conditions of test administration better psychologically for the candidates, such as obtaining optimum adjustment in the difficulty of the test, allowing review of the answers, and controlling the difficulty of the items to reduce anxiety.

### PROSPECTS AND CONCLUSIONS

From a technical perspective, there have been significant steps made in ability estimation methods. Likewise, the item-selection heuristics have reached a level of sophistication that makes it capable of guaranteeing the elaboration of tests that meet multiple requirements. In the next generation of CATs, new models will be used. Very soon we will be able to see comparative studies that analyse these new models that are multidimensional and can handle polytomous response data. However, the CATs elaborated from these models have yet to prove that their advantages are worth the additional effort their elaboration requires. This is especially true for multidimensional models.

Many practical problems have emerged with CAT going operational (Wise & Kingsbury, 2000), especially those related to test security and costs. Wainer (2000) provided a critical discussion of the supposed advantages attributed to CATs in the 1980s, from the experience accumulated on their massive use in the 1990s. His conclusion, though

not very favourable, is not discouraging. Wainer argues for more focus on areas where CATs will be useful: (a) when the construct cannot be measured easily without a computer, (b) when the test has to be continuously administered, and (c) when it is important for everyone involved to get the right measurement.

In the past few years, a growing tendency to extend the use of CATs to the Internet, using its Web service (CAT-Web) has been appreciated. This tendency basically responds to the interest of having the distance learning system also offering an individualized assessment. In this way, more ITS destined to the Internet are continuously being released, and some of them already include a CAT-Web.

Finally, two challenges that CATs will face in the future will be to offer diagnostic information of quality on multiple abilities and to substantially reduce the costs associated with the elaboration of the item pool. In the first place, CATs have to go further than the unidimensional dichotomous IRT models, and especially to solve in an efficient way the problem of multidimensionality. Moreover, according to the objectives of the test, offering quantitative scoring may not be enough. The solution could be far from the IRT. The possibilities offered by the models of measurement based on knowledge, like the Bayesian inference network or the knowledge space theory, would have to be seriously considered. In the second place, an alternative to *online* calibration and the automatic generation of items that could serve to reduce costs is to elaborate *instruments of measurement that learn to measure*. The necessary elements would be a theoretical model of the construct that is well supported, a psychometric model, a group of experts on the subject to obtain the initial parameters and an algorithm of learning. The test will modify the initial estimates of the experts from the empirical information collected, and from its execution in activities in which it could be trained through simulation. The algorithm of learning would bring the values of the parameters up to date so they would adapt to the predictions of the theoretical model and the available empirical evidence. The uses of the scoring would be conditioned to the degree of competence achieved by the test. Although it may seem far-fetched, some attempts

are being made in this direction in CATs of some ITS.

## References

- Almond, R.G. & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 223–237.
- Cheng, P.E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, 24(3), 257–265.
- Hockemeyer, C. & Albert, D. (1999). The adaptive tutoring system RATH: a prototype. In Auer, M.E. & Ressler, U. (Eds.), *ICL99 Workshop Interactive Computer Aided Learning: Tools and Applications*. Austria: Willach
- Lord, F.M. (1970). Some test theory for tailored testing. In Holtzman, W.H. (Ed.), *Computer Assisted Instruction, Testing and Guidance* (pp. 139–183). New York: Harper and Row.
- Sands, W.A., Waters, B. K. & McBride, J. R. (1997). *Computerized Adaptive Testing: From Inquiry to Operation*. Washington: American Psychological Association.
- Segall, D.O. (2000). Principles of multidimensional adaptive testing. In van der Linden, W.C. & Glas, C.A.W. (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 53–73). Boston, MA: Kluwer-Nijhoff.
- Sympson, J.B. & Hetter, R.D. (1985). *Controlling Item Exposure Rates in Computerized Adaptive Testing*. Paper presented at the meeting of the Military Testing Association, San Diego, CA.
- Tatsuoka, K.K. & Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: effect on remedial instruction as empirical validation. *Journal of Educational Measurement*, 34(1), 3–20.
- van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2), 201–216.
- Wainer, H. (2000). CATS: whither and whence. *Psicológica*, 21, 121–133.
- Wainer, H., Dorans, N.J., Flaugher, R. Green, B.F., Mislevy, R.J., Steinberg, L. & Thissen, D. (2000). *Computerized Adaptive Testing: A Primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wang, T. & Vispoel, W.P. (1998) Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 109–135.
- Wise, S.L. & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, 21, 135–155.

Vicente Ponsoda and Julio Olea

## Related Entries

COMPUTER-BASED TESTING, ITEM BANKING, ITEM RESPONSE THEORY: MODELS AND FEATURES

# A

## AMBULATORY ASSESSMENT

## INTRODUCTION

*Ambulatory assessment* designates a new orientation in behavioural and psychophysiological assessment. This approach relates to everyday life ('naturalistic' observation) and claims the ecological validity of research findings. Methods of recording psychological data in everyday life have a long history in differential psychology and clinical psychology. Event recorders for the timed registration of stimuli and responses, 'beeper' studies in which a programmable wristwatch prompts the subject to respond to a questionnaire, self-ratings on diary cards, and electronic data loggers have all been used for this purpose. The arrival of pocket-sized (hand-held, palm-top)

computers has eased the acquisition of data considerably. Computer-assisted methodologies facilitate investigations in real-life situations where relevant behaviour can be much more effectively studied than in the artificial environment of laboratory research. Such field studies are essential, for example, in research on stress-strain or in research on mechanisms that trigger off psychological and psychophysiological symptoms.

Ambulatory assessment originated from a number of previously rather independent research orientations with specific objectives. *Clinical (bedside) monitoring* was introduced as a means of continuously observing a patient's vital functions, e.g. respiratory and cardiovascular

parameters under anaesthesia, during intensive care and in perinatal condition. If relevant changes occur, i.e. if certain critical values are exceeded, an alarm is set off. *Biotelemetry* employs transmitter–receiver systems (radio-telemetry) in order to measure physical functions in real life, e.g. cardiovascular changes during intense strain at the workplace or during athletic performance. Radio equipment basically makes two-sided communication possible, i.e. feedback, telestimulation and telecommand, in addition to telemetry. *Ambulatory monitoring* means continuous observation of free-moving subjects (patients) in everyday life as compared to stationary, bedside (‘wired’) monitoring. Ambulatory monitoring can be conducted either by biotelemetry or by a portable recording system. This methodology is appropriate for patients who exhibit significant pathological symptoms which, for a number of reasons, cannot be reliably detected in the physician’s office or hospital as compared to a prolonged observation in everyday life. Such cases include ventricular arrhythmia, ischaemic episodes, sleep apnea, and epileptic seizures. Here, ambulatory monitoring furthers valid diagnoses as well as the stabilization of medication. *Field research* comprises observation in natural settings in contrast to the laboratory. Field research is an essential methodology in cultural anthropology, social research, and ethology. Likewise, in psychology and psychophysiology some research issues require field studies to obtain valid data (see Kerlinger & Lee, 2000; Patry, 1982). *Behavioural assessment* methods include, besides laboratory observation, a variety of in-vivo (in-situ) tests, simulated and quasi-naturalistic settings, such as behavioural approach/avoidance tests (BATs) which were designed to assess behaviour disorders and clinical symptoms.

*Ambulatory assessment* brings together those research orientations that correspond to each other in their basic ecological perspective. Ambulatory assessment involves the acquisition of psychological data and/or physiological measures in everyday life according to an explicit assessment strategy which relates data, theoretical constructs, and empirical criteria specific to the given research issue. Such field studies are not solely concerned with the ambulatory monitoring of patients, but rather include a wide spectrum of objectives and applications. Common features

are: recordings in everyday life, computer-assisted methodology, attempts to minimize method-dependent reactivity, maintaining ecological validity and, therefore, outstanding practical utility for various objectives – such as monitoring and self-monitoring, screening, classification and selection, clinical diagnosis, and evaluation – in many areas of psychology and psychophysiology (de Vries, 1992; Fahrenberg & Myrtek, 1996, 2001a,b; Littler, 1980; Miles & Broughton, 1990; Pawlik & Buse, 1982, 1996; Pickering, 1991; Suls & Martin, 1993).

## ACQUISITION OF PSYCHOLOGICAL DATA BY HAND-HELD PC

In psychology the hand-held PC so far has been predominantly used for recording self-reports on mood and other aspects of subjective state, including physical complaints and symptoms, that is, as an ‘electronic diary’ (e.g. job stress diary, pain diary). There are other kinds of data, which can be obtained in field studies: objective features of a behaviour setting, behaviour observations, behaviour and performance measures (psychometric testing), self-measurements of various kinds, for example, blood pressure data, and, possibly, ratings of environmental aspects. Potential contents of a computer-assisted protocol may further include, for example, individual comments or self-evaluation in connection with events.

## Advantages and Limitations

The application of a programmable pocket PC in ambulatory assessment has many advantages:

- alarm functions for prompting the subject at predefined intervals and a built-in reminder signal;
- reliable timing of input, delay of input, and duration of input;
- flexible layout of questions and response categories;
- branching of questions and tailor-made sequential or hierarchical strategies;
- concealment of previously recorded responses from the subjects;
- convenience and ease of transfer of data to a stationary PC for statistical analysis.

A higher technical reliability and ecological validity of computer-assisted recordings can be generally assumed compared with paper-and-pencil questionnaires and diaries that lack flexibility in data acquisition and exactness when timing responses. The versatility and wide acceptance of computer-assisted data acquisition is evident, although there are limitations and obvious restrictions. All participants of such studies will need sufficient practical training in the basic features of the PC and the program, at least, to avoid malfunctions and missing data. In spite of the obvious increase in computer literacy within the general population, there are sub-populations which are less familiar with such devices or may experience problems.

Following the progress made in pocket-sized computers, software to facilitate the use of hand-held PCs in field studies has been developed in many institutions, more or less geared to the needs of certain studies. More flexible software systems suited to the requirements of a variety of applications are still an exception (AMBU for in-field performance testing, cf. Buse and Pawlik, Hogrefe Verlag, Göttingen, Germany; MONITOR, a flexible software system for ambulatory assessment, Psychology Department, University of Freiburg, Germany). The OBSERVER (Behaviour Observation System, Noldus Information Technology, AG Wageningen, NL, Noldus, 1991) was introduced to ease the recording of behaviour observations in field studies in animal and human ethology.

Computer-assisted self-reports require a hand-held PC with certain features: a large display, easy handling of basic controls, clock, beeper with volume control, sufficient capacity of storage, low power consumption, and a low weight. For many applications, a comparatively large alphanumeric keyboard (complete QWERTY) is also preferable in order to ease recording and, especially, to record verbal responses. The latter may involve, for example, recording reports and comments about specific events, or reporting more precisely the occurrence of physical and psychological symptoms, which in either case hardly fit pre-defined categories. For some applications it may suffice to record only 'yes' or 'no' responses or numbers. In this case, a smaller hand-held PC, e.g. the Palm<sup>™</sup> series, may be preferable, although small keys or

a stylus may present a problem for some subjects or patients.

## PHYSIOLOGICAL MEASUREMENT AND MONITORING

A wide selection of physiological variables have been measured in daily life, using mostly non-invasive methods. The ECG and blood pressure enjoy by far the largest number of references in the literature to ambulatory monitoring. The application is predominantly in the medical field and only to a much smaller extent in the behavioural sciences, for example, in psychophysiology or behavioural medicine. The advances in micro-processor technology and storage capacity paved the way for multi-channel recordings and – another innovative step – led to the on-line analysis of medically important changes, for example, the immediate detection of ST-depression in the ECG.

The recording of posture and motion is another basic issue in the methodology of behaviour observation and performance measurement. Piezoresistive sensors (multi-site calibrated accelerometry) allow for:

- continuous recording and automatic detection of changes in posture and movement;
- assessment of movement disorders, such as hand tremor, restless legs syndrome;
- detection of head movement, e.g. nodding during a conversation, measured by a small accelerometer placed beneath the chin;
- estimation of gross physical activity and energy expenditure.

To assist in objective behaviour analysis, a range of interesting variables could be measured continuously:

- voice signal recorded via a throat sensor (micro);
- the temporal pattern of speech;
- ambient conditions recorded via suitable sensors for light, noise, and temperature.

Some hand-held PCs allow for audio recordings up to a number of minutes, depending on storage capacity. Digital dictating systems have a capacity up to 240 minutes in long play mode. In psychological and psychophysiological research, so far, little use has been made of digital

mini-cams or web-cams for recordings of the videostream of behaviour.

### Recorder–Analyser

Today, more than a dozen recorder–analyser systems are available from international manufacturers – not to mention the even greater number of long-term ECG recorders–analysers and the long-term-BP recorders. Only a few systems have a multi-channel design, the advantage of which is that they can be applied to a variety of research questions that require different recording channels (for an overview of selected multi-purpose recorder systems, see Fahrenberg, 2001). Besides the devices suitable for ambulatory recordings and their use in 24-h or long-term monitoring, a wide range of portable (mobile) equipment designed for in-field measurement does exist.

## ISSUES IN AMBULATORY ASSESSMENT

### Assessment strategies, designs, and data analysis

In psychological research various designs for computer-assisted ambulatory assessment have already been employed, whereby some of these assessments lasted for many days or weeks. In psychophysiology and in medicine, the restriction to a single 24-h recording appears to be the preferred format due to the costly equipment. Ambulatory assessment requires the elaboration of specific designs and strategies, for example, the strategic use and integration of time and event sampling, and the development of appropriate statistical models for multi-level analyses and for rather short time series (for a discussion, see Fahrenberg & Myrtek, 2001a; Schwarz & Stone, 1998; Stemmler, 1996). It would be oversimplified to state methodological advantages of the laboratory experiment as obstacles in field studies and vice versa, i.e. to retain the notion of basically different research strategies instead of a wider perspective that includes laboratory and field as complementary approaches.

*Laboratory–field comparisons* are designed to examine the validity of findings obtained in the laboratory to predict performance in real life. In the development of psychological tests such empirical

validation studies play an important role. More recently, it has been questioned whether certain diagnostic techniques and measurements in the physician’s office or in the psychophysiological laboratory, e.g. blood pressure measurement, reliably predict individual differences in real life. Laboratory–field comparisons revealed significant discrepancies. Office hypertension is a good example of how certain features of the setting and their meaning to the subject may play an important role in assessing individual differences: blood pressure readings are elevated if the measurement is made by the physician, but normal readings are obtained in everyday life.

*Laboratory–field comparisons* were valuable in the evaluation of methodological issues as well as practical aspects. Field studies, apparently, are more suited for prolonged observation that may extend over days and weeks. Accordingly, there is more chance for the detection of rare events and symptoms that occur at low frequencies or only in certain settings. Generally, larger response magnitudes and more realistic effect sizes may be expected in natural settings. Prolonged observation periods make the averaging/aggregation of measurements possible so that reliability and stability of measures may increase substantially. But field studies can be seriously threatened by the confounding of multiple effects which tend to produce ‘noise’ and, eventually, require relatively large subject samples in order to obtain valid estimates for main effects.

*Psychophysiological monitoring.* Multi-modal psychophysiological 24-h monitoring methodologies were developed in many fields, especially in research on blood pressure reactivity. This method consists of multi-channel recordings of blood pressure, heart rate, physical activity, and – concurrent to each blood pressure measurement – obtained a computer-assisted self-report on setting, behaviour, emotional state, and experience.

*Controlled monitoring.* Recordings obtained in everyday life will often include multiple effects. Therefore, investigators may wish to control for unwanted variance, such as blood pressure changes caused by physical activity. Concurrent recordings of physical activity provided means for a segmentation of recordings according to high or low activity. Furthermore, standardized or semi-standardized measurement periods were included which served as a reference for inter- and intra-individual



comparison. As part of the standard protocol in 24-h monitoring, the subjects performed specific tasks: climbing a staircase, performing a mental test, and participating in a short interview.

*Interactive monitoring.* The development of recorder equipment suitable for physiological and psychophysiological recordings and on-line (real-time) analysis led to innovative research strategies. Contingent to changes of certain physiological parameters, a patient can be prompted by a beeper signal to record specific events, activities, or symptoms. Myrtek et al. (1988) developed a new methodology for interactive monitoring of 'additional heart rate' indicative for emotional states.

### Acceptance, Compliance, and Reactivity

From the beginning, there have been concerns raised about the acceptance of hand-held PCs, and the validity of monitoring in daily life has been questioned. Ambulatory assessment with a pocket PC or recorder depends on the favourable attitude of the participating subjects. It is essential that the equipment is readily accepted and that good compliance to instructions is established and sustained. If the ambulatory monitoring is part of a diagnostic process or a treatment programme, the patient's compliance may be higher than in research projects. The ambulatory assessment should, of course, not cause major problems with the social environment.

The method of observation and measurement itself may cause unwanted variance because of specific interactions such as awareness, adaptation, sensitization, and coping tendencies. Three aspects of reactivity appear to be specific to ambulatory assessment. Subjects may: (1) tend to steer clear of certain settings during the recording in order to avoid being monitored there; (2) tend to unintentionally or deliberately manipulate the recording systems, shift settings of the PC and may even try to get access to the program; and (3) try to test their capacities or the equipment by unusual patterns of behaviour, exercise or vigorous movements. A comprehensive post-monitoring interview is recommended in order to obtain information on these essential aspects.

*Ethical issues* that are specific to ambulatory monitoring studies have hardly been discussed yet. Appropriate data protection is but one aspect, as

ambulatory assessment may violate privacy more easily than other methods. Furthermore, significant others and bystanders may become involved when the observation and the evaluation of settings are demanded. Obtaining the subject's informed consent before the recording starts is essential, but may be problematic since the exact course of daily activities and events cannot be anticipated.

*Acceptance and impact of computer-assisted monitoring methodology in psychophysiology and psychology.* The ambulatory monitoring of BP and ECG are now indispensable routine methods in medicine. The ever more widespread application of the new methodology can be attributed to its practical usefulness which was evident in the increased validity of diagnosis and in the external validity of therapy outcome evaluation. In contrast, computer-assisted monitoring and assessment still appear to have had little impact in psychophysiology and psychology. Standard textbooks on behavioural research methods and assessment in clinical psychology hardly refer to the new methodologies based on computer-assisted data acquisition and monitoring in the natural environment.

### PERSPECTIVES

Computer-assisted ambulatory assessment is an emerging new methodology. Progress is obvious not only in instrumentation, but in assessment strategies as well. Ambulatory assessment, like any other method, has problematic aspects, in particular how to account for multiple effects in the recordings, but the benefits are evident:

- *recording* of relevant data in natural settings;
- *real-time measurement* of behavioural and physiological changes;
- *real-time assessment and feedback* by reporting physiological changes to the subject;
- *concurrent assessment* of psychological and physiological changes (detection of events, episodes);
- *correlation and contingency (symptom-context) analysis* across systemic levels as suggested in triple-response models (multimodal assessment);
- *ecological validity* of findings and suitability for direct application.

Genuine research findings in relevant fields suggest further development and application of ambulatory assessment methodology. The expectation is that the hand-held PCs and the recorder-analyser for physiological measures will in future become smaller, cheaper and more refined. Such developments may include new strategies in controlled or interactive monitoring and on-line feedback, monitoring and concurrent recording of audio and video signals (intelligently pre-processed before stored), setting-dependent sampling, new strategies in self-monitoring and self-management in chronic illness.

A hand-held PC may be useful in the diagnostic assessment of a variety of behaviour disorders, for example, the assessment and self-management of drinking, smoking, and of eating disorders, and in facilitating self-management in chronic illness. Computer programs that are based on a hand-held PC can be used as a component of behavioural therapy (cf. a pilot study by Burnett et al., 1985).

There are noticeable developments which probably exert an essential influence on the computer-assisted methods in medicine and the behavioural sciences: the arrival of the wireless application protocol WAP, mobile phone short message systems SMS, the web-based mobile telecommunication (IMT-2000 and UMTS) and the new patient monitoring equipment, which appears to revolutionize the way in which patient information is transmitted and used in the healthcare system. At present, we may only speculate about the consequences of such developing information technologies for the healthcare system and, to some extent, on subsequent developments in applied fields of psychology.

## CONCLUSIONS

During the last two decades, a fast development in microprocessor technology has enabled the design of new instrumentation and, accordingly, new methodologies in medicine and the behavioural sciences. Multi-channel recorders-analysers and special purpose devices for physiological measures and convenient hand-held PCs for acquisition of psychological data are available. Such systems allow innovative research and

practical application in many fields and essential findings have been obtained.

## References

- Burnett, K.F., Taylor, C.B. & Agras, W.S. (1985). Ambulatory computer assisted therapy for obesity: a new frontier for behavior therapy. *Journal of Consulting and Clinical Psychology*, 53, 698–703.
- de Vries, W. (Ed.). (1992). *The Experience of Psychopathology. Investigating Mental Disorders in Their Natural Settings*. Cambridge: Cambridge University Press.
- Fahrenberg, J. (2001). Origins and developments of ambulatory monitoring and assessment In Fahrenberg, J. & Myrtek, M. (Eds.). *Progress in Ambulatory Assessment: Computer-Assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies* (pp. 587–614). Seattle, WA: Hogrefe and Huber.
- Fahrenberg, J. & Myrtek, M. (Eds.) (1996). *Ambulatory Assessment: Computer-Assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies*. Seattle, WA: Hogrefe and Huber.
- Fahrenberg, J. & Myrtek, M. (Eds.) (2001a). *Progress in Ambulatory Assessment: Computer-Assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies*. Seattle, WA: Hogrefe and Huber.
- Fahrenberg, J. & Myrtek, M. (2001b). Ambulantes Monitoring und Assessment. In Rösler, F. (Ed.), *Enzyklopädie der Psychologie. Serie Biologische Psychologie. Band 4: Grundlagen und Methoden der Psychophysiologie* (pp. 657–798). Göttingen: Hogrefe.
- Kerlinger, F.N. & Lee, H. B. (2000). *Foundations of Behavioral Research*. Fort Worth, TX: Harcourt.
- Littler, W.A. (Ed.). (1980). *Clinical and Ambulatory Monitoring*. London: Chapman and Hall.
- Miles, L.E. and Broughton, R.J. (Eds.) (1990). *Medical Monitoring in the Home and Work Environment*. New York: Raven Press.
- Myrtek, M., Brügger, G., Fichtler, A., König, K., Müller, W., Foerster, F. & Höppner, V. (1988). Detection of emotionally induced ECG changes and their behavioral correlates: a new method for ambulatory monitoring. *European Heart Journal*, 9 (Supplement N), 55–60.
- Noldus, L.P.J.J. (1991). The Observer: a software system for collection and analysis of observational data. *Behavior, Research Methods, Instruments & Computers*, 23, 415–429.
- Patry, J.L. (Ed.) (1982). *Feldforschung. Methoden und Probleme sozialwissenschaftlicher Forschung unter natürlichen Bedingungen*. Bern: Huber.
- Pawlik, K. & Buse, L. (1982). Rechnergestützte Verhaltensregistrierung im Feld: Beschreibung und erste psychometrische Überprüfung einer neuen Erhebungsmethode. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 3, 101–118.

- Pawlik, K. & Buse, L. (1996). Verhaltensbeobachtung in Labor und Feld. In Pawlik, K. (Ed.), *Enzyklopädie der Psychologie. Differentielle Psychologie und Persönlichkeitsforschung. Band 1. Grundlagen und Methoden der Differentiellen Psychologie* (pp. 359–394). Göttingen: Hogrefe.
- Pickering, T. G. (1991). *Ambulatory Monitoring and Blood Pressure Variability*. London: Science Press.
- Schwarz, J.E. & Stone, A.A. (1998). Strategies for analysing ecological momentary assessment data. *Health Psychology, 17*, 6–16.
- Stemmler, G. (1996). Strategies and designs in ambulatory assessment. In Fahrenberg, J. and Myrtek, M. (Eds.), *Ambulatory Assessment: Computer-Assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies*. (pp. 257–268). Seattle, WA: Hogrefe and Huber.
- Suls, J. & Martin, R.E. (1993). Daily recording and ambulatory monitoring methodologies in behavioral medicine. *Annals of Behavioral Medicine, 15*, 3–7.

Jochen Fahrenberg

### Related Entries

PSYCHOPHYSIOLOGICAL EQUIPMENT AND MEASUREMENT, APPLIED FIELDS: PSYCHOPHYSIOLOGICAL EQUIPMENT FOR ASSESSING BASIC PROCESS, FIELD SURVEY: PROTOCOLS DEVELOPMENT, OBSERVATIONAL METHODS (GENERAL)

## A ANALOGUE METHODS

### INTRODUCTION

Analogue behavioural observation (ABO) involves a situation designed by, manipulated by, or constrained by an assessor that elicits a measured behaviour of interest. Observed behaviours comprise both verbal and non-verbal emissions (e.g. motor actions, verbalized attributions, observable facial reactions).

ABO exists on a continuum of naturalism, ranging from highly contrived situations (e.g. How quickly do people walk down the hallway after being exposed to subconsciously presented words about ageing? Bargh et al., 1996) to naturalistic situations arranged in unnatural ways or settings (e.g. How do couples talk with one another when asked to discuss their top problem topic? Heyman, 2001) to naturalistic situations with some (but minimal) experimenter-dictated restrictions (e.g. family observations in the home; Reid, 1978).

### WHY USE ABO?

ABO is used as a hypothesis-testing tool for three purposes: (a) to observe otherwise unobservable behaviours, (b) to isolate the determinants of behaviour, and (c) to observe dynamic qualities of social interaction. Although naturalistic

observation might be preferable (i.e. generalizability inferences are minimized), the first two purposes require controlled experimentation, necessitating ABO; for the third purpose, ABO is often preferable because it allows the observer to “stack the deck” to make it more likely that the behaviours (and/or functional relations) of interest will occur when the assessor can see them.

### DOMAINS

ABO comprises two main assessment domains: individual/situation interactions and social situations. The goals of individual/situation interaction experiments are to manipulate the setting and test individual differences in response. This domain comprises a wide variety of tasks in developmental psychology (e.g. strange situation experiments; Ainsworth et al., 1978), social psychology (e.g. emotion regulation experiments; Tice et al., 2001) and clinical psychology (e.g. functional analysis of self-injurious behaviour; Iwata et al., 1994; social anxiety assessment; Norton & Hope, 2001).

The social situation domain employs ABO mostly as a convenience in assessing quasi-naturalistic interaction. The goal of such assessment is typically to understand behaviour and its determinants in dynamic, reciprocally influenced systems

(e.g. groups, families, couples). Understanding generalizable factors that promote or maintain problem behaviours in such systems typically requires more naturalistic approaches than those used in the other domain. Thus, although experimentation is often extremely useful in understanding causal relations in social situations (e.g. whether maternal attributions affect mother-child interactions; Slep & O'Leary, 1998), most such ABO investigations aim for quasi-naturalism.

## CLINICAL ASSESSMENT

ABO is a useful tool in clinical assessment, although relatively few ABO paradigms have been developed specifically with this application in mind. To be clinically useful, ABO must efficiently provide reliable, valid, and non-redundant (but cost-effective) information.

An apt analogy for research-protocol based assessment vs. field-realistic assessment might be found in the treatment literature. In recent years, a distinction has evolved between efficacy studies (i.e. those studying interventions under tightly controlled, idealized circumstances, such as a trial of treatment for major depressive disorder that eliminates all potential participants with comorbid disorders) and effectiveness studies (i.e. those studying interventions under real-world conditions). Because we do not have an adequate research body of effectiveness studies, clinicians in the field, urged to use empirically validated treatments, are expected to adapt such protocols to meet real-world demands. Similarly, clinicians should be urged to use empirically validated ABO when it would be appropriate, but should be expected to adapt ABO protocols in a cost-effective but still clinically-informative manner.

### ABO Protocols

Space limitations preclude a summary of the wide variety of ABO protocols. We note, however, literatures on parent-child interaction (e.g. Roberts, 2001), couple interaction (e.g. Heyman, 2001), social anxiety and social interaction skills (Norton & Hope, 2001), fear (e.g. McGlynn & Rose, 1998), self-injurious behaviour in those with developmental disabilities (e.g. Iwata et al., 1994), the effect of alcohol consumption on family

interaction (e.g. Leonard & Roberts, 1998), cooperation and competition (e.g. the prisoner's dilemma paradigm; Sheldon, 1999), and aggression (e.g. Bandura, 1986).

### Psychometric Considerations

Each ABO paradigm and its accompanying coding systems must be separately considered for reliability, validity, and utility. Like all psychological assessments, ABO psychometrics depend 'on the goals of assessment, the assessment settings, the methods of assessment, the characteristics of the measured variable, and the inferences that are to be drawn from the obtained measures' (Haynes & O'Brien, 2000: 201).

The psychometrics of ABO paradigms and coding systems has received little direct attention (see a special issue of *Psychological Assessment*, March 2001, for a notable exception). The validity of the ABO paradigms is implied by the results of studies using that paradigm. As such, ABO paradigms and their coding systems often have excellent validity and reported inter-rater agreement.

### Coding

Although we have described ABO as a hypothesis testing tool, in reality it is a hypothesis testing setting; coding the observed behaviours turns ABO into a true tool. Creation or use of a coding system is a theoretical act, and the following questions should be answered before proceeding: Why are you observing? What do you hope to learn? How will it impact your hypotheses (i.e. either research questions or case-conceptualization questions)? This is especially true because coding of many ABO target behaviours is difficult to do in a reliable, valid, and cost-effective manner. Interested readers should consult several excellent resources for more complete coverage (e.g. Bakeman & Gottman, 1997; Haynes & O'Brien, 2000).

### Sampling

The major sampling strategies are event sampling (the occurrence of behaviour is coded, ideally in sequential fashion), duration sampling (the length of each behaviour is recorded), interval sampling (the ABO period is divided into time blocks; during each time block, the occurrence of each

code is noted), and time sampling (intermittent observations are made, typically in a duration or interval sampling manner). Advantages and disadvantages of each are discussed in Bakeman and Gottman (1997) and Haynes and O'Brien (2000).

### *Choosing What to Code*

Some behaviours are so concrete that the observer serves more as a recorder than a coder (e.g. duration of a discrete behaviour). Other behaviours require at least some degree of inference. Such coding necessitates the use of culturally sensitive raters, using specified decision rules, to infer that a combination of situational, linguistic, paralinguistic, or contextual cues amounts to a codeable behaviour. Concrete codes are not necessarily better than social informant-inferred codes; sometimes one allows for a more valid measurement of a construct, sometimes the other does. In accord with Occam's razor, coding should be as simple as possible to reliably capture the behavioural constructs of interest.

Global (i.e. molar) coding systems make summary ratings for each code over the entire ABO (or across large time intervals). Codes tend to be few, representing behavioural classes (e.g. negativity). Microbehavioural (i.e. molecular) systems code behaviour as it unfolds over time, and tend to have many fine-grained behavioural codes (e.g. eye contact, criticize, whine).

Topographical coding systems measure the occurrence of a behaviour (including, potentially, its duration). Dimensional coding systems measure the intensity of the behaviour. Microbehavioural systems tend to be topographical; although global systems tend to be use-rating scales, they may summarize frequency rather than intensity. Dimensional coding of intensity, especially on a point-by-point basis, has been used sparingly in ABO.

### **Analyses**

ABO frequently uses single subject multiple baseline designs. Data are plotted and visually inspected for trends.

Statistical analysis of ABO data uses standard statistical tools. Between-groups hypotheses about behavioural frequencies are tested with ANOVA, continuous association hypotheses are tested with correlations or regressions.

When functional relations are of interest, testing how interactions unfold across time becomes important. Functional relation hypotheses can be addressed with conditional probabilities or with sequential analysis, which is similar to conditional probability analysis but which allows for significance testing. Dimensional data assessed continuously would use time-series analysis instead of sequential analysis.

### **CONCLUSIONS**

ABO can be a good theory-testing tool because (depending on exactly how it is employed) it minimizes inferences needed to assess behaviour, it can facilitate formal or informal functional analysis, provide the assessor with experimental control of situational factors, facilitate the observation of otherwise unobservable behaviours, and provide an additional mode of assessment in a multimodal strategy (e.g. questionnaires, interviews, observation). Finally, because the assessor can set up a situation that increases the probability that behaviours of interest will occur during the observation period, ABO can be high in clinical utility and research efficiency.

Like any tool, however, ABO's usefulness depends on its match to the resources and needs of the person considering using it. ABO can be a time, labour, and money-intensive assessment strategy. The use of research-tested protocols/coding is often impractical in clinical settings; adaptations of empirically supported ABO methodology in clinical settings may render them unreliable and of dubious validity. The conditional nature of validity may make it difficult to generalize ABOs to the broad variety of real-world settings. Finally, the less naturalistic the ABO situation, the more nagging the concerns about external validity.

### *Acknowledgements*

Preparation of this chapter was supported by the National Institutes of Mental Health (Grant R01MH57779) and National Center for Injury Prevention and Control, Centers for Disease Control and Prevention (Grant R49CCR218554-01)

## References

- Ainsworth, M.D.S., Blehar, M.C., Waters, E. & Wall, S. (Eds.) (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Hillsdale, NJ: Erlbaum.
- Bakeman, R. & Gottman, J.M. (1997). *Observing Interaction: An Introduction to Sequential Analysis* (2nd ed.). New York: Cambridge University Press.
- Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bargh, J.A., Chen, M. & Burrows, L. (1996). Automaticity of social behaviour: direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71, 230–244.
- Haynes, S.N. & O'Brien, W.H. (2000). *Principles and Practice of Behavioural Assessment*. New York: Kluwer.
- Heyman, R.E. (2001). Observation of couple conflicts: clinical assessment applications, stubborn truths, and shaky foundations. *Psychological Assessment*, 13, 5–35.
- Iwata, B.A., Pace, G.M., Dorsey, M.F., Zarcone, J.R., Vollmer, B., & Smith, J. (1994). The functions of self-injurious behaviour: an experimental-epidemiological analysis. *Journal of Applied Behaviour Analysis*, 27, 215–240.
- Leonard, K.E., & Roberts, L.J. (1998). The effects of alcohol on the marital interactions of aggressive and nonaggressive husbands and their wives. *Journal of Abnormal Psychology*, 107, 602–615.
- McGlynn, F.D. & Rose, M.P. (1998). Assessment of anxiety and fear. In Bellack, A.S. & Hersen, M. (Eds.), *Behavioural Assessment: A Practical Handbook* (4th ed., pp. 179–209). Needham Heights, MA: Allyn & Bacon.
- Norton, P.J. & Hope, D.A. (2001). Analogue observational methods in the assessment of social functioning in adults. *Psychological Assessment*, 13, 59–72.
- Reid, J.B. (Ed.) (1978). *A Social Learning Approach, Vol. 2: Observation in Home Settings*. Eugene, OR: Castalia.
- Roberts, M.W. (2001). Clinic observations of structured parent-child interaction designed to evaluate externalizing disorders. *Psychological Assessment*, 13, 46–58.
- Sheldon, K.M. (1999). Learning the lessons of tit-for-tat: even competitors can get the message. *Journal of Personality and Social Psychology*, 77, 1245–1253.
- Slep, A.M.S. & O'Leary, S.G. (1998). The effects of maternal attributions on parenting: an experimental analysis. *Journal of Family Psychology*, 12, 234–243.
- Tice, D.M., Bratslavsky, E. & Baumeister, R.F. (2001). Emotional distress regulation takes precedence over impulse control: if you feel bad, do it! *Journal of Personality and Social Psychology*, 80, 53–67.

Richard E. Heyman and Amy M. Smith Slep

## Related Entries

OBSERVATIONAL METHODS (GENERAL), OBSERVATIONAL TECHNIQUES IN CLINICAL SETTINGS

# A

## ANGER, HOSTILITY AND AGGRESSION ASSESSMENT

## INTRODUCTION

Over the last 25 years, interest in measuring the experience, expression, and control of anger has been stimulated by evidence that anger, hostility and aggression were associated with hypertension and cardiovascular disease (Williams, Barefoot, & Shekelle, 1985; Dembroski, MacDougall, Williams, & Haney, 1984). While definitions of anger-related constructs are often inconsistent and ambiguous, the experience and expression of anger are typically encompassed in definitions of hostility and aggression. Clearly, anger is the most fundamental of these overlapping constructs.

On the basis of a careful review of the research literature on anger, hostility and aggression, the following definitions of these constructs were proposed by Spielberger *et al.* (1983: 16):

Anger usually refers to an emotional state that consists of feelings that vary in intensity, from mild irritation or annoyance to intense fury and rage. Although hostility involves angry feelings, this concept has the connotation of a complex set of attitudes that motivate aggressive behaviours directed toward destroying objects or injuring other people. The concept of aggression generally implies destructive or punitive behaviour directed towards other persons or objects.

The physiological and behavioural manifestations of anger, hostility and aggression have been investigated in numerous studies, but until recently; angry feelings have been largely ignored in psychological research. Consequently, psychometric measures of anger, hostility and aggression generally do not distinguish between feeling angry, and the expression of anger and hostility in aggressive behaviour. Most measures of anger-related constructs also fail to take the state–trait distinction into account, and confound the experience and expression of anger with situational determinants of angry behaviour. A coherent theoretical framework that recognizes the difference between anger, hostility and aggression as psychological constructs, and that distinguishes between anger as an emotional state and individual differences in the experience, expression and control of anger as personality traits, is essential for guiding the construction and cross-cultural adaptation of anger measures.

#### ASSESSMENT OF ANGER: MEASURING STATE–TRAIT AND THE EXPRESSION AND CONTROL OF ANGER

The State–Trait Anger Expression Inventory (STAXI) was developed to measure the experience, expression and control of anger (Spielberger et al., 1985; Spielberger, Krasner, & Solomon, 1988). The State–Trait Anger Scale (STAS) was constructed to assess the intensity of anger as an emotional state, and individual differences in anger proneness as a personality trait (Spielberger et al., 1983). State anger was defined as ‘... an emotional state marked by subjective feelings that vary in intensity from mild annoyance or irritation to intense fury or rage, which is generally accompanied by muscular tension and arousal of the autonomic nervous system’. Trait Anger refers to individual differences in the disposition to experience angry feelings. The STAS Trait-Anger Scale evaluates how frequently State anger is experienced.

Recognition of the importance of distinguishing between the experience and expression of anger stimulated the development of the Anger Expression (AX) Scale (Spielberger et al., 1985). The AX Scale assesses how often anger is suppressed (anger-in) or expressed in aggressive behaviour (anger-out). The instructions for responding to the AX Scale differ markedly

from the traditional trait-anger instructions. Rather than directing subjects to respond according to how they generally feel, they are instructed to report on how often they react or behave in a particular manner when they feel “angry or furious” (e.g. ‘I say nasty things’; ‘I boil inside, but don’t show it’) by rating themselves on the same 4-point frequency scale that is used with the Trait-Anger Scale.

The identification of anger control as an independent factor stimulated the construction of a scale to assess the control of angry feelings (Spielberger et al., 1988). The content of three of the 20 original AX Scale items (e.g. control my temper, keep my cool, calm down faster), which were included to assess intermediate levels of anger-expression as a unidimensional bipolar scale, guided the generation of additional anger control items (Spielberger et al., 1985).

The last stage in the construction of the STAXI was stimulated by the research of psycholinguists, who identified English metaphors for anger, which called attention to the need to distinguish between two different mechanisms for controlling anger expression (Lakoff, 1987). The prototype of the anger metaphor was described as a hot liquid in a container, where blood was the hot liquid and the body was the container. The intensity of anger as an emotional state is considered analogous to the variations in the temperature of the hot liquid. The metaphor, boiling inside, has the connotation of an intense level of suppressed anger; blowing off steam connotes the outward expression of angry feelings; keeping the lid on implies controlling intense anger by preventing the outward expression of aggressive behaviour. Thus, Lakoff’s (1987) anger metaphors suggested two quite different mechanisms for controlling anger: keeping angry feelings bottled up to prevent their expression, and reducing the intensity of suppressed anger by cooling down.

In the original STAXI scale, the content of all but one of the eight Control items was related to controlling anger-out (e.g. ‘I Control my temper’). Therefore, a number of new items were constructed to assess the control of anger-in by reducing the intensity of suppressed anger (Sydeman, 1995). The content of these items described efforts to calm down, cool off or relax when a person feels angry or furious. Factor analyses of the responses of large samples of male and female adults to the anger-control items identified two anger control factors for both sexes: Anger Control-In and Control-Out.

## OTHER MEANS OF MEASURING ANGER

- (a) *Novaco Anger Inventory* (Novaco, 1975, 1977): this inventory is made up of 80 anger-provoking situations. Its reliability coefficient is rather high at 0.96, within a sample of 353 students (Biaggio, 1980). This inventory has shown remarkable differences between psychiatric patients with anger problems and normal population (Novaco, 1977).
- (b) *Multidimensional Anger Inventory – MAI* (Siegel, 1985): it is made up of 38 items, with a five-point ‘Likert’ scale. It measures anger-in with ruminations, anger-out with ruminations, anger-incited situations and hostile attitudes. It also provides a comprehensive index of anger in all its manifestations.
- (c) *Harburg Anger In/Anger Out Scale* (Harburg Erfurt, Chape, Hauenstein, Schull & Schork, 1973): this scale consists of a series of hypothetical interpersonal situations which may generate anger. It is a two-dimensional scale: it measures anger-in and anger-out, whereas at the same time it also provides a means of measuring resentment and reasoning.
- (d) *Anger Self-Report Scale – ASR* (Zelin, Alder & Myerson, 1972): it consists of 74 items with a six-point ‘Likert’ scale. It measures anger awareness and anger expression. The anger expression scale makes a distinction between different sub-scales or levels of expression. This test has shown an average reliability coefficient in samples of psychiatric patients and students.
- (e) *Anger Control Inventory*: this test consists of 134 items combining ten anger-provoking situations and six scales of anger response which describe cognitive, physiological and behavioural characteristics. Its reliability coefficient varies from 0.55 to 0.89 (Hoshmand & Austin, 1987).
- (f) *Framingham Anger Scale*: these are self-report scales developed during the Framingham Project (Haynes, Levine, Scotch, Feinleib & Kannel, 1978). These scales are used to measure anger symptoms, anger-in and anger-out, and anger expression.
- (g) *Subjective Anger Scale – SAS* (Knight, Ross, Collins & Parmenter, 1985): it measures the patient’s proneness to experience anger by means of nine different situations and four scales of anger response.
- (h) The *Anger Situation Scale* and the *Anger Symptom Scale* (Deffenbacher, Demm & Brander, 1986). They describe in detail the two worst, ongoing angering situations, and also, the two most salient physical signs of anger.

## ASSESSMENT OF HOSTILITY

- 1 Cook-Medley Ho Scale (Cook and Medley, 1954). The Ho scale is a part of the MMPI. This scale which has been widely used to measure hostility, in research on Health Psychology. However, its development has been shaped through research on rapport between teachers and students. Barefoot, Dodge, Peterson, Haney & Williams (1989) identified two subsets of items, which represent cognitive, affective and behavioural manifestations of hostility. Another subset of items reflects the tendency to elicit hostile intent from other people’s behaviour. The remaining subset of items identifies social avoidance. Its test-retest reliability has been of 0.84 in a four-year interval (Shekelle, Gale, Ostfeld & Paul, 1983).
- 2 The Buss–Durkee Hostility Inventory – BDHI (Buss & Durkee, 1957): This scale consists of 75 items, with a true–false response scheme. It is one of the most comprehensive instruments to measure hostility. It is made up of seven sub-scales: Assault, Indirect Hostility, Irritability, Negativity, Resentment, Suspicion and Verbal Hostility. The factorial analysis of these scales reveals two well-defined factors. One of them reflects hostile expression and the other experiential aspects of hostility. Its test-retest reliability, given a two-week interval, is 0.82 for the total hostility measurement (Biaggio, Supplee & Curtis, 1981).
- 3 Factor L: It is a sub-scale of a more general personality inventory Cattell’s 16 P.F. (Cattell, Eber & Tatsuoka, 1970). It is



described as a measure of suspiciousness versus trust.

### CROSS-CULTURAL ASSESSMENT OF ANGER, HOSTILITY AND AGGRESSION: THE SPANISH MULTICULTURAL STATE-TRAIT ANGER EXPRESSION INVENTORY

Spanish is spoken not only in Spain, but also in more than 20 countries in Central and South America and the Caribbean, and by more than 25 million native speakers of Spanish who reside in the United States. Although Spanish is the primary language in most of Latin America and for many Hispanic residents in the U.S., the indigenous cultures of these people often have profound effects on the Spanish they speak, and on the development of personality characteristics that influence their behaviour. Therefore, it is important to recognize the exceptionally complex social and cultural diversity of Hispanic populations, and the fact that language differences between these groups may outweigh the similarities. Consequently, in adapting English measures of emotion and personality for use in Spanish-speaking cultures, care must be taken to ensure that the key words and idiomatic expressions used for assessing anger-related concepts have essentially the same meaning in different Hispanic cultural groups.

The STAXI-2 (Spielberger, 1999) was adapted to measure the experience, expression and control of anger in culturally diverse populations in Latin America, and in Spanish-speaking sub-cultures in the United States (Moscoso & Spielberger, 1999a). Toward achieving this goal, the Spanish Multicultural State-Trait Anger Expression Inventory (STAXI-SMC) was designed to measure essentially the same dimensions of anger that are assessed with the revised STAXI-2. Scales and subscales were constructed to assess the following dimensions with the STAXI-SMC: (a) State Anger, with sub-scales for assessing Feeling Angry and Feel Like Expressing Anger; (b) Trait Anger, with sub-scales for measuring Angry Temperament and Angry Reaction; and (c) trait scales for measuring four dimensions of anger expression and control: anger-in, anger-out, and the control of anger-in and anger-out (Moscoso & Spielberger, 1999b).

Factor analyses of responses to the 56 preliminary STAXI-SMC items confirmed the hypothesized structural properties of the inventory. The eight factors that were identified corresponded quite well with similar factors in the STAXI-2. These included two S-Anger factors, two T-Anger factors, and four anger expression and control factors (Moscoso & Spielberger, 1999a). In separate factor analyses of the S-Anger items, two distinctive factors were identified for both males and females: "Feeling Angry" and "Feel Like Expressing Anger". However, gender differences in the strength of the item loadings on these factors raised interesting questions with regard to how Latin American men and women may differ in the experience of anger. For females, the "Feeling Angry" factor accounted for 73% of the total variance, while this factor accounted for only 19% of the variance for males. In contrast, the "Feel Like Expressing Anger" factor accounted for 70% of the total variance of the males, but only 13% for females.

The factor analyses of the T-Anger STAXI-SMC items also identified separate Angry Temperament and Angry Reaction factors, providing strong evidence that the factor structure for this scale was similar to that of the STAXI-2. Factor analyses of the STAXI-SMC anger expression and control items identified the same four factors as in the STAXI-2. The items designed to assess anger-in and anger-out, and the control of anger-in and anger-out, had high loadings on the corresponding anger expression and control factors, which were similar for both sexes. The alpha coefficients for the STAXI-SMC State and Trait Anger scales and sub-scales, and the anger expression and anger control scales, were reasonably high, indicating that the internal consistency of these scales was satisfactory.

In summary, the results of the factor analyses of responses of the Latin American subjects to the STAXI-SMC items of the Latin American subjects identified eight factors that were quite similar to those found for the STAXI-2. Factor analyses of the anger expression and control items also identified the same four factors that are found in the STAXI-2. Thus, the multi-dimensional factor structure of the STAXI-SMC for the Latin American respondents was remarkably similar to the factor structure of the English STAXI-2. The adaptation of the STAXI-2 test carried out in

Spain, using Spanish mainland natives (Miguel-Tobal, Cano-Vindel, Casado and Spielberger, 2001) is made up of 49 items with a similar factorial structure and the same sub-scales.

## FUTURE PERSPECTIVES AND CONCLUSIONS

Over the last quarter century, interest in measuring the experience, expression and control of anger has been stimulated by evidence that anger, hostility and aggression were associated with health problems and life-threatening disease. While definitions of anger-related constructs are often inconsistent and ambiguous, the experience and expression of anger are typically encompassed in definitions of hostility and aggression. Clearly, anger is the most fundamental of these overlapping constructs.

A sound theoretical framework that recognizes the difference between anger, hostility and aggression, and that distinguishes between anger as an emotional state and hostility in the experience, expression and control of anger as personality traits, is essential for guiding the construction of anger measures and cross-cultural adaptation.

In the cross-cultural adaptation of anger measures, it is essential to have equivalent conceptual definitions in the source and target languages that distinguish between the experience of anger as an emotional state, and hostility in the expression and control of anger as personality

traits. The construction and development of the Spanish Multicultural State-Trait Anger Expression Inventory was guided by definitions of state and trait anger and anger-expression and anger-control as these constructs were conceptualized in the STAXI-2.

Factor analyses of the items constructed for the STAXI-SMC identified eight factors that were quite similar to the factor structure of the STAXI-2. Research on the STAXI-2 and the STAXI-SMC clearly indicates that anger and hostility as psychological constructs can be meaningfully defined as emotional states that vary in intensity, and as complex personality traits with major components that can be measured empirically.

The importance anger and hostility have within the fields of Psychology, and particularly of Health, asks for precise means of assessment and measurement. Nowadays, there are some remarkable self-report tests available, which provide evidence of cross-cultural validity. However, in order to develop more accurate means of anger assessment, it is advisable to use and develop lesser known techniques of behavioural observation, such as self-monitoring (e.g. Meichenbaum & Deffenbaker, 1988) and interviewing. Also, research in the fields of physiological measurement and cognitive variables of anger (appraisals, attributions etc.) needs to be given a further boost. Measurement issues are a fundamental part of the research and the study of the hostility and the anger.

Table 1. Summary table of anger assessment scales

Scales	Assessment of					Hostility
	Anger	Expression of Anger	Anger-In	Anger-Out	Anger-Control	
STAXI	Yes	Yes	Yes	Yes	Yes	
Novaco Anger Inventory	Yes					
Multi-dimensional Anger Inventory	Yes		Yes	Yes		Yes
Harburg Anger-in /Anger-out	Yes		Yes	Yes		
Anger Self-Report Scale	Yes	Yes				
Anger Control Inventory	Yes					
Framingham Anger Scale	Yes	Yes	Yes	Yes		
Subjective Anger Scale	Yes					
Anger Situation Scale	Yes					
Anger Symptom Scale	Yes					
Cook-Medley Ho						Yes
Buss-Durkee Hostility Inventory						Yes
Factor L						Yes

## References

- Barefoot, J.C., Dodge, K.A., Peterson, B.L., Dahistrom, W.G. & Williams, R.B. (1989). The Cook-Medley hostility scale: item content and ability to predict survival. *Psychosomatic Medicine*, 51, 46–57.
- Biaggio, M.K., Suplee, K. & Curtis, N. (1981). Reliability and validity of four anger scales. *Journal of Personality Assessment*, 45, 639–648.
- Buss, A.H. & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, 42, 155–162.
- Cattell, R.B., Eber, H.W. & Tatsuoka, M.M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, IL: Institute for Personality and Ability Testing.
- Cook, W.W. & Medley, D.M. (1954). Propose hostility and pharasaic-virtue scales for the MMPI. *Journal of Applied Psychology*, 38, 414–418.
- Deffenbacher, J.L., Demm, P.M. & Brandon, A.D. (1986). High general anger. *Behaviour Research and Therapy*, 24, 481–489.
- Dembroski, T.M., McDougall, J.M., Williams, R.B. & Haney, T.L. (1984). Components of type A, hostility, and anger-in: relationship to angiographic findings. *Psychosomatic Medicine*, 47, 219–233.
- Harburg, E., Erfurt, J.C., Chape, C., Hauenstein, L.S., Schull, W.J. & Schork, M.A. (1973). Socio-ecological stressor areas and black-white blood pressure: Detroit. *Journal of Chronic Disease*, 26, 595–611.
- Haynes, S.N., Levine, S., Scotch, N., Feinleib, M. & Kannel, W.B. (1978). The relationship of psychological factors to coronary heart disease in the Framingham study: I. Methods and risk factors. *American Journal of Epidemiology*, 107, 362–363.
- Hoshmand, L.T. & Austin, G.W. (1987). Validation studies of a multifactor cognitive-behavioural Anger Control Inventory. *Journal of Personality Assessment*, 51, 417–432.
- Knight, R.G., Ross, R.A., Collins, J.I. & Parmenter, S.A. (1985). Some norms, reliability and preliminary validity data for an S-ROS: inventory of anger: the Subjective Anger Scale (SAS). *Personality and Individual Differences*, 6, 331–339.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What categories reveal about the mind*. Chicago: The University of Chicago Press.
- Meichenbaum, D.H. & Deffenbacher, J.L. (1988). Stress inoculation training. *The Counselling Psychologist*, 16, 69–90.
- Miguel-Tobal, J.J., Cano-Vindel, A., Casado, M.I. & Spielberger, C.D. (2001). *Inventario de Expresión de Ira Estado Rasgo – STAXI – 2: Spanish Adaptation*. Madrid: TEA
- Moscoso, M.S. & Spielberger, C.D. (1999a). Evaluación de la experiencia, expresión y control de la cólera en Latinoamérica. *Revista Psicología Contemporánea*, 6(1), 4–13.
- Moscoso, M.S. & Spielberger, C.D. (1999b). Measuring the experience, expression, and control of anger in Latin America: the Spanish multi-cultural State-Trait Anger Expression Inventory. *Interamerican Journal of Psychology*, 33(2).
- Novaco, R.W. (1975). *Anger Control: The Development and Evaluation of an Experimental Treatment*. Lexington: D.C. Heath
- Novaco, R.W. (1977). Stress inoculation: a cognitive therapy for anger and its application to a case of depression. *Journal of Consulting and Clinical Psychology*, 45, 600–608.
- Shekelle, R.B., Gale, M., Ostfeld, A.M. & Paul, O. (1983). Hostility, risk of coronary heart disease, and mortality. *Psychosomatic Medicine*, 45, 109–114.
- Siegel, S.M. (1985). The multidimensional anger inventory. In Chesney, M.A. & Rosenman, N.H. (Ed.), *Anger and Hostility in Cardiovascular and Behavioural Disorders*. Washington, DC: Hemisphere.
- Spielberger, C.D., (1999). *State-Trait Anger Expression Inventory – 2*. Odessa, FL: Psychological Assessment Resources.
- Spielberger, C.D., Krasner, S.S. & Solomon, E.P. (1988). The experience, expression and control of anger. In Janisse, M.P. (Ed.), *Health Psychology: Individual Differences and Stress* (pp. 89–108). New York: Springer Verlag.
- Spielberger, C.D., Johnson, E.H., Russell, S.F., Crane, R.J., Jacobs, G.A. & Worden, T.J. (1985). The experience and expression of anger: construction and validation of an anger expression scale. In Chesney, M.A. & Rosenman, R.H. (Eds.), *Anger and hostility in Cardiovascular and Behavioural Disorders*. New York: McGraw-Hill/Hemisphere.
- Spielberger, C.D., Jacobs, G.A., Russell, S.F. & Crane, R.S. (1983). Assessment of anger: the State-Trait Anger Scale. In Butcher, J.N. & Spielberger, C.D. (Eds.), *Advances in Personality Assessment* (Vol. 2, pp. 159–187). Hillsdale, NJ: Erlbaum.
- Sydemann, S.J. (1995). *The Control of Suppressed Anger*. Unpublished Master's Thesis, University of South Florida, Tampa.
- Williams, R.B., Barefoot, J.C. & Shekelle, R.B. (1985). The health consequences of hostility. In Chesney, M.A. & Rosenman, R.A. (Eds.), *Anger and Hostility in Cardiovascular and Behavioural Disorders* (pp. 173–185). New York: Hemisphere/McGraw-Hill.
- Zelin, M.I., Adler, G. & Myerson, P.G. (1972). Anger self-report: an objective questionnaire for the measurement of aggression. *Journal of Consulting and Clinical Psychology*, 39, 340.

Manolete S. Moscoso and  
Miguel Angel Pérez-Nieto

## Related Entries

TYPE A: A PROPOSED PSYCHOSOCIAL RISK FACTOR FOR CARDIO-VASCULAR DISEASES, DANGEROUS/VIOLENCE POTENTIAL BEHAVIUR

# A ANTISOCIAL DISORDERS ASSESSMENT

## INTRODUCTION

Using a broad definition, Antisocial Disorders may be defined as pervasive, maladaptive behaviours that violate the norms and rules of a group or society, causing social impairment or distress to others. Currently, the classification and assessment of antisocial disorders may follow: (a) the medical model or (b) the dimensional model:

- The medical model uses a categorical approach in which the presence of a variety of diagnostic criteria, such as persistent violations of social norms (including lying, stealing, truancy, inconsistent work behaviour and traffic arrests) is evaluated by experts (clinicians). This model relies on diagnostic criteria as outlined in the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, 1994) and ICD (International Classification of Diseases, 1993).
- The dimensional model evaluates antisocial disorders along a continuum of development, from normal to pathological, focusing on behavioural and trait dimensions, and identifying clusters of highly interrelated behaviours and traits.

There is agreement among researchers about the development of antisocial behaviour: it begins early in life (infancy) with aggressive and oppositional behaviours (e.g. conduct problems), gradually advances toward more significant expressions of antisocial acts (e.g. vandalism, stealing, truancy, lying, substance abuse) during adolescence, and lastly, progresses to extreme forms of delinquency in adult life. The most recent longitudinal and retrospective studies (Patterson, Reid & Dishion, 1992) suggest that the 'early starters' (childhood and preadolescence) are at greater risk for adult involvement in delinquent acts and are more likely to move toward more serious offences that lead to a 'criminal career' compared to the 'later starters' (adolescence).

A variety of methods are used for assessing antisocial disorders, these include: self-report instruments, others' ratings, clinical interviews (structured and semi-structured), and direct behavioural observation (see Table 1).

## CHILD AND ADOLESCENT ASSESSMENT

In order to establish the severity of antisocial behaviours during childhood and adolescence, it is important: (a) to determine the *age of onset*; (b) to evaluate the *frequency* of aggressive acts; (c) to establish the *variety* of antisocial behaviours; and (d) to observe them in *multiple settings* (family, peers, school and community). As a necessary complement to this assessment, it is also important to evaluate other aspects of the individual's functioning in order to rule out the co-occurrence of other psychological disturbances.

For children and adolescents, the terms *conduct disorders* and *conduct problems* (aggressive and oppositional behaviours) may be used interchangeably. It is important to note that conduct disorders have different prevalence rates for boys and girls: 6 to 16% for boys, and 2 to 9% for girls.

In recent years, more complete assessment procedures have been developed to cover a full range of childhood and adolescent behaviours directly and indirectly linked to antisocial behaviours in different contexts. The advantages of these assessment procedures are: (a) to have a complete picture of child and adolescent functioning for the purpose of differential diagnosis and (b) to collect data to provide empirical and theoretical support of the instruments used.

## Instruments For Child and Adolescent Assessment

Here we present only a few of the numerous instruments that can be used for measuring

Table 1. Assessment of antisocial disorder

Target	Informant	Model	Measure		
			Scale	Clinical Interviews	Observation
Children/ Adolescents	Self	Dimensional	<ul style="list-style-type: none"> <li>– Youth Self-Report (Achenbach &amp; Edelbrock, 1991c)</li> <li>– Behaviour Assessment System of Children-Self-report of Personality Scale (Reynolds &amp; Kamphaus, 1992)</li> </ul>		
	Others	Dimensional	<ul style="list-style-type: none"> <li>– Revised Behavior Problem Checklist (Quay &amp; Peterson, 1983)</li> <li>– Child Behavior Checklist (Achenbach, 1991a)</li> </ul>		<ul style="list-style-type: none"> <li>– Direct Observation Form (Achenbach, 1986)</li> <li>– Behavior Assessment System of Children – Student observation system (Reynolds &amp; Kamphaus, 1992)</li> <li>– Family Interaction Coding System (Reid, 1978)</li> <li>– Observation of Peer Interactions (Dodge, 1983)</li> </ul>
Children/ Adolescents	Self	Categorical	<ul style="list-style-type: none"> <li>– Teacher Report Form (Achenbach, 1991b)</li> <li>– Behavior Assessment System of Scale (Reynolds &amp; Kamphaus, 1992)</li> <li>– Devereux Scales of Mental Disorders (Naglieri, Lebuffe &amp; Pfeiffer, 1984)</li> <li>– Eysenck Personality Questionnaire (Eysenck &amp; Eysenck, 1993)</li> <li>– Minnesota Multiphasic Personality Inventory – II (Butcher et al., 1989)</li> </ul>		

*(continued)*

Table 1. Continued

Target	Informant	Model	Measure		
			Scale	Clinical Interviews	Observation
Adults	Self	Dimensional	<ul style="list-style-type: none"> <li>- Millon Clinical Multiaxial Inventory II (Millon, 1994)</li> <li>- Assessment of DSM-IV Personality Disorder Questionnaire</li> <li>- Antisocial Personality Questionnaire (Blackburn &amp; Fawcett, 1999)</li> </ul>	<ul style="list-style-type: none"> <li>- Hare Psychopathy Checklist - Revised (Hare, 1991)</li> </ul>	
		Categorical		<ul style="list-style-type: none"> <li>- International Personality Disorder Examination (Loranger, Sartorius &amp; Janca, 1996).</li> <li>- Structured Clinical Interview for DSM-IV (First et al., 1997)</li> <li>- Structured Interview For DSM-IV Personality Disorder (Pfohl, Blum &amp; Zimmerman, 1995)</li> </ul>	

antisocial behaviour. We included those that provide a comprehensive assessment of different psycho-social domains and those that are in some way representative of the field of antisocial behaviour, both at the level of research and intervention.

### *Revised Behaviour Problem Checklist (RBPC)*

The RBPC (Quay & Peterson, 1983) represents one of the first attempts to empirically classify childhood and adolescent disorders. The Revised Behaviour Problem Checklist covers the ages 5 to 17 years, and is available in two versions, one for teachers and one for mothers. It represents a revision of the original Behaviour Problems Checklist and now comprises six scales: Conduct Disorder, Socialized Aggression, Attention Problems-Immaturity, Anxiety-Withdrawal, Psychotic Behaviour, Motor Tension Excess. It allows one to distinguish between “socialized” and “under-socialized” conduct disorders. Socialized makes reference to antisocial behaviour within deviant peer group, unsocialized refers to impulsivity and irritability.

### *Child Behaviour Checklist (CBCL)*

The CBCL (Achenbach, 1991a) (parent form), together with Youth Self-report (YSR; Achenbach, 1991c), Teacher Report Form (TRF; Achenbach, 1991b) and Direct Observation Form (DOF; Achenbach, 1986), is one of the most comprehensive evaluation systems for childhood and adolescent psychopathology. It was developed by Achenbach in order to derive syndromes empirically and to allow for comparisons among different informants and cultures. The four forms share item content and can be used together to establish cross-contexts consistency.

They cover an age range of 4 to 18 years. The CBCL includes problem behaviour and social competence scales. Problem behaviour scales are: Aggressive behaviours, Delinquency, Anxiety/Depression, Somatic Complaints, Attention problems, Thought Problems and Social Withdrawal) and a social competence scale. In addition, there is a Sexual Problem Behaviour scale for children between 4 and 11 years old. It is also possible to derive two broader dimensions: Internalizing and Externalizing.

### *Behaviour Assessment System of Children (BASC)*

The BASC (Reynolds & Kamphaus, 1992) is a multi-method, multidimensional assessment instrument aimed at evaluating the behaviours and self-perceptions of children aged 4 to 18 years old. Similar to the CBCL, it has several different versions: self-report, teacher rating scale, parent rating scale, student observation system and structured developmental history. The *Self-Report of Personality Scale* (6–18 years) is comprised of the following subscales: Anxiety, Attitude to School, Attitude to Teachers, Atypicality, Depression, Interpersonal Relations, Locus of Control, Relations with Parents, Self-Esteem, Self-Reliance, Sensation Seeking, Sense of Inadequacy, Social Stress, and Somatization. The *Teacher and Parent Rating Scales* (different forms for 4–5 years, 6–11 years, and 12–18 years) are comprised of the following subscales: Aggression, Conduct Problems, Attention Problems, Hyperactivity, Anxiety, Atypicality, Depression, Somatization, Withdrawal, Learning Problems, Leadership, Social Skills, Study Skills, Adaptability. The *Student Observation System* assesses student’s behaviour in the classroom such as inappropriate movement, inappropriate attention and work on school subjects.

### *Devereux Scales of Mental Disorders (DSMD)*

The DSMD (Naglieri, Lebuffe & Pfeiffer, 1994) is designed to measure the risk for emotional and behavioural disorders in children between 5 and 18 years (5–12 years; 13–18 years). It relies on the DSM-IV, and has both teacher and parent forms. It includes scales to assess Problem behaviours, Delinquency, Attention, Depression and Anxiety, Autism and Acute Problems. It provides three different composites: Internalizing, Externalizing and Critical Pathology.

### *Diagnostic Interview Schedule for Children – Child Interview (DISC-C)*

The DISC-C (Costello, Edelbrock, Kalas, Kessler & Klaric, 1982) is a structured diagnostic interview that covers a broad range of DSM-IV diagnoses in children. Child, parent and teacher forms are available. Areas covered

include: Behaviour/Conduct Disorder, Attention Deficit Disorder, Affective/Neurotic Anxiety, Fears and Phobias, Obsessive-Compulsive Disorder, Schizoid/Psychotic Disorders, Affective (depression) Disorders.

### *Family Interaction Coding System (FICS)*

The FICS (Reid, 1978) is an assessment instrument used to register interactions between family members. This coding system enables researchers and family therapists to monitor clinical cases, systematically assess the outcome of family intervention programs, and builds a database for studying aggressive antisocial behaviours exhibited by children. It is composed of 29 categories, but the Total Aversive Behaviour score (such as physical negative, tease, noncompliance, destructiveness etc.) is mostly used (Reid, 1978).

### *Observation of Peer Interactions*

This instrument (Dodge, 1983) is used to register interactions among peers between the ages of 5 to 8 years. It has five categories: solitary active, interactive play, verbalizations, physical contacts with peers, and interactions with adult leaders within the group. This system is associated with dimensions of social status (rejection, popularity), and therefore may be useful to obtain a more complete assessment of peer interactions.

## ADULT ASSESSMENT

Albeit with some differences, antisocial disorders may correspond with the Antisocial Personality Disorder (APD) classification of DSM-IV and the Dissocial Personality Disorder classification of ICD-10. APD is characterized by criminal and antisocial behaviour, and also by deceitfulness, lack of remorse, disregard for the safety of others (DSM-IV, 1994), low tolerance for frustration and a low threshold for discharge of aggression (ICD-10, 1993). The emphasis is placed on a failure to conform to social norms, and on impulsivity and irresponsibility. Although it was excluded from recent classifications of mental disorders, the assessment of Psychopathy in adults with antisocial

disorders may be informative for differential diagnosis and treatment purposes. Psychopathy corresponds partially to the criteria of APD, but also includes emotional/interpersonal characteristics such as glibness, superficiality, egocentricity, grandiosity, lack of empathy, manipulateness, and shallow emotions. When assessing antisocial disorders, it is also important to evaluate the co-occurrence of substance abuse, anxiety disorders and depression.

## Instruments For Adult Assessment

As for children and adolescents, numerous instruments have been developed for the assessment of adult antisocial disorders. We have selected to present instruments that combine personality assessment (dimensional model) with classic diagnostic assessment (medical or categorical model), including interviews, checklists and questionnaires aimed at identifying the criteria for Antisocial Personality Disorders as presented in the DSM-V and the ICD-10.

### *Eysenck Personality Questionnaire (EPQ-R)*

The EPQ-R (Eysenck & Eysenck, 1993) is designed to measure the three traits of Eysenck's personality model: Extraversion (E), Neuroticism (N) and Psychoticism (P). This model links types, traits and behaviour into a hierarchical system. The P trait is the primary trait implicated in the development of antisocial behaviour, with elevations on E and N being secondary. In serious antisocial behaviour, the P trait has a primary role. When E is combined with high P, poor impulse control and a weakened association between behaviour and its consequences will exacerbate the P trait predisposition. Elevated E is more frequent among juvenile delinquents, and elevated N appears in adult criminals. The Eysenck Personality Inventory is also available in a form for adolescents.

### *Minnesota Multiphasic Personality Inventory – II (MMPI-II)*

The MMPI-II (Butcher et al., 1989) is the most frequently used clinical test. It is the revised version of the MMPI. It was originally intended



for use with an adult population. The MMPI-II has 10 clinical scales, 3 validity scales and 15 content scales. The clinical scales are Hypochondriasis, Depression, Hysteria, Psychopathic Deviate, Masculinity–Femininity, Paranoia, Psychasthenia, Schizophrenia, Hypomania, and Social Introversion. The clinical scales do not discriminate clinical groups from normal groups, as the labels might suggest. Subjects who score high on specific scales, show particular behaviours and tendencies. For example, subjects scoring high on the Psychopathic Deviate Scale show disregard for social custom, shallow emotions, and an inability to learn from experience. Content scales include internalizing symptoms (somatic disorder, strange beliefs and dysfunctional ways of thinking), aggressive tendencies (dysfunctional control of behaviour, cynicism), low self-esteem, family problems, work interference and negative treatment indicators. The content scales offer behavioural descriptions that are easier to interpret than the clinical scales.

The interpretation of subject profiles must be done by experienced clinicians. Recently, an adolescent version has been developed.

#### *Antisocial Personality Questionnaire (APQ)*

The APQ (Blackburn & Fawcett, 1999) is a recently developed, short, multi-trait, self-report inventory aimed at measuring intrapersonal and interpersonal aspects of emotional dysfunction, impulse control, deviant beliefs about the self and others, and interpersonal problem behaviours related to antisocial behaviours. It was derived from another instrument previously developed for mentally disordered offenders. It comprises the following measures: Self-Control, Self-Esteem, Avoidance, Paranoid Suspicion, Resentment, Aggression, Deviance and Extraversion. It is possible to derive two second-order scales: Hostile-Impulsivity and Social Withdrawal. These two scales reflect orientations towards others and the self, respectively.

#### *Hare Psychopathy Checklist – Revised (PCL-R)*

PCL-R (Hare, 1991) is a single construct rating scale that uses a semi-structured interview,

case-history information and specific diagnostic criteria to provide a reliable and valid estimate of the degree to which an offender or forensic psychiatric patient matches the traditional (prototypical) conception of a psychopath. The PCL-R evaluates emotional and interpersonal characteristics of psychopathy and social deviance.

#### *Millon Clinical Multiaxial Inventory II (MCMI-II)*

The MCMI-II (Millon, 1994) is composed of 24 self-administered scales, and is designed to measure 14 personality styles, grouped into (a) Clinical Personality Patterns (schizoid, avoidant (depressive), dependent, histrionic, narcissistic, antisocial (sadistic), compulsive and negativistic (masochistic)) and (b) Severe Personality Pathology (schizotypal, borderline and paranoid). The instrument was developed to match the DSM-IV personality disorder classifications. It also comprises 10 scales measuring other clinical syndromes (such as anxiety, depression, drug-dependence and thought disorders). This instrument also has an adolescent version.

#### *International Personality Disorder Examination (IPDE)*

The IPDE (Loranger, Sartorius Janca, 1996) is a semi-structured interview designed for the assessment of both DSM-IV and ICD-10 Personality Disorders (PD). The IPDE also combines the categorical and dimensional models. Questions are arranged in sections (e.g. background information, work, self, interpersonal relationships).

#### *Other Clinical Interview for DSM-IV*

The most frequently used clinical interviews for the diagnosis of Antisocial Personality Disorder are:

- Structured Clinical Interview for DSM-IV IV(Scid II; First et al., 1997)

The SCID II is a semi-structured diagnostic interview organized by disorder which includes all DSM-IV personality disorders. A computerized administration and scoring program is available.

- Structured Interview for DSM-IV (SIDP-IV) Personality Disorder (Pfohl, Blum & Zimmerman, 1995)

The SIDP-IV consists of 160 questions grouped under 16 thematic sections, such as relationships, emotions and reactions to stressful situations. Questions are asked regarding behaviours in the last 5 years.

## FUTURE PERSPECTIVES AND CONCLUSIONS

The assessment and diagnosis of antisocial disorders should be done by experienced mental health professionals. The assessment process should include multiple methods and informants, and use standardized instruments or structured diagnostic interviews, including complete information related to the ecology of the individual (family and social context) and individual functioning.

Based on the most relevant clinical research in the area of antisociality, we may conclude that in the future the assessment must focus more on both dysfunction and skills and try to integrate the two models, dimensional and categorical, in order to better direct the diagnostic process (screening, identification and placement for intervention).

## References

- American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: APA.
- Achenbach, T.M. (1986). *Manual for the Child Behaviour Checklist – Direct Observation Form*. Burlington, VT: University of Vermont Department of Psychiatry.
- Achenbach, T.M. (1991a). *Manual for the Child Behaviour Checklist/4–18 and 1991 Profile*. Burlington (VT): University of Vermont Department of Psychiatry.
- Achenbach, T.M. (1991b). *Manual for Teacher Report's Form and 1991 Profile*. Burlington (VT): University of Vermont Department of Psychiatry.
- Achenbach, T.M. (1991c). *Manual for the Youth Self-Report and 1991 Profile*. Burlington (VT): University of Vermont Department of Psychiatry.
- Blackburn, R. & Fawcett, D. (1999). The Antisocial Personality Questionnaire: an inventory for assessing personality deviation in offenders. *European Journal of Psychological Assessment*, 14–24.

- Butcher, J.N., Dahlstrom, W.G., Graham, J.R., Tellegen, A. & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2), Manual for Administration and Scoring*. Minneapolis: University of Minnesota Press.
- Costello, A.J., Edelbrock, C.S., Kalas, R., Kessler, M. & Klaric, S. (1982). *Diagnostic Interview Schedule for Children-Child Interview*. Bethesda, MD: National Institute of Mental Health.
- Dodge, K. (1983). Behavioural antecedents of peer social status. *Child Development*, 54, 1386–1399.
- Eysenck, H.J. & Eysenck, S.B.G. (1993). *Eysenck Personality Questionnaire – Revised*. San Diego: Educational and Industrial Testing Service.
- First, M.B., Gibbon, M., Spitzer, R.L. & Williams, J.B.W. (1997). *User's Guide for the Structured Clinical Interview for DSM-IV Axis II Personality Disorders*. Washington, DC: American Psychiatric Press.
- Hare, R.D. (1991). *Manual for the Hare Psychopathy Checklist – Revised*. Toronto: Multi Health System.
- Loranger, A.W., Sartorius, N. & Janca, A. (Eds.) (1996). *Assessment and Diagnosis of Personality Disorders: The International Personality Disorder Examination (IPDE)*. New York, NY: Cambridge University Press.
- Millon, T. (1994). *Manual for the Millon Clinical Multiaxial Inventory – III*. Minneapolis: National Computer Systems.
- Naglieri, J.A., Lebuffe, P.A. & Pfeiffer, S.I. (1994). *Devereux Scales of Mental Disorders*. New York: The Psychological Corporation.
- Patterson, G.R., Reid, J.B. & Dishion, T.J. (1992). *Antisocial Boys: A Social Interactional Approach*, Vol. 4. Eugene, OR: Castalia.
- Pfohl, B., Blum, N., Zimmerman, M (1995). *Structured Interview for DSM-IV Personality (SIDP-IV)*. Iowa City, IA: The University of Iowa.
- Quay, H.C. & Peterson, D.R. (1983). *Interim Manual for the Revised Behaviour Problem Checklist*. Coral Gables, FL: Author.
- Reid, J.B. (1978). *A Social Learning Approach, Observation in Home Settings*. Vol. 2: Eugene, OR: Castalia Publishing Company.
- Reynolds, C.R. & Kamphaus, R.W. (1992). *BASC Manual*. Circle Pines, MN: American Guidance Service.
- World Health Organization. (1993). *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. Geneva: WHO.

Concetta Pastorelli and Maria Gerbino

## Related Entries

APPLIED FIELDS: CLINICAL, DANGEROUS/VIOLENCE POTENTIAL BEHAVIOUR

# A ANXIETY ASSESSMENT

## INTRODUCTION

The first assessment of individual differences is reported in the Bible in the Book of Judges, Chapter 7 on Gideon. God asked Gideon, who was battling the Midianites, to thin out his troops by rejecting individuals who were both fearful and afraid of battle. However, too many men were left, so God instructed Gideon to lead his men down to the water and used the following selection procedure. Out of 10,000 persons, 300 lapped with water from their hands, with their tongues. They were selected. The ones who knelt to drink were not.

The present chapter will focus on the assessment of anxiety as an individual differences variable; the dimensional conceptualization of anxiety. Dimensionality arises from a personality psychology tradition, in which traits and behaviours are measured psychometrically. Traits are viewed as existing on a continuum, with low levels of a trait (e.g. anxiety) at one end and high levels of the trait at the opposite end of the same continuum. In contrast to the dimensional approach is the typological or categorical conceptualization of anxiety, consistent with the medical model (Endler & Kocovski, in press). Another entry in this encyclopaedia covers the assessment of anxiety disorders.

### Definition of Anxiety

Anxiety has been conceptualized as a stimulus, as a trait, as a motive, and as a drive and has been defined 'as an emotional state, with the subjectively experienced quality of fear as a closely related emotion' (Lewis, 1970: 77). Lewis notes that the emotion is unpleasant, future-oriented, disproportional to the threat and includes both subjective and manifest bodily disturbances. There are physiological, cognitive, and behavioural components to anxiety. These give rise to the various methods of the assessment

of anxiety. It is important to first distinguish between state anxiety and trait anxiety.

### State vs. Trait Anxiety

State anxiety is the momentary experience of anxiety. Trait anxiety is a predisposition or proneness to be anxious. The distinction between state and trait anxiety was first suggested by Cicero (Before the Common Era). Spielberger (1983) suggested that conceptual clarity could be achieved in the anxiety literature by distinguishing between state and trait anxiety. There are various methods to assess state anxiety. The assessment of trait anxiety has been conducted primarily through the use of self-report measures.

### Multidimensionality of State and Trait Anxiety

Trait anxiety and state anxiety are both multidimensional constructs (Endler, 1997; Endler, Edwards, & Vitelli, 1991). There are at least six facets of trait anxiety; social evaluation, physical danger, ambiguous, self-disclosure, separation and daily routines; and two facets of state anxiety; cognitive-worry and autonomic emotional (Endler & Flett, 2001). These facets of state and trait anxiety are presented in Table 1.

### Interaction Model of Anxiety

The distinction between state and trait anxiety has achieved wide recognition in the interaction

Table 1. Anxiety assessment techniques

Anxiety	Assessment Technique
State anxiety	Self-report Behavioural Cognitive Physiological
Trait anxiety	Self-report

model of anxiety, a subset of the interaction model of personality (Endler, 1997). According to the interaction model, increases in state anxiety will result only when a situational stressor is congruent with the facet of trait anxiety under investigation. Over 80% of the tests of the multidimensional interaction model of anxiety have yielded support for the model (Endler, 1997).

### Assessment Techniques

The use of questionnaire measures has been the primary assessment technique for trait anxiety. There are multiple techniques that can be used for the assessment of state anxiety. The assessment techniques are shown in Table 1 and include self-report, behavioural, cognitive, and physiological measures. The most comprehensive method of assessing state anxiety is through a combination of the available techniques as there are individual differences in the experience of anxiety.

### SELF-REPORT MEASURES

The majority of research in the area of personality is based on self-report measures,

despite the fact that personality theory also refers to observable behaviours. Self-report questionnaires have the following advantages: they are easy to administer, results are easy to analyse, results can be compared to normative data, and results can be subjected to factor analytic techniques (as well as other advanced statistical techniques).

Commonly used self-report measures are presented in Table 2. One of the first self-report anxiety measures is the Taylor Manifest Anxiety Scale (Taylor, 1953). Since then, numerous other scales have been developed. One commonly used self-report measure of anxiety is the State-Trait Anxiety Inventory (STAI; Spielberger, 1983). The STAI assesses both state and trait anxiety as unidimensional constructs. The state and the trait scales consist of 20 items each. These scales have been shown to have high internal consistency (approximately 0.90 for both the state and trait scales) and test-retest validity for the trait scale (Spielberger, 1983).

The Endler Multidimensional Anxiety Scales (EMAS) assess both state anxiety and trait anxiety as multidimensional constructs and assess the perception of the situation (Endler, Edwards, & Vitelli, 1991). Cognitive-worry and autonomic-emotional are the two components of state anxiety assessed by the EMAS-State

Table 2. Self report measures of anxiety

Name of scale	Author/year	Psychometric properties
Anxiety Sensitivity Index	Reiss et al. (1986)	Alpha reliability = 0.88; Test-retest reliability ranges from 0.75 to 0.85 (2 week interval)
Beck Anxiety Inventory	Beck et al. (1988)	Alpha reliability = 0.92; test-retest reliability = 0.75 (1 week interval)
Endler Multidimensional Anxiety Scales (EMAS)	Endler et al. (1991)	Alpha reliabilities range from 0.89 to 0.95; test-retest reliabilities for the trait scales range from 0.60 to 0.79 (2 week interval)
EMAS-Social Anxiety Scales (EMAS-SAS)	Endler & Flett (2001)	Alpha reliabilities range from 0.92 to 0.93; test-retest reliabilities range from 0.69 to 0.77 (1 week interval)
State Trait Anxiety Inventory (STAI)	Spielberger (1983)	Alpha reliabilities range from 0.91 to 0.93; test-retest reliabilities range from 0.71 to 0.75 for the trait scale (30 day interval)
Taylor Manifest Anxiety Scale	Taylor (1953)	Test-retest reliability = 0.88 (4 week interval)

measure (20 items in total). The EMAS-Trait measures assess a predisposition to experience anxiety in the following four situational domains (15 items each): social evaluation, physical danger, ambiguous, and daily routines. Recent research has resulted in the addition of the following two situational domains: self-disclosure (to family or to friends) and separation anxiety (Endler & Flett, 2001). The alpha reliabilities of these measures have been found to be highly acceptable (ranging from 0.89 to 0.95; Endler et al., 1991). Numerous studies have been conducted which have found support for the validity of the EMAS-State, Trait, and Perception scales (Endler et al., 1991; see Endler, 1997 for a review).

Another self-report instrument commonly used to assess anxiety is the Beck Anxiety Inventory (BAI; Beck, Epstein, Brown & Steer, 1988). The BAI consists of 21 items representing two factors: somatic symptoms and subjective anxiety symptoms. It has been shown to have a high internal consistency ( $\alpha = 0.92$ ). A weakness is that the BAI does not distinguish between state and trait anxiety. Respondents are asked to report the degree to which they have been bothered by the symptoms assessed over the past week. The BAI is primarily used in clinical settings. Finally, the Anxiety Sensitivity Index consists of 16 items and assesses the fear of experiencing anxiety (Reiss, Peterson, Gursky & McNally, 1986).

## BEHAVIOURAL MEASURES

Another anxiety assessment technique is the measurement of various behaviours. The presence and frequency of certain behaviours are rated by others (e.g. clinicians, experimenters). A review of ratings by others for the purposes of clinical evaluation is beyond the scope of the present chapter. The behaviours used to represent an indication of the level of anxiety an individual is experiencing depend upon the situational domain. For example, behavioural measures of social anxiety include measurement of the maintenance of eye contact, the number of conversations initiated or amount spoken during a social encounter, hand tremors, and fidgeting (Leary, 1986). Not all of these behavioural measures are relevant for other situational domains.

Types of interaction used in behavioural observation can be classified as either artificial (i.e. a role-play situation) or naturalistic (i.e. *in vivo* observation; Glass & Arnkoff, 1989). Behaviours are often recorded in role-play situations due to the impracticality of rating people in naturalistic environments. Even within the naturalistic category, waiting-room type interactions are often used (especially for the assessment of social anxiety).

Behavioural observation techniques are less subjective on the part of the examinee than the use of self-report measures. However, the presence of the examiner in an evaluative role may affect the level of anxiety, and additionally, the examiner is responsible for determining whether the examinee's actual behaviour constitutes the behaviour being assessed. Furthermore, in an interaction type behavioural observation assessment, the behaviour of the partner (or confederate) may represent a confound. The partner may respond differently to different participants depending on variables such as the social skill level of the participant (Glass & Arnkoff, 1989). Despite these criticisms, behavioural assessment techniques for performance situations have been shown to be highly reliable.

## COGNITIVE MEASURES

Anxiety also has a cognitive component. Cognitive measures examine the thoughts an individual has. This can be done through thought-listing procedures (Cacioppo & Petty, 1981) or via a questionnaire approach. Thought-listing techniques ask participants to record thoughts in paper and pencil format while they are in an anxious situation (Cacioppo & Petty, 1981). Participants are asked not to concern themselves with spelling or grammar and not to edit the thoughts as they arise. The list of thoughts is then analysed according to such indices as content or frequency. Variations of this technique include: (i) having participants state their thoughts aloud rather than recording them and (ii) having participants watch a video of their performance and state their thoughts during the viewing.

## PHYSIOLOGICAL MEASURES

Anxiety has a physiological component, which is largely determined by the septo-hippocampal

system (behavioural inhibition system; Gray, 1996), thus allowing for the assessment of anxiety through physiological means. Among the physiological measures are the measurement of heart rate, electrodermal activity, and respiration. Additionally, blushing, assessed with a photoplethysmograph, has been used to assess social anxiety. The different physiological measures do not, however, correlate well with one another or with self-report measures (Leary, 1986).

### **Heart Rate**

Heart rate is the most commonly used physiological measure of anxiety. It is assessed either via electrodes (which can be attached to the patient's skin to the right and left of the sternum) or via sensors. The unit of measurement typically used is the number of beats per minute. This can be determined by (i) counting the number of beats per minute or, alternatively, (ii) using equipment to determine the length of the interval between heart beats and then calculating beats per minute based on that figure. These two approaches typically yield different results; however, both are used in the assessment of heart rate as an indicator of state anxiety. Heart rate has been found to be strongly correlated with self-report state anxiety in a competitive sports situation and moderately correlated with self-report state anxiety (and one item in particular which assesses heart rate) in a performance anxiety situation (Kantor, Endler, Heslegrave & Kocovski, 2000).

### **Finger Pulse Volume**

Finger pulse volume is a measure of digital vasoconstriction (Bloom & Trautt, 1977). The use of finger pulse volume to assess anxiety is based on the premise that one of the responses of the sympathetic nervous system is decreased blood flow to peripheral areas of the body. Finger pulse volume has been shown to be a valid physiological measure in social-evaluation situations.

### **Electrodermal Activity**

Another physiological measure of anxiety is sweat gland activity. The eccrine sweat glands

are innervated by the sympathetic nervous system and are located throughout the surface of the body. The primary concentration of eccrine sweat glands is in the palms of the hands and the soles of the feet. Changes in the degree of sweat gland activity can be a result of state anxiety, however, there are other variables that also play a role. For example, room temperature affects the activity of the eccrine glands, as do person variables (e.g. gender). There are, therefore, numerous variables that can affect the internal validity of a study that uses sweat gland activity as an indication of anxiety as the dependent variable. These need to be considered both in research studies and in the assessment of an individual.

Clements and Turpin (1996) assessed the sweat gland activity of participants while giving a presentation and while being a member of the audience. Sweat gland activity was found to increase prior to and during the presentation and decrease upon completion of the presentation. Levels of state anxiety were also found to be elevated during the presentation. There was, however, no relationship found between the physiological measure (sweat gland activity) and each of state and trait anxiety.

### **Respiration**

Respiration rate can also be used as a tool in the assessment of anxiety. To measure respiration rate, a stretchable device attached to an equipment capable of measuring strain, is placed around the chest and the abdomen. Respiration rate has been shown to be positively related to self-reported anxiety.

Correlations among the various physiological measures of anxiety are generally found to be low. There are many factors that can account for this difference, including individual differences in the experience of anxiety and temporal factors. For example, Bloom and Trautt (1977) found that initially, participants were more anxious according to the finger pulse volume measure. However, according to heart rate, participants were more anxious later on. This provides support for the view that any measure of anxiety should be used along with other measures of anxiety. Various psychological, behavioural, and physiological processes are

involved in the experience of anxiety and there are individual differences.

## SUMMARY, CONCLUSIONS, AND DIRECTIONS FOR FUTURE RESEARCH

- 1 Most of the research has used self-report measures. Additional research can further investigate the reliability and validity of the various techniques in the assessment of anxiety.
- 2 Considerably more research has been conducted on the assessment of anxiety in the social evaluation situational domain (e.g. presentation situations, interaction situations) compared to other areas. This is especially the case with respect to the use of behavioural observation, cognitive measures, and physiological measures.
- 3 Future research can focus on the use of these techniques for the assessment of anxiety in situations other than social evaluation situations (i.e. physical danger situations, self-disclosure situations, separation situations, and ambiguous situations).
- 4 There are various techniques to assess state anxiety, the momentary experience of anxiety. Included among these are self-report instruments, behavioural observation methods, cognitive assessment techniques and physiological measures.
- 5 Trait anxiety, the predisposition to be anxious in different situations, is assessed through self-report instruments.
- 6 The reliability and validity of some techniques have been demonstrated to be higher than for other techniques.
- 7 There are individual differences in the qualitative experience of anxiety. It is therefore important to use diverse sets of assessment techniques that tap at the various facets of anxiety.
- 8 Self-report measures may be the most convenient method of anxiety assessment in terms of the time required for administration, the cost of administration, and data analyses. However, other factors (i.e. the validity of the assessment) are also important to consider.

## Acknowledgments

This chapter was supported, in part, by Grant No. 410-94-1473 from the Social Sciences and Humanities Research Council of Canada (SSHRC) to the first author and a SSHRC doctoral fellowship to the second author.

## References

- Beck, A.T., Epstein, N., Brown, G. & Steer, R.A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology*, 56, 893–897.
- Bloom, L.J. & Trautt, G.M. (1977). Finger pulse volume as a measure of anxiety: further evaluation. *Psychophysiology*, 14, 541–544.
- Cacioppo, J.T. & Petty, R.E. (1981). Social psychological procedures for cognitive response assessment: the thought-listing technique. In Merluzzi, T.V., Glass, C.R. & Genest, M. (Eds.), *Cognitive Assessment* (pp. 309–342). New York: Guilford.
- Clements, K. & Turpin, G. (1996). Physiological effects of public speaking assessed using a measure of palmar sweating. *Journal of Psychophysiology*, 10, 283–290.
- Endler, N.S. (1997). Stress, anxiety and coping: the multidimensional interaction model. *Canadian Psychology*, 38, 136–153.
- Endler, N.S., Edwards, J.M. & Vitelli, R. (1991). *Endler Multidimensional Anxiety Scales (EMAS): Manual*. Los Angeles, CA: Western Psychological Services.
- Endler, N.S. & Flett, G.L. (2001). *Endler Multidimensional Anxiety Scales – Social Anxiety Scales: Manual*. Los Angeles, CA: Western Psychological Services.
- Endler, N.S. & Kocovski, N.L. (in press). State and trait anxiety revisited. *Journal of Anxiety Disorders*.
- Glass, C.R. & Arnkoff, D.B. (1989). Behavioural assessment of social anxiety and social phobia. *Clinical Psychology Review*, 9, 75–90.
- Gray, J.A. & McNaughton, N. (1996). The neuropsychology of anxiety: Reprise. In Hope, D.A. (Ed.), *Nebraska Symposium on Motivation, 1995: Perspectives on Anxiety, Panic, and Fear. Current Theory and Research in Motivation*, Vol. 43 (pp. 61–134). Lincoln, NE: University of Nebraska Press.
- Kantor, L., Endler, N.S., Heslegrave R.J. & Kocovski, N.L. (2000). *Validating Self-Report Measures of State and Trait Anxiety with a Physiological Measure*. Manuscript submitted for publication.
- Leary, M.R. (1986). Affective and behavioural components of shyness: implications for theory, measurement, and research. In Jones, W.H., Cheek, J.M. & Briggs, S.R. (Eds.), *Perspectives on Shyness: Research and Treatment* (pp. 27–38). New York: Plenum.
- Lewis, A. (1970). The ambiguous word “anxiety”. *International Journal of Psychiatry*, 9, 62–79.
- Reiss, S., Peterson, R.A., Gursky, D.M. & McNally, R.J. (1986). Anxiety sensitivity, anxiety frequency, and the prediction of fearfulness. *Behaviour Research and Therapy*, 24, 1–8.

- Spielberger, C.D. (1983). *Manual for the State-Trait Anxiety Inventory (Form V)*. Palo Alto, CA: Consulting Psychologists Press.
- Taylor, J.A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology*, 48, 285–290.

Norman S. Endler and Nancy L. Kocovski

# A

---

## ANXIETY DISORDERS ASSESSMENT

### *Related Entries*

PERSONALITY (GENERAL), EMOTIONS, ANXIETY DISORDERS, TEST ANXIETY

## INTRODUCTION

### **The Multidimensional Nature of Anxiety**

Anxiety is one of the most common and universal emotions. This emotional reaction to the perception of threatening or dangerous stimuli occurs throughout an individual's lifetime. In fact, anxiety elicited by stimuli or situations such as animals, physical danger and separation is an early biological acquisition, whose function is to protect the child from potential dangers. In this sense, anxiety is undoubtedly of value in relation to the preservation of the human being.

The conceptualization of anxiety has varied considerably over recent decades. On the one hand, critics of the unidimensional view of anxiety have proposed a new multidimensional approach. From this perspective, anxiety is a combination of responses, including cognitive, physiological and behavioural (motor) reactions. These responses are provoked by identifiable cognitive-subjective, physiological or environmental stimuli. In spite of the lack of an accurate explanation of the contents of each system, and there being some discrepancies among authors on what might be understood by the responses of the cognitive system or, to a lesser extent, those of the physiological system (Cone & Hawkins, 1977; Fernández-Ballesteros, 1983), this classification of the different anxiety responses in three systems is widely accepted and used.

In addition, since the seminal works of Cattell or Spielberger in the 1960s, the differentiation between state and trait anxiety has become a classic one. State anxiety is conceptualized as a transitory emotional reaction to the individual's

perception of a threatening or dangerous situation, while trait anxiety is defined as a relative stable tendency to interpret situations as threatening or dangerous, and to react to them with anxiety. Recent works by Endler and his co-workers propose a multidimensional nature for trait anxiety, highlighting the existence of different facets (social evaluation, physical danger, etc.) closely related to specific situational areas.

With the aim of integrating the above-mentioned aspects, anxiety must be considered as an emotional response, or pattern of responses, that includes unpleasant cognitive aspects, physiological aspects characterized by high arousal of the Autonomous Nervous System, and inaccurate and less adaptive motor or behavioural reactions. The anxiety response may be provoked both by situational external and internal stimuli such as thoughts, ideas, images, etc., perceived by the individual as threatening or dangerous. Such anxiety-eliciting stimuli (external or internal) will be mainly determined by the subject's characteristics; thus, there are remarkable individual differences in relation to the tendency to manifest anxiety reactions in different situations (Miguel-Tobal, 1990).

### **ANXIETY AS DISORDER**

Up to now, we have considered anxiety as a normal emotional response of an individual to different situations or circumstances. However, when its frequency, intensity and duration are excessive, producing serious limitations in different facets of individuals' lives and reducing their ability to adapt to the environment, we must talk about pathological anxiety.



Anxiety is closely related to anxiety disorders, depression, disorders traditionally labelled as neurotic, many psychotic disorders, and a wide variety of psychophysiological problems such as cardiovascular disorders, peptic ulcers, headaches, premenstrual syndrome, asthma, skin disorders, and so on. It is also involved in sexual disorders, addictive behaviour and eating disorders; more recently, there are findings that relate anxiety to weakness of the immune system.

Due to the wide variety of problems in which this emotion plays an important role, anxiety must be considered a central aspect of psychopathology and health psychology. In fact, thousands of persons with anxiety problems seek attention in hospitals, health centres, etc., and this results in an important economic cost to public health services.

## Anxiety Disorders

Anxiety disorders constitute the most common psychopathology, followed by affective disorders and drugs and alcohol consumption. The life-prevalence rate accounts for 19.5% of females and 8% of males (Robins, Helzer & Weissman, 1984).

The classifications of anxiety disorders have varied over recent years. The most widely used are the ICD-10 (World Health Organization, 1992), the DSM-IV (American Psychiatric Association, 1994) and the DSM-IV-TR (American Psychiatric Association, 2000). The DSM-IV and DSM-IV-TR will be used as reference sources, and are shown in Table 1.

## Anxiety Disorders Assessment

Changes in the theoretical frameworks of anxiety research that occurred in the late 1960s have not been accurately reflected in assessment procedures which are instruments, especially for self-report measures, the most widely used. This has impeded the consolidation of a systematic research line focused on different aspects of anxiety in several anxiety disorders.

The works of Lacey (1967) and Lang (1968) proposed the multidimensional nature of anxiety responses and the existence of three relatively independent response systems (cognitive, physiological, and motor responses), while the interactive model (Endler, 1973) stressed the multidimensionality of trait anxiety (Endler & Magnusson, 1974,

Table 1. DSM-IV–DSM-IV-TR Classification

Codes	Anxiety Disorders
300.01	Panic disorder without agoraphobia
300.21	Panic disorder with agoraphobia
300.22	Agoraphobia without history of panic disorder
300.29	Specific phobia
300.23	Social phobia (Social Anxiety Disorder, added in DSM-IV-TR)
300.3	Obsessive-compulsive disorder
309.81	Post-traumatic stress disorder
308.3	Acute stress disorder
300.02	Generalized anxiety disorder (includes overanxious disorder of childhood in DSM-IV-TR)
293.84	Anxiety due to a medical condition
Variable	Substance-induced anxiety disorder
300.00	Anxiety disorder not otherwise specified

1976). Finally, the discovery of individual differences in relation to the tendency to experience anxiety in some situations, but not in others, led to theoretical advances that have not yet been sufficiently applied in research on anxiety disorders.

With the aim of including all of these theoretical advances in an assessment instrument, we developed the *Inventory of Situations and Responses of Anxiety* (ISRA, Miguel-Tobal & Cano Vindel, 1986, 1988 & 1994). The ISRA is a self-report instrument for a multidimensional and interactive assessment of anxiety that permits the evaluation of the three response systems (cognitive, physiological and motor responses), trait anxiety, and four situational areas or specific traits (test anxiety, interpersonal anxiety, phobic anxiety and daily life anxiety).

Several studies have explored differential anxiety characteristics, in both anxiety disorders and psychophysiological disorders, through the ISRA. Such studies indicate that there are characteristic profiles in different pathologies that can be relevant in both the research and clinical practice contexts (see Miguel-Tobal & Cano Vindel, 1995).

## INSTRUMENTS AND PROCEDURES

A large number of procedures and instruments have been used for the assessment of anxiety,

including self-reports, physiological procedures and behavioural methods. More information on this issue can be found in Endler and Kocovski's entry 'Anxiety Assessment' in this same volume. Here we shall focus especially on the instruments developed for the assessment of different anxiety disorders. It should be noted that procedures for the assessment of general anxiety are also commonly used in clinical practice.

### Broad Screening

Several structured interviews have been used in order to determine the onset of an anxiety disorder or to make a more accurate diagnosis. Two good examples are the *Anxiety Disorder Interview Schedule – Revised* (Di Nardo et al., 1985), and the *Structured Clinical Interview for DSM-IV Axis I disorders* (Spitzer, Gibbon & Williams, 1996).

With regard to specific disorders, some widely used instruments and procedures are:

#### *Panic Disorder Assessment*

The most widely used self-report instrument for the assessment of panic attacks is the *Panic Attack Questionnaire* (PAQ, Norton, 1988).

#### *Agoraphobia Assessment*

In the assessment of agoraphobia, both self-reports and behavioural measures have been used. Among self-reports, the *Agoraphobic Cognitions Questionnaire* (ACQ), along with its companion measure, the *Body Sensations Questionnaire* (BSQ), were devised to assess 'fear of fear' (Chambless, Caputo, Bright & Gallagher, 1984). Among behavioural measures, there are two kinds of devices: one type that measures avoidance behaviours, an example of which is the *Individualized Behavioural Avoidance Test* (IBAT, Agras, Leitenberg & Barlow, 1968), and another type for measuring the time and distance walked away from a 'safe' place as a cue for the intensity of agoraphobic reactions (see Emmelkamp, 1982). It should be noted that assessment instruments designed for phobia, social phobia, and panic attacks are also used in the evaluation of agoraphobia.

#### *Specific Phobia Assessment*

The most frequently used instruments are self-reports, such as the *Fear Survey Schedule I* (Lang & Lazovik, 1963) and *Fear Survey Schedule III* (Wolpe & Lang, 1964), for measuring the type and intensity of irrational fears and fear-eliciting stimuli. Also used are behavioural avoidance measures, such as the *Behavioural Avoidance Test* (Lang & Lazovik, 1963) and the *Behavioural Avoidance Slide Test* (Burchardt & Levis, 1977). It should be noted that some of these instruments are also used for the assessment of social phobia and agoraphobia.

#### *Social Phobia Assessment*

The *Social Avoidance and Distress Scale* (SADS), the *Fear of Negative Evaluation Scale* (FNE, Watson & Friend, 1969), the *Suinn Test Anxiety Behaviour Scale* (STABS, Suinn, 1969) and the *Social Reaction Inventory Revised* (SRI-R, Curran, Corriveau, Monti & Hagerman, 1980) are used for assessing social skills, while the *Social Phobia and Anxiety Inventory* (SPAI, Turner, Beidel, Dancu & Stanley, 1989) is also employed. Among behavioural measures, the *Social Interaction Test* (SIT, Trower, Bryant & Argyle, 1978) is designed for measuring social skills in a test anxiety-provoking situation by means of role-play procedures.

#### *Obsessive-Compulsive Disorder Assessment*

The most important self-report measures used are the *Leyton Obsessional Inventory* (LOI, Cooper, 1970), the *Compulsive Activity Checklist* (CAC, Philpott, 1975) and the *Maudsley Obsessional-Compulsive Inventory* (MOCI, Hodgson & Rachman, 1977).

#### *Post-Traumatic Stress Disorder Assessment*

There are several methods for the assessment of PTSD disorder, including clinical interviews, self-report instruments and psychophysiological measures. For the purpose of this entry we consider general-oriented instruments rather than special populations-oriented ones (combat

survivors, rape victims, etc.), except for psychophysiological measures. Two good examples of clinical interviews are the *Clinical-Administered PTSD Scale* (CAPS-1, Blake, Weathers, Nagy, Kaloupek, Klauminzer, Charney & Keane, 1990), and the *PTSD Symptom Scale Interview* (PSS-I, Foa, Riggs, Dancu & Rothbaum, 1993). Two other good examples of self-report instruments are the *Revised Impact of Events Scale* (RIES, Horowitz, Wilner & Alvarez, 1979), and the *PTSD Diagnostic Scale* (PDS, Foa, 1995). Finally, data from laboratory studies provide evidence that psychophysiological measurement is a valuable tool in the assessment of PTSD. Studies with combat populations reveal that *cardiovascular measures* (heart rate and blood pressure) have generally shown good specificity and sensitivity in PTSD classification (see Lating & Everly, 1995; Miguel-Tobal, González Ordi & Léppez Ortega, 2000).

### *Generalized Anxiety Disorder (GAD) Assessment*

Given the lack of specificity of GAD general anxiety assessment instruments, including the *State-Trait Anxiety Inventory* (STAI, Spielberger, Gorsuch & Lushene, 1970), the *Beck Anxiety Inventory* (BAI, Beck, Epstein, Brown, & Steer, 1988), the *Anxiety Sensitivity Index* (ASI, Reiss, Peterson, Gursky & McNally, 1986), the *Endler Multidimensional Anxiety Scales* (EMAS, Endler, Edwards & Vitelli, 1991) and, in Spain, the *Inventory of Situations and Responses of Anxiety* (ISRA, Miguel-Tobal & Cano Vindel, 1986, 1988, 1994) have been used for its evaluation.

As can be seen, there are very few references to physiological measures in this review since, though commonly used in clinical research, they have not generally shown enough specificity to discriminate between different anxiety disorders, except, as mentioned earlier, in the case of PTSD.

Finally, we should stress the appropriateness of using multiple instruments that allow the assessment of general anxiety on the one hand and the evaluation of a specific disorder or disorders on the other. Clinical practice reveals that it is hard to find a pure disorder, since, as Wittchen (1987) points out, the comorbidity rate for anxiety disorders is 68%: in other words, two

out of every three patients also present another anxiety disorder.

### SUMMARY, CONCLUDING REMARKS, AND DIRECTIONS FOR FUTURE RESEARCH

Anxiety disorder assessment has mainly been carried out using self-reports, and to a lesser extent behavioural measures. Physiological measures do not provide sufficient specificity to delimit or evaluate specific disorders; however, there is a promising line of research in relation to PTSD.

In addition to this lack of specificity with regard to anxiety disorders, due to the overlapping of their symptoms, it is also important to consider the problem of their high comorbidity (68% for anxiety disorders and 50% for depression). Taking these aspects into account, it is necessary to carry out a wide-spectrum assessment that includes general anxiety measures, specific disorder measures and measures of depression.

Theoretical advances in the study of anxiety and research on measurement procedures have fostered the multisystem-multimethod assessment, but such advances have been weakly reflected in anxiety disorder assessment research, and have had even less impact on clinical practice. This is one of the challenges for the future, which it is to be hoped will see the development of new multidimensional instruments through the integration of data derived from self-reports, physiological records and behavioural measures.

### References

- Agras, W.S., Leitenberg, H. & Barlow, D.H. (1968). Social reinforcement in the modification of agoraphobia. *Archives of General Psychiatry*, 19, 423–427.
- American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (4th ed. ). Washington, DC: APA.
- American Psychiatric Association (2000). *Desk Reference to the Diagnostic Criteria from DSM-IV-TR*. Washington, DC: APA.
- Beck, A.T., Epstein, N., Brown, G. & Steer, R.A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology*, 56, 893–897.

- Blake, D.D., Weathers, F.W., Nagy, L.M., Kaloupek, D.G., Klauminzer, G., Charney, D. & Keane, T.M. (1990). A clinical rating scale for assessing current and lifetime PTSD: the CAPS-1. *Behaviour Therapist*, 18, 187-188.
- Burchardt, C.J. & Levis, D.J. (1977). The utility of presenting slides of a phobic stimulus in the context of a behavioural avoidance procedure. *Behaviour Therapy*, 8, 340-346.
- Chambless, D.L., Caputo, G.C., Bright, P. & Gallagher, R. (1984). Assessment of fear of fear in agoraphobics: the Body Sensations Questionnaire and the Agoraphobic Cognitions Questionnaire. *Journal of Consulting and Clinical Psychology*, 52, 1090-1097.
- Cone, J.D. & Hawkins, R.P. (1977). *Behavioural Assessment: New Directions in Clinical Psychology*. New York: Brunner-Mazel.
- Cooper, J. (1970). The Leyton Obsessional Inventory. *Psychological Medicine*, 1, 48-64.
- Curran, J.P., Corriveau, D.P., Monti, P.M. & Hagerman, S.B. (1980). Social skill and social anxiety. *Behaviour Modification*, 4, 493-512.
- Di Nardo, P.A., Barlow, D.H., Cerny, J.A., Vermilyea, B.B., Vermilyea, J.A., Himadi, W.G. & Wadell, M.T. (1985). *Anxiety Disorders Interview Schedule-Revised (ADIS-R)*. Albany, NY: Center for Stress and Anxiety Disorders.
- Emmelkamp, P.M.G. (1982). *Phobic and Obsessive-Compulsive Disorders: Theory, Research and Practice*. New York: Plenum Press.
- Endler, N.S. (1973). The person versus the situation a pseudo issue? a response to others. *Journal of Personality*, 41, 287-303.
- Endler, N.S., Edwards, J.M. & Vitelli, R. (1991). *Endler Multidimensional Anxiety Scales (EMAS): Manual*. Los Angeles, CA: Western Psychological Services.
- Endler, N.S. & Magnusson, D. (1974). Interactionism, trait psychology, psychodynamics, and situationism. Report from the Psychological Laboratories. University of Stockholm, No 418.
- Endler, N.S. & Magnusson, D. (Eds.) (1976). *Interactional Psychology and Personality*. Washington, DC: Hemisphere Publishing Co.
- Fernández-Ballesteros, R. (1983). *Psicodiagn Óstico*. Madrid: UNED.
- Foa, E.B. (1995). *PSD (Posttraumatic Stress Diagnostic Scale). Manual*. Minneapolis: National Computer System.
- Foa, E.B., Riggs, D.S., Dancu, C.V. & Rothbaum, B.O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of Trauma Stress*, 6, 459-473.
- Hodgson, R.J. & Rachman, S. (1977). Obsessional-compulsive complaints. *Behaviour Research and Therapy*, 15, 389-395.
- Horowitz, M.J., Wilner, N. & Alvarez, W. (1979). Impact of Event Scale: A measure of subjective distress. *Psychosomatic Medicine*, 41, 207-218.
- Lacey, J.I. (1967). Somatic responses patterning and stress: some revisions of the activation theory. In Appley, M.H. & Trumbull, R. (Eds.), *Psychological stress: Issues in Research* (pp. 14-42). New York: Appleton-Century-Crofts.
- Lang, P.J. (1968). Fear reduction and fear behaviour: problems in treating a construct. In Shilen, J.M. (Ed.), *Research in Psychotherapy*; Vol. III (pp. 90-103). Washington, DC: American Psychological Association.
- Lang, P.J. & Lazovic, A.D. (1963). Experimental desensitization of a phobia. *Journal of Abnormal and Social Psychology*, 66, 519-525.
- Lating, J.M. & Everly, G.S. (1995). Psychophysiological assessment of PTSD. In Everly, G.S. & Lating, J.M. (Eds.), *Psychotraumatology: Key Papers and Core Concepts in Post-Traumatic Stress* (pp. 129-145). New York: Plenum Press.
- Miguel-Tobal, J.J. (1990). La ansiedad. In Mayor J. & Pinillos J.L. (Eds.), *Tratado de Psicología General. Vol. 8: Motivación y Emoción* (pp. 309-344). Madrid: Alhambra.
- Miguel-Tobal, J.J. & Cano Vindel, A.R. (1986). *Inventario de Situaciones y Respuestas de Ansiedad (Inventory of Situations and Responses of Anxiety)*. (1988 & 1994, 2nd and 3rd revisions, respectively). Madrid: TEA Ediciones
- Miguel-Tobal, J.J. & Cano Vindel, A. (1995). Perfiles diferenciales de los trastornos de ansiedad. *Ansiedad y Estrés*, 1, 37-60.
- Miguel-Tobal, J.J., González Ordi, H. & López Ortega, E. (2000). Estrés postraumático: hacia una integración de aspectos psicológicos y neurobiológicos. *Ansiedad y Estrés*, 6, 255-280.
- Norton, G.R. (1988). Panic Attack Questionnaire. In Hersen, M. & Bellack, A.S. (Eds.), *Dictionary of Behavioural Assessment Techniques* (pp. 332-334). New York: Pergamon Press.
- Philpott, R. (1975). Recent advances in the behavioural measurement of obsessional illness: Difficulties common to these and other instruments. *Scottish Medical Journal*, 20, 33-40.
- Reiss, S., Peterson, R.A., Gursky, D.M. & McNally, R.J. (1986). Anxiety sensitivity, anxiety frequency, and the prediction of fearfulness. *Behaviour Research and Therapy*, 24, 1-8.
- Robins, L.N., Helzer, J.E. & Weissman, M.M. (1984). Lifetime prevalence of specific psychiatric disorders in three sites. *Archives of General Psychiatry*, 41, 949-958.
- Spielberger, C.D., Gorsuch, R.L. & Lushene, R.E. (1970). *STAI. Manual for the State-Trait Anxiety Inventory (Self-Evaluation Questionnaire)*. Palo Alto, CA: Consulting Psychologists Press.
- Spitzer, R.L., Gibbon, M. & Williams, J.B.W. (1996). *Structured Clinical Interview for DSM-IV Axis I Disorders*. New York: New York State Psychiatric Institute, Biometrics Research Department.
- Suinn, R. (1969). The STABS, a measure of test anxiety for behaviour therapy: normative data. *Behaviour Research and Therapy*, 7, 335-339.
- Trower, P., Bryant, B. & Argyle, M. (1978). *Social Skills and Mental Health*. Pittsburgh: University of Pittsburgh Press.

- Turner, S.M., Beidel, D.C., Dancu, C.V. & Stanley, M.A. (1989). An empirically derived inventory to measure social fears and anxiety: The Social Phobia and Anxiety Inventory. *Psychological Assessment*, 1, 35–40.
- Watson, D. & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, 33, 448–457.
- Wittchen, H.U. (1987). Epidemiology of panic attacks and panic disorder. In Hand, I. & Wittchen, H.U. (Eds.), *Panic and phobias (1). Empirical Evidence of Theoretical Models and Long-Term Effects of Behavioural Treatments*. New York: Springer-Verlag.
- Wolpe, J. & Lang, P.J. (1964). A fear survey schedule for use in behaviour therapy. *Behaviour Research and Therapy*, 2, 27–30.
- World Health Organization (1992). *International Classification of Diseases and Related Health Problems. ICD-10*. Geneva: WHO.

Juan José Miguel-Tobal and  
Héctor González-Ordi

### Related Entries

APPLIED FIELDS: CLINICAL, ANXIETY, EMOTIONS, TEST ANXIETY

# A APPLIED BEHAVIOURAL ANALYSIS

## INTRODUCTION

Applied behaviour analysis is a science in which procedures derived from the principles of behaviour are systematically applied to improve socially meaningful behaviour that could be rigorously defined and objectively detected and measured (Cooper et al., 1987). As pointed out by Moore (1999), behaviour analysis has developed three components, as well as a philosophy of science: (1) the *experimental analysis of behaviour*, the basic science of behaviour, (2) *applied behaviour analysis*, the systematic application of behavioural technology, and (3) the *conceptual analysis of behaviour*, the philosophical analysis of the subject matter of behaviour analysis. The philosophy of science that guides behaviour analysis is called *radical behaviourism*. Even though, the link between the experimental and applied component of behaviour analysis is not as united as it should be, bridges are being built between basic and applied work, such as the work being conducted in the areas of establishing fluency and building momentum (Mace, 1996). The impact of bridge studies has been especially pronounced in functional analysis methodologies on aberrant behaviour (Wacker, 2000). This article will focus on important aspects of functional assessment.

## CHARACTERISTICS AND AREAS OF INTEREST

Baer, Wolf, and Risley (1968) list seven defining characteristics of applied behaviour analysis: behaviour or stimuli studied are selected because of their significance to society rather than their importance to theory (*applied*). The behaviour chosen must be the behaviour in need of improvement and it must be measurable (*behavioural*). It requires a demonstration of the events that can be responsible for the occurrence or non-occurrence of that behaviour (*analytic*). The interventions must be completely identified and described (*technological*). The procedure for behaviour change is described in terms of the relevant principles from which they are derived (*conceptual systems*). The behavioural techniques must produce significant effects for practical value (*effective*). The behavioural change must be stable over time, appear consistently across situations, or spread to untrained responses (*generality*).

The writings of B. F. Skinner have inspired behaviour analysts to develop basic concepts of reciprocal behaviour–environment interactions. Over fifty years of research and application have shown the usefulness of these basic concepts in understanding many forms of behaviour, as well as in guiding effective

behaviour-change strategies. The knowledge of *stimulus control* (when the presentation of a stimulus changes some measures of behaviour) and *reinforcement* (the process by which the frequency of an operant (class of responses) is increased) has been useful in the analysis and treatment of human behaviour problems, as well as creating novel behaviour since the inception of applied behaviour analysis. Applied behaviour analysis has played a prominent role in the treatment of individuals with autism and/or developmental disabilities. Though, the areas of interest have been expanding, e.g. school settings, treatment of habit disorders, paediatrics, troubled adolescent runaways, brain-injury rehabilitation, behavioural psychotherapy, organizational management, performance analysis, consultation, sport psychology, college teaching, and behavioural medicine (e.g. Austin and Carr, 2000).

## ASSESSMENT

The role of assessment in applied behaviour analysis has been described as the process of identifying a problem and identifying how to alter it for the better. Furthermore, it involves selecting and defining the behaviour (target behaviour) to be changed. Two questions have been essential in behavioural assessment: “(a) What types of assessment methodologies provide reliable and valid data about behavioural function, and how can they be adapted for use in a particular situation? and (b) How might the results of such assessments improve the design and selection of treatment procedures?” (Neef & Iwata, 1994: 211). As we shall examine further, behaviour is assumed to be a function of current environmental conditions – antecedent and consequent stimuli, and it is predicted to be stable as long as the specific environmental conditions remain stable. On the contrary, traditional approaches or non-behavioural therapies assume that the behaviour is a function of enduring, underlying mental states or personal variables. One premise is that the client’s verbal behaviour (what people talk about, what they do and why they do it) is considered important because it is believed to be reflective of a person’s inner state and the mental processes that govern a person’s behaviour (Cooper et al., 1987). This is

quite different from a behaviour analytic view where a distinction is made between what people say they do and what they do (Skinner, 1953), and the focus is on behaviour for its own sake.

## Function versus Structure

Behaviour could be classified either structurally or functionally. When we talk about a structural approach, behaviour is classified or analysed in terms of its form. For example within developmental psychology, the structural approach is a prominent approach in which researchers investigate what children do at specific stages of development, e.g. the behaviour is studied to draw inferences about cognitive abilities and so-called hypothetical structures, as object permanence or Piagetian schemes. In behaviour analysis, the topography or structure of a response is determined by the contingencies of this behaviour. Instead of inferring such cognitive abilities, the researchers consider the history of reinforcement to be responsible for the child’s capability (Pierce & Epling, 1999). Structural approaches to assessment are exemplified by diagnostic, personality and psychodynamic approaches to human behaviour, while functional explanations focus on the relationships between what happens to the organism (i.e. stimuli) and the behaviour of the organism (responses) (Sturme, 1996). The controversy between functional and structural approach is quite similar to debate in biology on the separation of physiology and anatomy, and also to Skinner’s treatment of verbal behaviour (function; without regard to modality (vocal, gestures etc.), the field of verbal behaviour is concerned with the behaviour of individuals and the functional units of their verbal behaviour function) versus language (structure; the consistencies of vocabulary and grammar) (Catania, 1998).

## Functional Assessment

Early in the development of behaviour analysis, Skinner (1938) argued that behaviour did not take place in a vacuum and a response must have a function. Empirical demonstrations of “cause-effect relationships” between environment and behaviour have been rendered possible by functional analysis (Skinner, 1953). Since then comprehensive methods to systematically

assess particular functions of different types of behaviour have been developed, and functional assessment is one of the most intense research areas in our field (see for example Iwata et al., 2000; Repp & Horner, 1999).

Functional assessment is an umbrella term and encompasses: (1) indirect assessments, which are characterized by interviews and questionnaires and behavioural functions. They are based on subjective verbal reports in absence of direct observation. Two recognized indirect methods are the *Motivation Assessment Scale* (Durand & Crimmins, 1988) and the *Motivation Analysis Rating Scale* (Wieseler et al., 1985), (2) descriptive assessments involve no manipulation of relevant variables and are based on direct observation, e.g. the antecedent-behaviour-consequence assessment (ABC) or scatter plot assessment, (3) functional experimental analyses or analogue functional assessment involve manipulation of suspected maintaining variables using experimental methodology to demonstrate control over responding (Desrochers et al., 1997). The first two approaches are approximations compared to the third because they do not elucidate functional relationships, and both are characteristically non-experimental. Furthermore, the functional experimental analysis is most effective in identifying the function of problem behaviour (Carr et al., 1999).

### Experimental functional analysis or analogue functional assessment

Since the prominent publication by Iwata et al. (1982) there has been a remarkable increase of publications concerning experimental functional analysis (See *Journal of Applied Behaviour Analysis*). Experimental functional analysis represents a simulation of the natural environment and will be the primary tool for demonstrating causal relationships (Carr et al., 1999). Experimental functional analysis methodologies can be used to identify: (1) antecedent conditions (setting events, establishing operations and/or discriminative stimuli) under which behaviour occurs, and these conditions may then be altered so that problem behaviours are less likely, (2) reinforcement contingencies that must be changed, (3) whether the same reinforcer that currently maintains the

behaviour problem may be used in establishing and strengthening alternative behaviours, and (4) those reinforcers and/or treatment components that are relevant (Iwata et al., 2000).

Results from the research on functional analysis methodologies have shown that functional analyses are effective in identifying environmental determinants of self-injurious behaviour (SIB), and subsequently, in guiding the process of treatment selection (Iwata et al., 1994). Furthermore, results have shown that the growing use of functional assessment based interventions have increased the number of studies using non-aversive procedures (Carr et al., 2000).

### Recording techniques

In applied behaviour analysis it is important to demonstrate that a particular intervention has been responsible for a particular behaviour change. Therefore, measurement is very important with respect to designing successful interventions and evaluating treatment changes. Automatic recording, permanent products, and direct observational recording are procedures used for measuring and recording behaviour. Direct observational recording include frequency or event, duration, or latency recording, and the recording could either be continuous, time sampling or interval (Cooper et al., 1987). Objectivity, clarity and completeness have been set forth as three criteria of an adequate response definition (Kazdin, 1982).

### Experimental designs

In experimental functional analyses various experimental designs have been used to rule out the possibility that changes in extraneous variable(s) other than in the independent variable could be responsible for the change in dependent variable, e.g. eliminating rival explanations. Thus, these experimental designs have been used to study the functional relationships between environmental changes and changes in target behaviour. Typical experimental design N=1 design (within-subject manipulation, single-case research design) have been used in applied behaviour analysis, and the designs have been categorized as ABAB designs, multiple baseline designs, multiple treatment designs and changing criterion designs (Kazdin, 1982). The multi-element design (multiple treatment designs) has

typically been used in experimental functional analysis (e.g. Iwata et al., 1982).

In single-case research, replication, either direct or systematic, is crucial for evaluating generality of intervention effects across subjects. The term direct replication has been used when the same procedures have been used across a number of different subjects, while systematic replication indicate that features (e.g. types of subjects, intervention, target behaviour) of the original experiment vary. By replicating in this way, knowledge will be accumulated, and behaviourists will be pyramid builders.

## CONCLUSION AND FUTURE PERSPECTIVES

Different aspects regarding behavioural assessment as indirect assessment, descriptive assessment and experimental functional analysis have been discussed. Extension and refinement of behavioural assessment and functional analysis technologies will, hopefully, provide for even more effective methods in establishing behaviour and treating maladaptive behaviour. In addition, the advancement of computer technology allows for more simplified assessment techniques. Until now functional assessment technologies have primarily focused on non-compliance and self-injurious and aggressive behaviour in persons with disabilities and autism, but advancements in these procedures will include their applications on other types of behaviour and a larger diversity of problem behaviour in populations other than persons with autism and disabilities.

## References

- Austin, J. & Carr, J.E. (Eds.) (2000). *Handbook of Applied Behaviour Analysis*. Reno, Nevada: Context Press.
- Baer, D.M., Wolf, M.M. & Risley, T.R. (1968). Some current dimensions of applied behaviour analysis. *Journal of Applied Behaviour Analysis*, 1, 91–97.
- Carr, E.G., Langdon, N.A. & Yarbrough, S.C. (1999). Hypothesis-based intervention for severe problem behaviour. In Repp, A.C. & Horner, R.H. (Eds.), *Functional Analysis of Problem Behaviour* (pp. 9–31). Belmont, CA: Wadsworth Publishing Company.
- Carr, J.E., Coriarty, S. & Dozier, C.L. (2000). Current issues in the function-based treatment of aberrant behaviour in individuals with developmental disabilities. In Austin, J. & Carr, J.E. (Eds.), *Handbook of Applied Behaviour Analysis* (pp. 91–112). Reno, Nevada: Context Press.
- Catania, A.C. (1998). *Learning* (4th ed.). Englewood Cliffs, New Jersey: Prentice-Hall.
- Cooper, J.O., Heron, T.E. & Heward, W.L. (1987). *Applied Behaviour Analysis*. Merrill Publications: Columbus.
- Desrochers, M.N., Hile, M.G. & Williams-Moseley, T.L. (1997). Survey of functional assessment procedures used with individuals who display mental retardation and severe problem behaviours. *American Journal on Mental Retardation*, 5, 535–546.
- Durand, M. & Crimmins, D.B. (1988). Identifying the variables maintaining self-injurious behaviour in a psychotic child. *Journal of Autism and Developmental Disorders*, 18, 99–117.
- Iwata, B.A., Dorsey, M.F., Slifer, K.J., Bauman, K.E. & Richman, G.S. (1982). Toward a functional analysis of self-injury. *Analysis and Intervention in Developmental Disabilities*, 2, 3–20.
- Iwata, B.A., Kahng, S.W., Wallace, M.D. & Lindberg, J.S. (2000). The functional analysis model of behavioural assessment. In Austin, J. & Carr, J.E. (Eds.), *Handbook of Applied Behaviour Analysis* (pp. 61–89). Reno, Nevada: Context Press.
- Iwata, B.A., Pace, G.M., Dorsey, M.F., Zarcone, J.R., Vollmer, T.R., Smith, R.G., Rodgers, T.A., Lerman, D.C., Shore, B.A., Mazelski, J.L., Goh, H.-L., Cowdery, G.E., Kalsner, M.J., McCosh, K.C. & Willis, K.D. (1994). The functions of self-injurious behaviour: an experimental-epidemiological analysis. *Journal of Applied Behaviour Analysis*, 27, 215–240.
- Kazdin, A.E. (1982). *Single-Case Research Designs*. New York: Oxford University Press.
- Mace, F.C. (1996). In pursuit of general behavioural relations. *Journal of Applied Behaviour Analysis*, 29, 557–563.
- Moore, J. (1999). The basic principles of behaviourism. In Thyer, B.A. (Ed.), *The Philosophical Legacy of Behaviourism* (pp. 41–68). Dordrecht: Kluwer Academic Publishers.
- Neef, N.A. & Iwata, B.A. (1994). Current research on functional analysis methodologies: an introduction. *Journal of Applied Behaviour Analysis*, 27, 211–214.
- Pierce, W.D. & Epling, W.F. (1999). *Behaviour Analysis and Learning* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall, Inc.
- Repp, A.C. & Horner, R.H. (Eds.) (1999). *Functional Analysis of Problem Behaviour*. Belmont, CA: Wadsworth Publishing Company.
- Skinner, B.F. (1938). *The Behaviour of Organisms*. Acton, Massachusetts: Copley Publishing Group.
- Skinner, B.F. (1953). *Science and Human Behaviour*. New York: Free Press.
- Sturmey, P. (1996). *Functional Analysis in Clinical Psychology*. Baffins Lane, Chichester, UK: John Wiley & Sons.
- Wacker, D.P. (2000). Building a bridge between research in experimental and applied behaviour analysis. In



Leslie, J.C. & Blackman, D.E. (Eds.), *Experimental and Applied Analysis of Human Behaviour* (pp. 205–234). Reno, Nevada: Context Press.

Wieseler, N.A., Hanzel, T.E., Chamberlain, T.P. & Thompson, T. (1985). Functional taxonomy of stereotypic and self-injurious behaviour. *Mental Retardation*, 23, 230–234.

Erik Arntzen

# A

---

## APPLIED FIELDS: CLINICAL

### Related Entries

BEHAVIOURAL TECHNIQUES, OBSERVATIONAL TECHNIQUES (GENERAL), OBSERVATIONAL TECHNIQUES IN CLINICAL SETTING, THEORETICAL PERSPECTIVE: BEHAVIOURAL

## INTRODUCTION

Psychological assessment is utilized in clinical psychology primarily for purposes of differential diagnosis, treatment planning, and outcome evaluation. Differential diagnosis involves drawing on assessment information to describe an individual's psychological characteristics and adaptive strengths and weaknesses. These descriptions provide a basis for determining (a) what type of disorder an individual may have, (b) the severity and chronicity of this disorder and the circumstances in which it is likely to be manifest, and (c) the kinds of treatment that are likely to provide the individual relief from this disorder. With respect to further treatment planning, adequate assessment information helps to guide treatment strategies and anticipate possible obstacles to progress in therapy. As for outcome evaluation, pre-treatment assessments establish an objective baseline against which treatment progress can be monitored in subsequent evaluations, and by which the eventual benefits of the treatment can be judged at its conclusion. These clinical contributions of psychological assessment can be implemented during each of four sequential phases in delivering psychological treatment: deciding on therapy, planning therapy, conducting therapy, and evaluating therapy.

## DECIDING ON THERAPY

The first step in the clinical utilization of assessment information consists of deciding

whether a patient needs treatment and is likely to benefit from it. Accurate differential diagnosis identifies pathological conditions (e.g. depression, paranoia) and maladaptive characteristics (e.g. passivity, low self-esteem) for which treatment is usually indicated, and adequate psychological evaluation helps to distinguish such conditions and characteristics from normal range functioning that does not call for professional mental health intervention. Assessment methods also provide valuable information concerning two factors known to predict whether people are likely to become involved in and profit from psychotherapy: their motivation for treatment and their accessibility to being treated.

Motivation for treatment usually corresponds to the amount of subjectively felt distress that people are experiencing. Accessibility to psychological treatment typically depends on how willing people are to examine themselves, to express their thoughts and feelings openly, and to make changes in their customary beliefs and preferred ways of conducting their lives. Information derived from appropriate assessment procedures can provide clinicians with objective indices of each of these variables, and these assessment data can in turn be used as a basis for determining whether to recommend and proceed with some form of treatment.

## PLANNING THERAPY

Planning therapy for patients who need and want to receive psychological treatment involves

(a) deciding on the appropriate setting in which to deliver the treatment, (b) estimating the duration of the treatment, and (c) selecting the particular type of treatment to be given. With respect to deciding on the treatment setting, assessment data provide reliable information concerning the severity of a patient's disturbance, the patient's ability to distinguish reality from fantasy, and his or her likelihood of becoming suicidal or dangerous to others, all of which bear on whether the person requires residential care or can be treated safely and adequately as an outpatient. The more severely disturbed people are, the farther out of touch with reality they are, and the greater their risk potential for violence, the more advisable it becomes to care for them in a protected environment.

Regarding treatment duration, clinical experience and research findings consistently indicate that mild and acute problems of recent onset can usually be treated successfully in a shorter period of time than severe and chronic problems of long-standing duration. A variety of psychodiagnostic measures provide clues to the chronicity as well as the severity of symptomatic and characterological mental and emotional problems, and pretreatment data obtained with these measures can accordingly help clinicians formulate some expectation of how long a treatment is likely to last. Having available such assessment-based information on expected duration in turn assists clinicians in presenting treatment recommendations to prospective patients.

As for treatment selection, people who are relatively psychologically minded, self-aware, and interested in gaining fuller self-understanding are relatively likely to respond positively to an uncovering, insight-oriented, and conflict-focused treatment approach. Patients whose preference is to feel better without having to examine themselves closely, on the other hand, are more likely to become actively engaged in supportive and symptom-focused approaches to treatment than in exploratory psychotherapy. Psychologically minded people are inclined to feel dissatisfied with supportive treatment, because it does not get at the root of their problems, whereas relief-minded people tend to feel uncomfortable in uncovering treatment, because it makes unwelcome demands on them. Additionally, there is reason to believe that some

kinds of conditions and difficulties, especially in people who are problem-oriented, respond relatively well to cognitive-behavioural forms of treatment, whereas other kinds of disorders and maladaptive tendencies, especially in people who are interpersonally-oriented, respond better to psychodynamic-interpersonal than cognitive-behavioural therapy.

Psychological mindedness and preferences for problem-oriented or interpersonally-oriented approaches to life situations are among a vast array of personality characteristics that can be measured with assessment methods. Accordingly, adequately conceived pretherapy psychological assessment can facilitate treatment planning by differentiating among psychological states and orientations of the individual that have known implications for successful response to particular treatment approaches.

## CONDUCTING THERAPY

Psychological assessment can play a key role in conducting therapy by helping to identify in advance: (a) treatment targets on which the therapy should be focused and (b) possible obstacles to progress towards these treatment goals. Appropriately collected assessment data, and particularly the results of a multimethod test battery, typically contain many normal range findings and often some indications as well of notably good personality strengths and especially admirable personal qualities. At the same time, especially in people who are being evaluated for symptoms or difficulties that have led them to seek professional help, test data are likely to reveal specific adaptive shortcomings and coping limitations. One person may show a penchant for circumstantial reasoning and poor judgement; another person may give evidence of poor social skills and interpersonal withdrawal; a third may exhibit considerable emotional inhibition with restricted capacity to express feelings and feel comfortable in emotionally charged situations. In short, any assessment findings that fall outside of an established normal range and are known to indicate specific types of cognitive dysfunction, affective distress, coping deficit, personal dissatisfaction, or interpersonal inadequacy in turn assist therapists and

their patients in deciding on the objectives of their work together and directing their efforts accordingly.

Some psychological characteristics of patients that constitute targets in their treatment may also pose obstacles to their becoming effectively engaged in therapy and making progress toward their goals. For example, people who are set in their ways and characteristically rigid and inflexible in their views often have difficulty reframing their perspectives or modifying their behaviour in response even to well-conceived and appropriately implemented treatment interventions. People who are interpersonally aversive or withdrawn may be slow or reluctant to form the kind of working alliance with their therapist that facilitates progress in most forms of therapy. People who are relatively satisfied with themselves and not experiencing much subjectively felt distress may have little tolerance for the demands of becoming seriously engaged in a course of psychological treatment. Characteristics of these kinds do not preclude effective psychotherapy, but they can result in slow progress, and they may cause patients and therapists to become discouraged and terminate prematurely a treatment that does not appear to be going well. Pretreatment assessment data serve to alert therapists in advance to possible treatment obstacles, which can help them understand and be patient with initially slow progress and also guide them in dealing directly with these obstacles, as by concentrating in the early phases of therapy on encouraging flexibility and open-mindedness, building a comfortable and trusting treatment relationships, or generating some motivation for the patient's involvement in the therapy.

## EVALUATING THERAPY

Psychological assessment provides valuable data for monitoring the progress of therapy and measuring its eventual benefit. For this potential benefit of assessment to be realized, it is vital for assessment data to be collected from patients prior to their beginning treatment. In addition to helping to identify treatment targets and the long-term objectives of therapy, pretreatment data provide an objective baseline for

comparison with the results of subsequent assessments. Periodic re-evaluations can then shed light on whether the treatment is making a difference, how close it has come to meeting its aims, in what way the focus of continued treatment should be adjusted, and whether a termination point has been reached.

For example, if a reliable test index shows abnormally high anxiety, low self-esteem, poor self-control, or excessive anger, and a retest during treatment shows the same or a worse result for any of these treatment targets, there is objective evidence that no progress has been made on this front. Such results can then lead to an informed decision to alter the type or focus of the treatment, change the therapist, or await the next re-assessment before making any change. On the other hand, should retesting show an index closer to an adaptive range than initially, there is reason to conclude that progress is being made on the treatment target related to that index but that further improvement remains to be made in that area. When an initially abnormal test result is found on retesting to be in an adaptive range, then therapists and their patients can conclude with confidence that they have achieved the objective to which this result relates and do not need to address it further. At the point when retesting indicates that most or all of the treatment targets have reached or are approaching as much resolution as could realistically be expected, then the assessment process helps to indicate that an appropriate termination point has been reached.

Assessments conducted at the conclusion of psychotherapy, when compared with initial baseline evaluations, provide an objective basis for evaluating the overall benefit of the treatment that has been provided. Evaluations of treatment benefit made possible by pre-therapy and post-therapy assessments serve important research and practical purposes in clinical psychology. With respect to research issues, assessment data bearing on treatment benefit facilitates comparison studies of the relative effectiveness of different types and modalities of therapy. For practical purposes, retest findings demonstrating treatment benefit bear witness to the value of psychological interventions, particularly as weighed against the financial cost of these services.

## WIDELY USED INSTRUMENTS

Surveys of clinical psychologists and the contents of standard handbooks concerning psychological assessment identify several instruments as being among those most widely used by clinicians in the United States for purposes of differential diagnosis, treatment planning, and outcome evaluation. Four of these measures are relatively structured self-reports inventories on which conclusions are derived from what respondents are able and willing to say about themselves: the *Minnesota Multiphasic Personality Inventory*, the *Millon Clinical Multiaxial Inventory*, the *Sixteen Personality Factors Questionnaire*, and the *Personality Assessment Inventory*. Four of them are relatively unstructured performance-based measures in which the key data consist not of what respondents say about themselves but how they deal with various kinds of somewhat ambiguous tasks that are assigned to them: the Rorschach Inkblot Method, the Thematic Apperception Test, several types of figure drawing tasks, and some alternative sentence completion methods.

## CONCLUSIONS AND FUTURE PERSPECTIVES

Psychological assessment has been an integral part of clinical psychology since its inception and continues to the present day to provide practitioners with valuable information to guide their evaluation and treatment of persons who seek their help. At times, failure to appreciate the benefits of preceding treatment with thorough assessment has led to insufficient teaching and learning of psychodiagnostic methods by clinical psychologists, as has the regrettable and short-sighted devaluing of diagnostic procedures by health insurance providers. However, the future application of psychodiagnostic methods in clinical psychology appears to rest safely in the hands of practitioners and researchers whose knowledge from their experience and data how useful assessment can be in facilitating good clinical decisions.

## References

- Beutler, L.E. & Harwood, T.M. (1995). How to assess clients in pre-treatment planning. In Butcher, J.N. (Ed.), *Clinical Personality Assessment* (pp. 59–77). New York: Oxford.
- Blatt, S.L. & Ford, R.Q. (1994). *Therapeutic Change*. New York: Plenum.
- Garfield, S.L. (1994). Research on client variables in psychotherapy. In Bergin, A.E. & Garfield, S.L. (Eds.), *Handbook of Psychotherapy and Behaviour Change* (4th ed.; pp. 190–228). New York: Wiley.
- Greencavage, L.M. & Norcross, J.C. (1990). What are the commonalities among the therapeutic factors? *Professional Psychology, 21*, 372–378.
- Hayes, S.C., Nelson, R.O. & Jarrett, R.B. (1987). The treatment utility of assessment. *American Psychologist, 42*, 963–974.
- Horvath, O., & Greenberg, L.S. (Eds.) (1994). *The Working Alliance*. New York: Wiley.
- Hurt, S.W., Reznikoff, M. & Clarkin, J.F. (1991). *Psychological Assessment, Psychiatric Diagnosis, & Treatment Planning*. New York: Brunner/Mazel.
- Kubiszyn, T.W., Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R. & Eisman, E.J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology, 31*, 119–130.
- Maruish, M.E. (Ed.) (1999). *The Use of Psychological Testing for Treatment Planning and Outcome Assessment* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Shectman, F. & Smith, W.H. (Eds.) (1984). *Diagnostic Understanding and Treatment Planning*. New York: Wiley.
- Weiner, I.B. (1983). The future of psychodiagnosis revisited. *Journal of Personality Assessment, 47*, 451–461.
- Weiner, I.B. & Exner, J.E. (1991). Rorschach changes in long-term and short-term psychotherapy. *Journal of Personality Assessment, 56*, 453–465.

Irving B. Weiner

## Related Entries

APPLIED BEHAVIOURAL ANALYSIS, CHILD AND ADOLESCENT ASSESSMENT IN CLINICAL SETTINGS, CLINICAL JUDGEMENT, COUPLE ASSESSMENT IN CLINICAL SETTINGS, INTERVIEW IN BEHAVIOURAL CLINICAL AND HEALTH SETTINGS, OBSERVATIONAL TECHNIQUES IN CLINICAL SETTING, GOAL ATTAINMENT SCALING, PSYCHOPHYSIOLOGICAL EQUIPMENT AND MEASUREMENTS, OUTCOME ASSESSMENT/OUTCOME TREATMENT, PREDICTION: CLINICAL VS. STATISTICAL



## INTRODUCTION

The role of assessment and evaluation in education has been crucial, probably since the earliest approaches to formal education. However, change in this role has been dramatic in the last few decades, largely due to wider developments in society. The most dramatic change in our views of assessment is represented by the notion of assessment as a tool for learning. Whereas in the past, we have seen assessment only as a means to determine measures and thus certification, there is now a realization that the potential benefits of assessing are much wider and impinge on in all stages of the learning process. In this contribution, we will outline some of the major developments in educational assessment, and we will reflect on the future of education within powerful learning environments, where learning, instruction and assessment are more fully integrated.

### Consequences of the Developments in Society

Economic and technological change, which brings significant changes in the requirements of the labour market, poses increasing demands on education and training. For many years, the main goal of education has been to make students knowledgeable within a certain domain. Building a basic knowledge store was the core issue. Students taking up positions in modern organizations nowadays need to be able to analyse information, to improve their problem-solving skills and communication and to reflect on their own role in the learning process. People increasingly have to be able to acquire knowledge independently and use this body of organized knowledge in order to solve unforeseen problems. As a consequence, education should contribute to the education of students as lifelong learners.

### Paradigm Change: From Testing Towards Assessment

Many authors (Mayer, 1992; De Corte, 1990) have pointed to the importance of instruction to promote student's abilities as thinkers, problem-solvers and inquirers. Underlying this goal is the view that meaningful understanding is based on the active construction of knowledge and often involves shared learning. It is argued that a new form of education requires reconsideration about assessment (Dochy, Segers & Slujsmans, 1999). Changing towards new forms of learning, with a status quo for evaluation, undermines the value of innovation. Students do not invest in learning that will not be honoured. Assessment is the most determining factor in education for the learning behaviour of students. Traditional didactic instruction and traditional assessment of achievement are not suited to the modern educational demands. Such tests were generally designed to be administered following instruction, rather than to be integrated with learning. As a consequence, due to their static and product-oriented nature, these tests not only lack diagnostic power but also fail to provide relevant information to assist in adapting instruction appropriately to the needs of the learner (Campion & Brown, 1990; Dochy, 1994). Furthermore, standard test theory characterizes performance in terms of the difficulty level of response choice items and focuses primarily on measuring the amount of declarative knowledge that students have acquired.

This view of performance is at odds with current theories of cognition. Achievement assessment must be an integral part of instruction, in that they should reflect, shape, and improve student learning. Assessment procedures should not only serve as a tool for crediting students with recognized certificates, but should also be used to monitor progress and, if needed to direct students to remedial learning activities. The view that the evaluation of student's achievements is

something which happens at the end of the process of learning is no longer widespread; assessment is now represented as a tool for learning (Dochy & McDowell, 1997).

The changing learning society has generated the so-called assessment culture as an alternative to the testing culture. The assessment culture strongly emphasizes the integration of instruction and assessment. Students play far more active roles in the evaluation of their achievement. The construction of tasks, the development of criteria for the evaluation of performance, and the scoring of the performance may be shared or negotiated among teachers and students. The assessment takes all kinds of forms such as observations, text- and curriculum-embedded questions and tests, interviews, performance assessments, writing samples, exhibitions, portfolio assessment, and project and product assessments. Several labels have been used to describe subsets of these alternatives, with the most common being 'direct assessment', 'authentic assessment', 'performance assessment' and 'alternative assessment'.

### **New Methods of Assessment**

Investigations of new approaches (e.g. Birenbaum & Dochy, 1996; Nitko, 1995; Shavelson et al., 1996) illustrate the development of more 'in context' and 'authentic' assessment (Archbald & Newmann, 1992; Hill, 1993). Nisbet (1993) defines the term authentic assessment as 'methods of assessment which influence teaching and learning positively in ways which contribute to realizing educational objectives, requiring realistic (or "authentic") tasks to be performed and focusing on relevant content and skills, essentially similar to the tasks involved in the regular learning processes in the classroom' (p. 35).

Assessment of such 'authentic' tasks is highly individual and contextualized. The student gets feedback about the way he solved the task and about the quality of the result. Evaluation is given, on the basis of different 'performance tasks', performed and (reviewed) assessed at different moments. The evaluation criteria have to be known in advance. When students know the criteria and know how to reach them, they will be more motivated and achieve better results. This form of evaluation gives a more

complete and realistic picture of the student's ability (achievement). It evaluates not only the product, but also the process of learning. Students get feedback about their incorrect thinking strategies.

Within the new forms of 'new assessment', much attention is paid to authentic problem-solving, case-based exams, portfolios and the use of co- peer-, and self-assessment (Birenbaum, 1996).

In traditional education, the question 'who takes up the exam and who defines the criteria' is seldom asked. Most of the time, it is the teacher. New forms of education do pose this question. Students themselves, other students or the teacher and students together are responsible for assessment. The type of student self-assessment referred to most frequently in the literature is a process, which involves teacher-set criteria and where students themselves carry out the assessment and marking. Another form of student self-assessment is the case where a student assesses herself or himself, on the basis of criteria which she or he has selected, the assessment being either for the student's personal guidance or for communication to the teacher or others. According to Hall (1995) there are two critical factors for genuine self-assessment: the student not only carries out the assessment, but also selects the criteria on which the assessment is based. Similarly, peer-assessment can indicate that fellow students both select the criteria and carry out the assessment. Any situation where the tutor and students share in the selection of criteria and/or the carrying-out of the assessment is more accurately termed co-assessment (Hall, 1995). However, it is still frequently the case that teachers control the assessment process, sometimes assisted by professional bodies or assessment experts, whereas students' assessments and criteria are taken seriously but considered to be additional to the assessment undertaken by the teacher or professor rather than replacing it (Rogers, 1995). Implementing forms of self-, peer- and co-assessment may decrease the time-investment professors would otherwise need to make in more frequent assessment. In addition to that advantage, using these assessment forms assists the development of certain skills for the students, e.g. communication skills, self-evaluation skills, observation skills, self-criticism.

## ASSESSING NEW ASSESSMENT FORMS: DEVELOPMENTS IN EDMETRICS

Judgements regarding the cognitive significance of an assessment begin with an analysis of the cognitive requirements of tasks as well as the ways in which students try to solve them (Glaser, 1990). Two criteria by which educational and psychological assessment measures are commonly evaluated are validity and reliability. One can say that based on these criteria, the results above are not yet consistent and depending upon the assessment form there is a larger or smaller basis to state that the evaluation is acceptable.

It is however important to note that Birenbaum (1996) mentions that the meaning of validity and reliability has recently expanded. Dissatisfaction with the available criteria, which were originally developed to evaluate indirect measures of performance, is attributed to their insensitivity to the characteristics of a direct assessment of performance.

The most important element of new assessment models is the reflection of the competencies, required in real-life practice. The goal is to ensure that the success criteria of education or training processes are the same as those used in the practice setting. Hence, as notions of fitness of purpose change, and as assessment of more qualitative areas are developed, the concepts of validity and reliability encompassed within the instruments of assessment must also change accordingly. This means that we should widen up our view and search for other and more appropriate criteria. It should not be surprising that a new learning society and consequently a new instructional approach and a new assessment culture cannot be evaluated on the basis of the pre-era criteria solely.

### *Validity Related Issues*

Although performance assessment appears to be a valid form of assessment, in that it resembles meaningful learning tasks, this measure may be no more valid than scores derived from response choice items (Linn et al., 1991). Evidence is needed to assure that assessment requires the high-level thought and reasoning processes that they were intended to evoke.

The authors of the 1985 Standards define *test validity* as 'a unitary concept, requiring multiple lines of evidence, to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores.' (AERA, APA, NCME, 1985: 9) All validity research should be guided by the principles of scientific inquiry reflected in *construct validity*.

Within the construct validity framework, almost any information, gathered in the process of developing and using an assessment is relevant, when it is evaluated against the theoretical rationale underlying the proposed interpretation and inferences, made from test scores (Moss, 1995). Thus, validation embraces all the experimental statistical and philosophical means by which hypotheses are evaluated. Validity conclusions then, are best presented in the form of an *evaluative argument*, which integrates evidence to justify the proposed interpretation against plausible alternative interpretations.

Kane's *argument-based approach* is in line with Cronbach's view on validity. According to Kane (1992), to validate a test-score interpretation is to support the plausibility of the corresponding interpretative argument with appropriate evidence: (1) for the inferences and assumptions, made in the proposed interpretative argument and (2) for refuting potential counter arguments. The core issue is not that we must collect data to underpin validity, but that we should formulate transparent, coherent, and plausible arguments to underpin validity.

Authors like Kane and Cronbach use validity principles from interpretative research traditions, instead of psychometric traditions, to assist in evaluating less-standardized assessment practices.

Other criteria suggested for measuring validity of new assessment forms are the transparency of the assessment procedure, the impact of assessment on education, directness, effectiveness, fairness, completeness of the domain description, practical value and meaningfulness of the tasks for candidates, and authenticity of the tasks (Haertel, 1991). According to Messick (1994), these validation criteria are, in a more sophisticated form, already part of the unifying concept of validity, which he expressed in 1989. He asserted that validity is an evaluative summary of both evidence for and the actual as well as potential consequences of score interpretation and use. The more traditional conception of

validity as ‘evidence for score interpretation and use’ fails to take into account both evidence of the value implications of score interpretation and the social consequences of score use.

Messick’s unifying concept of validity encompasses six distinguishable parts – content, substantive, structural, external, generalizability, and consequential aspects of construct validity – that conjointly function as general criteria for all educational and psychological assessment. The *content* aspect of validity means that range and type of tasks, used in assessment must be an appropriate reflection (*content* relevance, representativeness) of the construct-domain. Increasing achievement levels in assessment tasks should reflect increases in expertise of the construct-domain. The *substantive* aspect emphasizes the consistency between the processes required for solving the tasks in assessment, and the processes used by domain-experts in solving tasks (problems). Further, the internal *structure* of assessment – reflected in the criteria, used in assessment tasks, the interrelations between these criteria and the relative weight placed on scoring these criteria, should be consistent with the internal structure of the construct-domain. If the content aspect (relevance, representativeness of content and performance standards) and the substantial aspect of validity is guaranteed, score interpretation, based on one assessment task should be *generalizable* to other tasks, assessing the same construct. The *external* aspect of validity refers to the extent that the assessment scores’ relationship with other measures and non-assessment behaviours reflect the expected high, low, and interactive relations. The *consequential* aspect of validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use (Messick, 1994).

In line with Messick’s conceptualization of consequential validity, Frederiksen and Collins (1989) proposed that assessment has ‘systematic validity’ if it encourages behaviours on the part of teachers and students that promote the learning of valuable skills and knowledge, and allows for issues of transparency and openness, that is to access the criteria for evaluating performance. Encouraging deep approaches to learning is one aspect, which can be explored in considering the consequences. Another is the impact which assessment has on teaching. Dochy

and McDowell (1997) argue that assessing high-order skills by means of authentic assessment will lead to the teaching of such high-order knowledge and skills.

With today’s emphasis on high-stakes assessment, two threats to test validity are worth mentioning: construct under-representation and construct-irrelevance variance. In the case of *construct-irrelevance variation*, the assessment is too broad, containing systematic variance that is irrelevant to the construct being measured. The threat of *construct-underrepresentation* means that the assessment is too narrow and fails to include important dimensions of the construct being measured.

### *Special Points of Attention For New Assessment Forms*

The above implies in our view that other criteria suggested for measuring validity of new assessment forms will need to be taken into account, i.e. the transparency of the assessment procedure, the impact of assessment on education, directness, effectiveness, fairness, completeness of the domain description, practical value and meaningfulness of the tasks for candidates, and authenticity of the tasks.

In addition, predictable difficulties will have to be taken into account, such as those outlined in the following paragraphs.

Authentic assessment tasks are more sensitive to construct-underrepresentation and construct-irrelevance variation, because they are often loosely structured, so that it is not always clear to which construct-domain inferences are drawn. Birenbaum (1996) argues that it is important to specify accurately the domain and to design the assessment rubrics, so they clearly cover the construct-domain. Messick (1994) advises to adopt a construct-driven approach to the selection of relevant tasks and the development of scoring criteria and rubrics, because it makes salient the issue of construct-underrepresentation and construct-irrelevance variation.

Another difficulty with authentic tasks, with regard to validity, is concerning rating authentic problems. Literature reveals that there is much variability between raters in scoring the quality of a solution. Construct-underrepresentation in rating is manifested as omission of assessment



criteria or idiosyncratic weighting of criteria such that some aspects of performance do not receive sufficient attention. Construct-irrelevance variance can be introduced by rater's application of extraneous, irrelevant or idiosyncratic criteria (Heller et al., 1998). Suggestions for dealing with these problems in literature include constructing guidelines, using multiple raters and selecting and training raters.

### *Reliability Related Issues*

Reliability in classical tests is concerned with the degree in which the same results would be obtained on a different occasion, in a different context or by a different assessor. Inter- and intrarater agreement is used to monitor the technical soundness of performance assessment rating. However, when these conventional criteria are employed for new assessments (for example using authentic tasks), results tend to compare unfavourably to traditional assessment, because of a lack of standardization of these tasks.

The unique nature of new forms of assessment has affected the traditional conception of reliability, resulting in the expansion of its scope and a change in weights attached to its various components (Birenbaum, 1996). In new assessment forms, it is not about achieving a normally distributed set of results. The most important question is to what extent the decision 'whether or not individuals are competent' is dependable (Martin, 1997). Differences between ratings sometimes represent more accurate and meaningful measurement than would absolute agreement. Measures of interrater reliability in authentic assessment then, do not necessarily indicate whether raters are making sound judgement and do not provide bases for improving technical quality. Measuring the reliability of new forms of assessment stresses the need for more evidence in a doubtful case, rather than to rely on making inferences from a fixed and predetermined set of data (Martin, 1997).

In line with these views on reliability is Moss' idea (1992) about reliability. She asserts that a hermeneutic approach of 'integrative interpretations based on all relevant evidence' is more appropriate for new assessment because it includes the value and contextualized knowledge of the reader, than the psychometric

approach that limits human judgement 'to single performances', results of which are then aggregated and compared with performance standards.

### CONCLUSION: FUTURE RESEARCH AND DEVELOPMENTS

The assessment culture leads to a change in our instructional system from a system that transfers knowledge into students' heads to one that tries to develop students who are capable of learning how to learn. The current societal and technological context requires education to make such a change. The explicit objective is to interweave assessment and instruction in order to improve education. A number of lessons can be learned from the early applications of new assessment programs.

First, one should not throw the baby out with the bath water. Objective tests are very useful for certain purposes, such as high-stake summative assessment of an individual's achievement, although they should not dominate an assessment program. Increasingly, measurement specialists recommend the so-called balanced or pluralistic assessment programs, where multiple assessment formats are used. There are several motives for these pluralistic assessment programs (Birenbaum, 1996; Messick, 1984): a single assessment format cannot serve several different purposes and decision-makers; and each assessment format has its own method variance, which interacts with persons.

There is a need to establish a system of assessing the quality of new assessment and implement quality control. Various authors have recently proposed ways to extend the criteria, techniques and methods used in traditional psychometrics. Others, like Messick (1995), oppose the idea that there should be specific criteria, and claim that the concept of construct validity applies to all educational and psychological measurements, including performance assessment.

### *References*

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985).

- Standards for Educational and Psychological Testing*. Washington, D.C.
- Archbald, D.A. & Newmann, F.M. (1992). Approaches to assessing academics achievement. In Berlak, H., Newmann, F.M., Adams, E., Archbald, D.A., Burgess, T., Raven, J. & Roberg, T.A. (Eds.), *Toward a New Science of Educational Testing and Assessment*. (pp. 139–180). Albany: State University of New York Press.
- Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In Birenbaum, M. & Dochy, F. (Eds.), *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge*. Boston: Kluwer Academic.
- Birenbaum, M. & Dochy, F. (Eds.) (1996). *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge*. Boston: Kluwer Academic.
- Campione, J.C. & Brown, A.L. (1990). Guided learning and transfer : Implications for approaches to assessment. In Frederiksen, N., Glaser, R., Lesgold, A.A. & Shafto, M.G. (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition* (pp. 141–172). Hillsdale, NJ: Lawrence Erlbaum Associates.
- De Corte E. (1990). Toward powerful learning environments for the acquisition of problem solving skills. *European Journal of Psychology of Education* 5(1), 519–541.
- Dochy, F. (1994). Prior knowledge and learning. In Husén, T. & Postlethwaite, T.N. (Eds.), *International Encyclopedia of Education* (2nd. ed., pp. 4698–4702). Oxford/New York: Pergamon Press.
- Dochy, F. & McDowell, L. (1997). Introduction: assessment as a tool for learning. *Studies in Educational Evaluation* 23(4), 279–298.
- Dochy, F., Segers, M. & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education*, 24(3), 331–350.
- Frederiksen, J.R. & Collins, A. (1989). A system approach to educational testing. *Educational Researcher* 18(9), 27–32.
- Glaser, R. (1990). *Testing and Assessment; O Tempora! O Mores!* Horace Mann Lecture, University of Pittsburgh, LRDC, Pittsburgh, Pennsylvania.
- Haertel, E.H. (1991). New forms of teacher assessment. In Grant, G. (Ed.), *Review of Research in Education*, 17, 3–29.
- Hall, K. (1995). *Co-assessment: Participation of Students with Staff in the Assessment Process. A Report of Work in Progress*. Paper given at the 2nd European Electronic Conference on Assessment and Evaluation, EARLI-AE list European Academic & Research Network (EARN) (listserv.surfnet.nl/archives/earli-ae.html).
- Heller, J.I., Sheingold, K. & Myford, C.M. (1998). Reasoning about evidence in portfolios: cognitive foundations for valid and reliable assessment. *Educational Assessment*, 5(1), 5–40.
- Hill, P.W. (1993). *Profiles and the VCE: Authentic Assessment in a High Stakes Environment*. Paper presented to the VCTA Comview Conference, Melbourne, 1 December.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Linn, R.L., Baker, E. & Dunbar, S. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Martin, S. (1997). Two models of educational assessment: a response from initial teacher education: if the cap fits. *Assessment and Evaluation in Higher Education*, 22(3), 337–343.
- Mayer, R.E. (1992) *Thinking, Problem Solving, Cognition* (2nd ed.). New York: Freeman.
- Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, 21, 215–238.
- Messick, S. (1994). The interplay of evidence and consequences in the validation performance assessments. *Educational Researcher*, 23(2), 13–22.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.
- Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research*, 62(3), 229–258.
- Moss, P.A. (1995). Themes and variations in validity theory. *Educational Measurement*, 2, 5–13.
- Nisbet, J. (1993). Introduction. In *OECD-Curriculum Reform: Assessment in Question*, 25–38. Paris: Organisation for Economic Cooperation and Development.
- Nitko, A. (1995) Curriculum-based continuous assessment: a framework for concepts, procedures and policy. *Assessment in Education*, 2, 321–337.
- Rogers, P. (1995). *Validity of Assessments*. Contribution to the 2nd EECAE Conference (European Electronic Conference on Assessment and Evaluation), EARLI-AE list, March 10–14.
- Shavelson, R.J., Xiaohong, G. & Baxter, G. (1996). On the content validity of performance assessments: centrality of domain-specifications. In Birenbaum, M. & Dochy, F. *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge* (pp. 131–142). Boston: Kluwer Academic.

Filip Dochy

### Related Entries

DIAGNOSTIC TESTING IN EDUCATIONAL SETTINGS, EVALUATION IN HIGHER EDUCATION (INCLUDES EVALUATING QUALITY IN HIGHER EDUCATION), INSTRUCTIONAL STRATEGIES, LEARNING DISABILITIES, LEARNING STRATEGIES, PSYCHOEDUCATIONAL TEST BATTERIES, REPORTING TEST RESULTS IN EDUCATION, STANDARD FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

# A

---

## APPLIED FIELDS: FORENSIC

### INTRODUCTION

Psychological forensic assessment aims to contribute to rational problem-solving in a forensic context when judgements have to be made about conditions or consequences of human behaviour brought to (criminal or civil) court. We describe a decision-oriented model of the process of psychological assessment that can serve as a general framework for psychological assessment concerning forensic questions. Frequently asked forensic questions relate to (1) psychological problems of parental custody and contact with children after divorce, (2) credibility of witness statements, and (3) prognosis of offence recidivism.

### GENERAL CONCEPT

Modern psychological forensic assessment is conceived as an aid for optimizing forensic problem solving in a scientific process of hypotheses-testing. The assessment process can be regarded as a sequence of decisions. Decisions during planning have a crucial impact on assessment results: mistakes in planning may cause invalid results. Additionally, many decisions must be made while realizing the assessment plan and combining the data into results. Explicit rules to aid these decisions are explained and compiled in checklists by Westhoff and Kluck (1998).

This approach is in contrast to the – outdated – trait-oriented comprehensive ‘portraying’ of the personality. According to this general concept, it is not the personality that has to be evaluated, but the conditions and the course of a person’s actions, or the relations between individuals, in the past, present and in the future. There are six sets of conditions influencing human behaviour: (1) environment; (2) organism; (3) cognition; (4) emotion; (5) motivation; and (6) social variables; and their interactions.

In a single case, all empirically relevant conditions and behavioural variables are checked for their contributions to the forensic question put to the psychological expert. In order to test the resulting hypotheses, different sources of information have to be selected, e.g. according to their psychometric properties. Data can be gathered from systematically planned interviews, observation of behaviour, biographical files and standardized procedures (e.g. tests or questionnaires). Assessors balance the costs of a special assessment procedure, e.g. a test, and its benefits. Of course, they take into consideration not only material, but also immaterial costs and benefits for each participant in the assessment process. A competent realization of the assessment plan requires the up-to-date knowledge and skills of a well-trained psychologist. This expert will use the most objective methods of documentation, e.g. tape recording of interviews.

Data from all relevant sources of information are weighted according to the single case and combined in order to reach a decision about each of the initial hypotheses. In a second step the outcomes of these decisions are integrated, in order to answer the forensic question(s) posed by the judicial system. The conclusions are always stated as probabilistic ‘if-then’-statements.

The structure of a psychological report according to this assessment process corresponds to the international scientific publication standards and the Guidelines for the Assessment Process (GAP) of the European Association of Psychological Assessment (Fernández-Ballesteros et al., 2001):

- 1 Client’s question (and client)
- 2 Psychological questions (= hypotheses)
- 3 Plan and sequence of the investigation (including the names of all investigators, all appointments; duration and locations of meetings)
- 4 Data

- 5 Results
- 6 Recommendations and suggestions (if asked for in the client's question)
- 7 References
- 8 Appendix (including psychometric calculations)
- 9 Signature (of the responsible psychologist)

## JUDICIAL SYSTEM AND FORENSIC QUESTIONS PUT TO THE PSYCHOLOGICAL EXPERT

The roles and the tasks of all the participants in legal proceedings differ according to the different judicial systems in Western societies. Consequently, the questions put to forensic psychological experts, and their working conditions, differ as well. Nevertheless, there are common basic forensic-psychological concepts and methods. The following sections will deal with them. They will be illustrated by sketches of the forensic questions most frequently put to psychological experts.

### Psychological Reports in Family Law

Writing a psychological report on questions of parental custody and contact of parents with their children after divorce is a very complex task which, primarily, needs thorough planning. Preparation of such a report aims to support the parents' readiness of communication with each other and their educational competence. The results of the psychological expert's work help the judge at the family court to decide in the 'best interest of the child'. Psychological experts optimize this assessment process by using explicit rules. Westhoff, Terlinden-Arzt, and Klueber (2000) explain every single decision that has to be made in this process. Additionally they give checklists containing rules to help avoid errors and mistakes and to minimize judgement biases.

To enable the parents and/or the judge to decide in the 'best interest of the child' requires the operationalization of this hypothetical construct. The psychological expert has to test the following sets of (psychological) variables:

- 1 the personal attachments of the child;
- 2 the continuity of personal care and the continuity environment of the child;

- 3 fostering the development of the child;
- 4 the attitudes of the child to possible solutions;
- 5 parents' readiness for communication with each other regarding the child;
- 6 their readiness to support the personal attachments of the child;
- 7 strategies of the family to cope with their divorce-related problems.

The psychological expert has to select the most useful, objective, reliable, and valid instruments for gathering the necessary data. There are only very few standardized procedures that match the questions asked by the family court. Most of the relevant data for psychological assessment in family court problems are obtained from systematic, partly standardized interviews and from the systematic observation of relevant behaviour (e.g. 'the strange situation' designed by Ainsworth et al. (1978) for the assessment of attachment quality). The Family Relations Test by Bene and Anthony (1985) can be very useful as a supporting instrument for the systematic interviewing of even young children: it helps the children to verbalize their incoming and outgoing emotions about each members of their family. The still widely used projective techniques as well as trait-oriented personality questionnaires are not validated for answering family court questions: the constitutional right to have or to rear children is not limited by a particular degree of any personality trait. Therefore, personality trait scores cannot be meaningful criteria in deciding with which of the parents the children should live or whether they should have contact with the other party.

### Statement Credibility

In criminal investigation, psychological experts may be asked to assess the credibility of statements by witnesses of a crime. Expert knowledge is mainly required in cases of sexual abuse and maltreatment or other violent crimes, especially when children are victims and/or witnesses of such offences and where there is no other evidence than the victim's/witness's statement. Nevertheless, the principal logic and the basic procedure of conducting an expert assessment is not limited to minors or to particular kinds of crime.

The assessment process here is again a hypothesis-testing procedure: starting with the assumption that the statement is not based on a real-life experience of the witness, the expert has to look out for data that rule out this hypothesis. Only if there is strong evidence for the alternative hypothesis, 'the statement is based on an experienced real-life event', can this alternative hypothesis be accepted. In contrast to this, the presupposition that the alleged event has actually occurred would only need very weak supporting evidence to be accepted and would therefore lead to an extremely false-positive bias.

Assessing the credibility of a witness's statement does not rely on 'general trustworthiness' as a kind of personality construct, but refers only to the assessment of the veracity of the specific testimony in a particular case. The general question of credibility assessment can, therefore, be stated as follows: 'Is this individual witness, under the given conditions of the investigation and the possible influences of other people, capable of making this particular statement even if it is not based on real-life experience?' (translated from Steller & Volbert, 1999).

The basic working hypothesis for analysing the content of a witness's statement was developed by Undeutsch (1967); it says that a statement that is based on real-life experience differs systematically from one that lacks this experience. For credibility assessment, this means that the witness's statement has to be analysed according to quality criteria applied to its content, which differentiate between reality-based statements and others. Reality criteria have been described since the beginning of the 20th century in German psychological and juridical literature. Undeutsch (1967) was the first to describe a comprehensive set of reality criteria. Steller and Koehnken (1989) refer to former approaches proposed by several authors and describe a system of five categories of reality criteria (pp. 221); these are:

- 1 general characteristics: logical structure, unstructured production, quantity of details;
- 2 specific contents: contextual embedding; descriptions of interactions; reproduction of conversation, unexpected complications during the incident
- 3 peculiarities of content: unusual details, superfluous details, accurately reported

details misunderstood, related external associations, accounts of subjective mental state; attribution of perpetrator's mental state;

- 4 motivation-related contents: spontaneous corrections, admitting lack of memory, raising doubts about one's own testimony, self-deprecation; pardoning the perpetrator;
- 5 elements specific to the offence: details characteristic of the offence.

This integrative expert system has experienced-enhanced theoretical foundation, (Ceci & Bruck, 1995). During the last fifteen years many studies of empirical validation of the system have been conducted in field studies and as well as in experimental studies. The criteria system has turned out to be a useful assessment instrument for scientific research and for practical use in assessing the credibility of a witness's statement.

Criteria Based Content Analysis (CBCA) can only lead to a valid credibility assessment if it takes into account certain characteristics of the witness as preconditions for a reliable and valid testimony. These are perception parameters, memory conditions and verbalization. In addition, there are motivational aspects to be considered like readiness to testify, goals, expectations, desires and fears connected with giving true or false testimony.

Furthermore there must be a test of whether there are or ever have been situational conditions that influence the statements so that they can even be made without that particular experience in real life. Statements by very young children in particular are susceptible to inductive and suggestive influences and questions, whether these are intentional or unintentional. Therefore, the 'history' of the statement and its development has to be explored, as well as the cognitive, emotional, and social developmental status of a young child witness.

The complete process of credibility assessment described here is called Statement Validity Assessment (SVA). In 1999, the Federal Supreme Court of Germany decided that expert opinions on the credibility of (child) witnesses are not acceptable in forensic contest unless they meet the standards of an SVA (Bundesgerichtshof, 1999).

Appropriate data for testing the above hypotheses for SVA are mainly obtained from biographical interviews; psychometric tests would have

to be selected with regard to their ecological validity for the special aspects of the abilities in the forensic context mentioned above. While severe limitations of sensory perception and developmental delays can be easily observed or assessed by psychometric or otherwise standardized methods, an appropriate test of 'memory' for SVA would have to test 'episodic' memory; a test of 'logical thinking' would have to refer to 'understanding social context'. Special tests of this kind are not yet available.

Consequently, the most important procedure for gathering data to run an SVA is therefore a non-suggestive, systematic interview of the witness (for interviewing strategies, see Milne & Bull, 1999). Observation of overt behaviour can be helpful in certain aspects, but most nonverbal cues (e.g. facial expression or illustrators during speaking) are ambiguous with regard to the veracity of a witness's statement (Koehnken, 1990).

### Prognosis of Offender Recidivism

Predicting the risk of recidivism of criminal offenders can very much influence the sentence and – in the case of mentally disordered offenders – the kind and duration of correctional treatment. This prediction task has to balance the severe consequences of false positive and of false negative judgements, both from the viewpoint of the individual offender and of the community.

Prognoses of offender recidivism are fraught with many specific and difficult problems: absolute certainty cannot be achieved by logical reason: the available data for prediction are incomplete; the only data about recidivism risks are those obtained about the individual offender; the important situational conditions can only be vaguely rated.

The process of psychological (and/or psychiatric) prognosis requires four steps of assessment (Rasch, 1999, Dahle, 1999): (1) analysis of the former criminal offences of the individual; (2) assessment of his present mental state (including possible mental disorders or illnesses); (3) analysis of the psychological development of the offender since the latest offence; (4) the general framework (situations, persons, chances) of his prospective living conditions. All these criteria are assessed according to the base rate of

individuals, where a similar constellation of conditions is observed.

Data for this prognosis task come from prison, hospital or therapy records, from some standardized psychodiagnostic questionnaires which have proven themselves as being reliable and valid predictors for criminal recidivism (such as the HCR-20 by Webster et al., 1994 and the Level of Service Inventory (LSI-R) by Andrews et al., 1995). Nevertheless, the most important method is the systematic interview with the offender based on the topics of the prognosis criteria.

### OTHER TOPICS OF FORENSIC ASSESSMENT

The three topics of forensic assessment described here are only examples. In different countries there are many other forensic questions that are put to the psychological expert. These concern for example: (1) assessment of criminal responsibility, (2) 'lie detection' by psychophysiological methods, (3) assessment of the effects of victimisation (4) and (other) special problems in civil law. The structure of the assessment process described above does not differ, however, for any forensic question whatsoever put to the forensic psychological expert.

### References

- Ainsworth, M.D.S., Blehar, M.C., Waters, E. & Wall, S. (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Hillsdale, NJ: Erlbaum.
- Andrews, D.A. (1995). The psychology of criminal conduct and effective treatment. In Mc Guire, J. (Ed.), *What Works: Reducing Reoffending* (pp. 35–62). Chichester: Wiley.
- Bene, E. & Anthony, J. (1985) *Family Relations Test, Children's Version, 1985 Revision*. Windsor: The NFER – Nelson Publishing Co. Ltd.
- Bundesgerichtshof (1999) 'Wissenschaftliche Anforderungen an aussagepsychologische Begutachtungen (Glaubhaftigkeitsgutachten). BGH, Urteil vom 30.7.1999 – 1 StR 618/98 (LG Ansbach)', *Neue Juristische Wochenschrift*, 2746–2751.
- Ceci, S.J. & Bruck, M. (1995) *Jeopardy in the Courtroom*. Washington, DC: APA.
- Dahle, K.-P. (1999). Psychologische Begutachtung zur Kriminalprognose. In Kröber, H.-L. & Steller, M. (Eds.), *Psychologische Begutachtung im Strafverfahren* (pp. 77–111). Darmstadt: Steinkopff.

- Fernández-Ballesteros, R., De Bruyn, E.E.J., Godoy, A., Hornke, L.F., Ter Laak, J., Vizcarro, C., Westhoff, K., Westmeyer, H. & Zaccagnini, J.L. (2001). Guidelines for the assessment process (GAP): a proposal for discussion. *European Journal of Psychological Assessment*, 17(3), 178–191.
- Koehnken, G. (1990). *Glaubwürdigkeit*. München: Psychologie Verlags Union.
- Milne, R. & Bull, R. (1999). *Investigative Interviewing – Psychology and Practice*. New York: Wiley.
- Rasch, W. (1999). *Forensische Psychiatrie* (2nd ed.) Stuttgart: Kohlhammer (1st ed., 1986).
- Steller, M. & Koehnken, G. (1989). Criteria-based statement analysis. Credibility assessment of children's statements in sexual abuse cases. In Raskin, D.C. (Ed.), *Psychological Methods for Investigation and Evidence* (pp. 217–245). New York: Springer.
- Steller, M. & Volbert, R. (1999). Forensisch-aussagepsychologische Begutachtung (Glaubwürdigkeitsbegutachtung), Gutachten für den BGH. *Praxis der Rechtspsychologie*, 9, 46–112.
- Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Aussagen. In Undeutsch, U. (Ed.), *Handbuch der Psychologie. Forensische Psychologie*, Band 11 (pp. 26–181). Göttingen: Hogrefe.
- Webster, C., Harris, G., Rice, M., Cormier, C. & Quinsey, V. (1994). *The Violence Prediction Scheme: Assessing Dangerousness in High Risk Men*. Toronto: Centre of Criminology, University of Toronto.
- Westhoff, K. & Kluck, M.-L. (1998). *Psychologische Gutachten schreiben und beurteilen* (3rd ed.) Berlin: Springer (1st ed., 1991).
- Westhoff, K., Terlinden-Arzt, P. & Klueber, A. (2000). *Entscheidungsorientierte Psychologische Gutachten für das Familiengericht*. Berlin: Springer.

Marie-Luise Kluck and Karl Westhoff

### Related Entries

PROCESS (THE ASSESSMENT PROCESS), CHILDREN CUSTODY

## A APPLIED FIELDS: GERONTOLOGY

### THE NEED FOR PSYCHOLOGICAL ASSESSMENT OF OLDER ADULTS

Older adults and particularly those frequently described as the 'oldest old' (85+) represent the fastest growing population subgroup in most (industrialized) countries around the world. Although high competence characterizes the majority of today's elders (Lehr & Thomae, 2000), a whole gamut of critical situations related to ageing, and particularly to very old age, underscores the need for psychological assessment in older adults. Psychological assessment provides a rational, scientific means for making decisions in these situations, prototypical examples of which are residential decisions (e.g. relocation to an institution or within institutions), treatment decisions (e.g. early diagnosis of dementia coupled with a promising cognitive training intervention), or rehabilitation decisions (e.g. the estimation an individual's rehabilitation potential and remaining plasticity).

In order to define the content of this article, we first draw from Lawton and Storandt (1984), who suggested a broad conception of

assessment: 'An attempt to evaluate the most important aspects of the behaviour, the objective, and the subjective worlds of the person (...)' (p. 258). Second, we argue for a theoretical framework to organize the different types of assessment and numerous instruments found in this rapidly evolving field of gerontology. Our suggestion is to roughly distinguish between three assessment approaches: (1) Person-oriented (P) assessment is aimed to address the older person's cognitive and behavioural competence, personality, and psychological aspects of health. (2) Environment-oriented (E) assessment addresses the social and the physical environment of the ageing person. (3) Finally, the assessment of P×E outcomes evaluates the impact of person–environment transactions on major domains of life quality such as subjective well being, affect, and mental health. Below, we use this line of thinking to review psychological assessment in gerontology. The challenges of assessing older persons in terms of application and theoretical–methodological issues are discussed shortly thereafter. We end this entry with some general conclusions and the consideration of future perspectives.

## MAIN APPROACHES IN THE ASSESSMENT OF OLDER PERSONS AND THEIR ENVIRONMENTS

The following overview draws from both old and new treatments of the assessment of older adults (e.g. Kane & Kane, 2000; Lawton & Storandt, 1984; Lawton & Teresi, 1994). Due to space limitations, each theoretical domain is illustrated using a small number of prototypical instruments that essentially reflect the construct or family of constructs in question (see also Table 1).

### Person-oriented Assessment

#### *Cognitive and Behavioural Competence*

Cognition is a major aspect of behavioural competence which undergoes particular decline in the later years. However, two reservations are warranted: First, this is true only for speed-dependent cognitive abilities ('fluid intelligence' in contrast to 'crystallized' intelligence); second, pronounced interindividual variability in performance is characteristic for old age. To test an individual's intellectual ability against the norm, the well-known *Wechsler Adult Intelligence Scale* (WAIS) is a classic in the field of ageing (Wechsler, 1981). Also, while there is a high correlation between cognitive functioning and the so-called 'Activities of Daily Living' (ADL; basic activities such as eating, washing, or dressing) as well as the 'Instrumental Activities of Daily Living' (IADL; more complex activities such as preparing meals, using the phone, or shopping), a separate assessment of ADL and IADL is nevertheless recommended to afford a comprehensive picture of the everyday competence of the older person. Respective assessment procedures (e.g. the classic scale proposed by Lawton and Brody, 1969) have proven to be powerful predictors of institutionalization and mortality. To further complement the evaluation of everyday competence, an additional assessment of leisure activities using an activity list or diary is helpful (Mannell & Dupuis, 1994).

#### *Personality*

There has been some debate in psychological gerontology regarding the question of whether personality traits such as the 'Big Five'

(neuroticism, extraversion, openness to experience, agreeableness, conscientiousness; Costa & McCrae, 1985) remain stable across the adult lifespan. Moderate stability has been widely confirmed, with a tendency toward lower stability over correspondingly longer observation periods. From a practical perspective, a recurring question is whether so-called 'problem behaviours' (such as antisocial behaviour, health-related risk behaviours, or the nonuse of existing competencies) may be better explained by individual differences in personality. In this regard, the *NEO Personality Inventory* (Costa & McCrae, 1985) is a classic assessment device that has been used intensively with elders. Reservations have to be made regarding the practical utility of these and other personality instruments with respect to the very old and those suffering from mild cognitive impairments; short scales with easily understood items are still rare. Besides standardized testing, a careful semi-structured exploration of the biography and major (and often critical) turning points therein is essential for an in-depth understanding of an older person's current strengths and weaknesses (Lehr & Thomae, 2000).

In a process-oriented perspective of personality, two constructs are particularly useful to explain situation-specific outcomes such as subjective well-being: coping and control. A classic coping instrument is the *Ways of Coping Checklist*, which has also been proved as useful in a shortened version, helpful for assessing the very old (Folkman, Lazarus, Pimley & Novacek, 1987). For measurement of perceived control, we recommend a short instrument newly developed within the context of the Berlin Aging Study (Smith & Baltes, 1999; Smith, Marsiske & Maier, 1999).

#### *Health*

Gaining clarity on the influences of health impairments is important for psychological assessment in any age group. However, this is particularly true for older persons. Chronic conditions and multimorbidity occur frequently in later life and are among the most influential explanations of subjective well being, depression, and the loss of independence. From a psychological perspective, the subjective evaluation of health based on a single-item assessment ('How would you rate your overall health at the present



Table 1. Recommendation of assessment instruments for use with older adults<sup>a</sup>

Assessment domain	Prototypical instrument	Application issues and selected psychometric information <sup>b</sup>
<i>Person-oriented assessment</i>		
Cognitive and behavioural competence	Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1981)	Very widely used; takes about 1.5 hours to administer <sup>c</sup> ; Cronbach's alpha of all subscales >0.70; broad evidence underlining validity.
	Activities/Instrumental Activities of Daily Living Scale (ADL/IADL) (Lawton & Brody, 1969)	Very widely used; takes about 5 minutes to administer; Cronbach's alpha of both scales >0.80; inter-rater $r$ 0.61 (ADL) and 0.91 (IADL); broad evidence underlining validity.
Personality	NEO Personality Inventory (Costa & McCrae, 1985)	Very widely used; takes about 20 minutes to administer; Cronbach's alpha of all subscales >0.70; broad evidence underlining validity.
	Ways of Coping Checklist (Short) (Folkman et al., 1987)	Frequently used; takes about 10 minutes to administer; Cronbach's alpha of subscales 0.47–0.74; some evidence underlining validity.
	Perceived Control (Smith et al., 1996 <sup>d</sup> )	Instrument introduced in the Berlin Aging Study; takes about 10 minutes to administer; some evidence underlining reliability and validity.
Health (psychological aspects)	SF-36 (Ware & Sherbourne, 1992)	Frequently used; takes about 10 minutes to administer; Cronbach's alpha of subscales 0.57–0.94; some evidence underlining validity.
<i>Environment-oriented assessment</i>		
Social environment	Social Networks in Adult Life Survey (Kahn & Antonucci, 1980)	Frequently used; administration time depends on persons nominated as social network members; on an average about 30 minutes; reasonable degree of convergence between respondents' and significant others' report; some evidence underlining validity.
	UCLA Loneliness Scale (Russel et al., 1980)	Frequently used; takes about 10 minutes to administer; Cronbach's alpha >0.90; some evidence underlining validity.
	Burden Interview (Zarit et al., 1980)	Frequently used; takes about 10 minutes to administer; Cronbach's alpha >0.70; some evidence underlining validity.
Physical environment	The Housing Enabler (Iwarsson, 1999)	Recently developed instrument; takes about 1.5 hours to administer; inter-rater reliability mean kappas for the different domains assessed 0.68–0.87; some evidence underlining validity.
	Multiphasic Environmental Assessment Procedure (Moos & Lemke, 1996)	Frequently used; data-collection time depends on the size of the institution to be assessed; can take up to about 1 week; Cronbach's alpha of subscales 0.44–0.96; some evidence underlining validity.

(continued)

Table 1. Continued

Assessment domain	Prototypical instrument	Application issues and selected psychometric information <sup>b</sup>
<i>Assessment of person × environment outcomes</i>		
Subjective well-being and affect	Philadelphia Geriatric Center Morale Scale (PGCMS) (Lawton, 1975)	Very widely used; takes about 10 minutes to administer; Cronbach's alpha >0.80 (total score); broad evidence underlining validity.
	Scales of Psychological Well-being (Ryff, 1989)	Frequently used; takes about 20 minutes to administer; Cronbach's alpha of all subscales >0.70; some evidence underlining validity.
	Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988)	Frequently used; takes about 5 minutes to administer; Cronbach's alpha >0.70; some evidence underlining validity.
Mental health	Center of Epidemiological Studies of the Elderly Depression Scale (CES-D) (Radloff, 1977)	Very widely used; takes about 10 minutes to administer; Cronbach's alpha >0.80; broad evidence underlining validity.
	Mini-Mental-State Examination (MMSE) (Folstein et al., 1975)	Very widely used; takes about 10 minutes to administer; inter-rater $r >0.80$ ; broad evidence underlining validity.

<sup>a</sup>See also additional description of these instruments in the text.

<sup>b</sup>The psychometric information given here is based on additional published evidence, which is not explicitly cited in this article due to space limitation.

<sup>c</sup>The estimation of duration always refers to the administration with old and very old persons.

<sup>d</sup>We recommend direct contact with the authors of this instrument for more information.

time: excellent, good, fair, or poor') has proven to be a powerful predictor of subjective well-being in many studies. A multi-item assessment of this construct as well as other health-related aspects is provided by the now classic *SF-36* (Ware & Sherbourne, 1992). Frequently overlooked in its impact on everyday life and well-being, the assessment of pain and its psychosocial impact is recommended as a must for any comprehensive health evaluation of older adults (Parmelee, 1994).

## Environment-Oriented Assessment

### *Social Environment*

Aspects of the social environment include the objective size of the social network, the amount of real and perceived social support, interpersonal conflicts, and overall social network evaluations, such as loneliness. Caregivers are a significant part of elders' social environments. A classic instrument to measure social network size as well as some of its major qualitative characteristics is the *Social Networks in Adult Life Survey*

(Kahn & Antonucci, 1980). This instrument defines social network membership using concentric circles, an approach that has proven to be very helpful in differentiating members of the social network in terms of closeness and importance. Another well-established tool to assess the existing network is the *UCLA Loneliness Scale* (Russell, Peplau, & Cutrona, 1980) addressing how often the respondent feels isolated and misunderstood and wishes to be involved in more social relationships. Caregivers persons deserve the attention of psychologists as well, given the extensive strain associated with this task and the increased risk of becoming physically and mentally ill. An instrument for assessing the stress of caregivers is the *Burden Interview* suggested by Zarit and colleagues in the early 1980s (Zarit, Reever, & Bach-Person, 1980).

### *Physical Environment*

Physical environments optimally adapted to the needs of frail elders can take on powerful supportive and stimulating functions in old age (for a review of the according empirical literature,

see Wahl, in press). Gitlin (1998) concluded in her review of checklists providing a comprehensive assessment of the home environment that the psychometric properties of most of these devices are at best unclear. Among the rare strictly tested instruments, we would recommend the 'The Housing Enabler' as a promising tool that carefully considers the physical home environment as well as the functional profile of older persons acting within these environments (Iwarsson, 1999). Although many different suggestions have been tossed around, there is no single device with well-proven psychometric properties currently available. In contrast, the assessment of institutional environments serving the elderly has found much attention and more canalized research efforts. A comprehensive measurement device is the *Multiphasic Environmental Assessment Procedure* (MEAP), which is based on a wide-ranging research program conducted by Moos and associates (Moos & Lemke, 1996) and has also been transferred to other countries (e.g. Fernandez-Ballesteros et al., 1991).

### Assessment of Person × Environment Outcomes

#### *Subjective Well-being and Affect*

Subjective well-being, or the cognitive and affective evaluation of the past and present life, has been regarded as a major indicator of successful ageing. The most highly renowned instrument probably is the *Philadelphia Geriatric Center Morale Scale* (PGCMS) (Lawton, 1975). This relatively easy-to-use 17-item scale covers three dimensions of subjective well-being, i.e. agitation, satisfaction with the ageing process, and general life-satisfaction. Due to the clinical nature of this instrument with many items addressing negative thoughts and emotions, it is particularly useful in the clinical, psychological evaluation of an older person, while other instruments more thoroughly address the positive facets of subjective well-being (e.g. Ryff, 1989).

Compared to subjective well-being, the measurement of affect has not yet found very much empirical attention (Labouvie-Vief, 1999). The term "affect" includes emotions, moods, and feeling states, all of which can be assessed in terms of intensity, frequency, and duration.

A promising assessment tool for use with elders is the *Positive and Negative Affect Schedule* (PANAS) suggested by Watson, Clark, & Tellegen (1988).

#### *Mental Health*

Within the spectrum of mental health threats in later life, depression is, besides dementia, the major disease, whose optimal detection requires a combination of expertise from clinical psychology and psychiatry. The *Center of Epidemiological Studies of the Elderly Depression Scale* (CES-D) introduced by Radloff (1977) is widely used, has proven psychometric properties, and works well in elderly populations. Although a score of 17 is widely accepted as an indication of a depressive illness, it is wise to always include at least one other source of information (such as a clinical expert rating) before a final diagnostic decision is made. In addition, because severe cognitive impairments substantially increase as people age – with estimated dementia rates of about 25% beyond the age of 85 years – dementia assessment should be included as a routine part of every older person's clinical evaluation. A classic screening test in this regard is the *Mini-Mental State Examination* (MMSE), originally suggested by Folstein, Folstein, and McHugh (1975). A major advantage of this widely used device is its scoring system, which is well known among clinicians and thus significantly facilitates communication (a score of 23 is generally recommended as indicative of cognitive dysfunction).

### SPECIFIC CHALLENGES OF ASSESSING OLDER ADULTS

A number of factors can threaten the internal and external validity of assessing older persons. In the following, only a selective overview of these issues can be provided.

Two messages are important in terms of practical test application: on the one hand, old age is associated with a slowing in fine motor functioning and reaction time, the loss of sensory functioning, and cognitive impairment. One consequence of this is that performance tests that require motor behaviour may not be adequate, at least in some elderly subpopulations (such as geriatric patients).

Furthermore, scales which are normally self-administered (e.g. personality tests) must frequently be administered by a third person, which means, as compared to other age groups, a substantial change in the social psychology of the test situation, for instance in terms of self-disclosure. The length of the instrument is particularly critical in case of very old persons. Furthermore, the response format should remain stable within testing sessions and should be as simple as possible (not more differentiated than a 5-point Likert-type scale). Also, motivational issues, including fatigue, must be considered when creating optimal test circumstances. On the other hand, test strategies found to be very effective and economic in younger persons, such as phone and computer-based assessment, can, in many cases, be transferred to older people as well. With respect to demented elders, the use of observational methods is frequently the only well-functioning assessment procedure for evaluating behaviour and inner states.

A major theoretical-methodological challenge of assessing older persons is the issue of construct invariance. For instance, constructs such as depression or pain might have a fundamentally different semantic at the age of twenty than at the age of ninety years. Moreover, measures might have age-related characteristics with respect to response bias, response format, or the production of missing data. These and other issues as well as tentative solutions have intensively been addressed by Teresi and Holmes (1994).

To conclude, we urge researchers and practitioners to adopt an attitude of 'constructive caution' in interpreting and using test results gathered in elderly populations.

## FUTURE PERSPECTIVES AND CONCLUSIONS

The assessment of older persons is an important field of gerontology in terms of research and application. Due to the multitude of measurement instruments suggested in the gerontological literature, it is essential to carefully check the proven psychometric properties and practical usefulness of these devices for making adequate instrument selections. Standardized tests, semi-structured assessments, and observational methods should serve as complementary tools in any comprehensive clinical evaluation. An important

task of future research is, as is so often the case, replicative research including different subgroups of elders and the revision of existing devices in order to improve the critical mass of good instruments. The assessment procedures so developed should provide a broad, reliable, and valid description of both the positive and negative sides of the ageing individual.

## Acknowledgement

Comments of David Burmedi and Mike Martin on an earlier draft of this entry are very much appreciated.

## References

- Costa, P.T. & McCrae, R.R. (1985). *The NEO-Personality Inventory. Manual Form S and Form R*. Odessa, FL: Psychological Assessment Resources.
- Fernandez-Ballesteros, R. et al. (1991). Evaluation of residential programs for the elderly in Spain and United States. *Evaluation Practice*, 12, 159–164.
- Folkman, S., Lazarus, R.S., Pimley, S. & Novacek, J. (1987). Age differences in stress and coping processes. *Psychology and Aging*, 2, 171–184.
- Folstein, M.F., Folstein, S.E. & McHugh, P.R. (1975). Mini mental state: a practical method of grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189–198.
- Gitlin, L.N. (1998). Testing home modification interventions: issues of theory, measurement, design, and implementation. In Schulz, R., Maddox, G. & Lawton, M.P. (Eds.), *Focus on Interventions Research with Older Adults*, Vol. 18 (pp. 190–246). New York: Springer.
- Iwarsson, S. (1999). The housing enabler: an objective tool for assessing accessibility. *British Journal of Occupational Therapy*, 62, 491–97.
- Kahn, R.L., & Antonucci, T.C. (1980). Convoys over the life course: attachment, roles, and social support. In Baltes, P.B. & Brim, O.G. (Eds.), *Life-span Development and Behaviour* (pp. 253–286). New York: Academic Press.
- Labouvie-Vief, G. (1999). Emotions in adulthood. In Bengtson, V.L. & Schaie, K.W. (Eds.), *Handbook of Theories of Aging* (pp. 253–267). New York: Springer Publishing.
- Kane, R.L. & Kane, R.A. (Eds.) (2000). *Assessing Older People: Measures, Meaning and Practical Applications*. New York, NY: Oxford University Press.
- Lawton, M.P. (1975). The Philadelphia geriatric center morale scale: a revision. *Journal of Gerontology*, 30, 85–89.
- Lawton, M.P. & Brody, E.M. (1969). Assessment of older people: self-maintaining and instrumental activities of daily living. *The Gerontologist*, 9, 179–185.

- Lawton, M.P. & Storandt, M. (1984). Assessment of older people. In Reynolds, P.M. & Chelune, G.J. (Eds.), *Advances in Psychological Assessment*, Vol. 6 (pp. 236–276). San Francisco: Jossey-Bass.
- Lawton, M.P. & Teresi, J.A. (Eds.). (1994). *Focus on Assessment Techniques* Vol. 14. New York: Springer.
- Lehr, U.M., & Thomae, H. (2000). *Psychologie des Alterns* [Psychology of ageing] (9th ed.). Wiebelsheim: Quelle & Meyer.
- Mannell, R.C. & Dupuis, S.L. (1994). Leisure and Productive Activity. In Lawton, M.P. & Teresi, J.A. (Eds.), *Focus on Assessment Techniques*, Vol. 14 (pp. 125–141). New York: Springer.
- Moos, R.H., & Lemke, S. (1996). *Evaluating Residential Facilities: The Multiphasic Environmental Assessment Procedure*. Thousand Oaks, CA: Sage.
- Parmele, P.A. (1994). Assessment of pain in the elderly. In Lawton, M.P. & Teresi, J.A. (Eds.), *Focus on Assessment Techniques*, Vol. 14 (pp. 281–301). New York: Springer.
- Radloff, L.S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Journal of Applied Psychological Measurement*, 1, 387–393.
- Russell, D.W., Peplau, L.A. & Cutrona, C.E. (1980). The revised UCLA loneliness scale: concurrent and discriminant validity evidence. *Journal of Personality and Social Psychology*, 39, 472–480.
- Ryff, C.D. (1989). Happiness in everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, 57, 1069–1081.
- Smith, J. & Baltes, P.B. (1999). Trends and profiles of psychological functioning in very old age. In Baltes, P.B. & Mayer, K.U. (Eds.), *The Berlin Aging Study. Aging from 70 to 100* (pp. 197–226). Cambridge: Cambridge University Press.
- Smith, J., Marsiske, M., & Maier, H. (1996). *Differences in Control Beliefs from Age 70 to 105*. Unpublished manuscript. Max Planck Institute for Human Development, Berlin.
- Teresi, J.A., & Holmes, D. (1994). Overview of Methodological Issues in Gerontological and Geriatric Measurement. In Lawton, M.P. & Teresi, J.A. (Eds.), *Focus on Assessment Techniques*, Vol. 14 (pp. 1–22). New York: Springer.
- Wahl, H.-W. (in press). Environmental influences on ageing and behaviour. In Birren, J.E. & Schaie, K.W. (Eds.), *Handbook of the Psychology of Aging* (5th ed.). San Diego: Academic Press.
- Ware, J.E., & Sherbourne, C.D. (1992). The MOS 36-item short-form healthy survey (SF-36). *Medical Care*, 30, 473–483.
- Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale – Revised Manual*. New York: The Psychological Corporation.
- Zarit, S.H., Reever, K.E., & Bach-Peterson, J. (1980). Relatives of the impaired elderly: correlates of feelings of burden. *The Gerontologist*, 20, 649–655.

Hans-Werner Wahl and Ursula Lehr

### Related Entries

DEMENTIA, QUALITY OF LIFE, HEALTH, DYNAMIC ASSESSMENT (LEARNING POTENTIAL, TESTING-THE-LIMITS), COGNITIVE PLASTICITY, COGNITIVE DECLINE/IMPAIRMENT, FLUID AND CRYSTALLIZED INTELLIGENCE, AUTO-BIOGRAPHY, INTELLIGENCE THROUGH COHORT AND TIME, CAREGIVER BURDEN.

## A APPLIED FIELDS: HEALTH

### INTRODUCTION

Health psychology is a field within psychology that is devoted to understanding psychological influences on health-related processes, such as why people become ill, how they respond to illness, how they recover from a disease or adjust to chronic illness, and how they stay healthy in the first place (Schwarzer & Gutiérrez-Doña, 2000). Health psychologists conduct research on the origins and

correlates of diseases. They identify personality or behavioural antecedents that influence the pathogenesis of certain illnesses. Health psychologists analyse the adoption and maintenance of health behaviours (e.g. physical exercise, nutrition, condom use, or dental hygiene) and explore the reasons why people adhere to risk behaviours (e.g. why they continue to smoke or drink alcohol). Health promotion and the prevention of illness are, therefore, agendas for research and practice, as is

the improvement of the health care system in general.

In health psychology, a multitude of variables are assessed, such as physical conditions, health behaviours, quality of life, coping with stress or illness, coping resources, and premorbid personality. Since health behaviours dominate the discipline, the following contribution will focus on this particular subarea.

## HEALTH BEHAVIOURS

Many health conditions are caused by such behaviours as problem drinking, substance use, smoking, reckless driving, overeating, or unprotected sexual intercourse. Health behaviours are often defined as behaviours that people engage in to maintain or improve their current health and to avoid illness. They include any behaviour a person performs in order to protect, promote, or maintain his or her health, whether or not such behaviours are objectively effective towards that end (Conner & Norman, 1996; Schwarzer & Renner, 2000).

People are inconsistent in the way they practice multiple health behaviours. For example, a person who exercises regularly does not necessarily adhere to a healthy diet. One reason people's current health habits are not more consistent is that they differ on a number of dimensions (see Table 1).

For a valid and reliable measurement of health behaviours, it is essential to distinguish between these dimensions and to define clearly the subject matter under investigation.

## ASSESSMENT OF HEALTH BEHAVIOURS

There are various methods of assessing health behaviours (Renner, in press). *Questionnaires* that

assess the frequency of past behaviour are the most commonly used methods. There are numerous questionnaires that ask for the average or typical quantity and frequency of alcohol consumption (for an overview, see Sobell and Sobell, 1995), dietary habits, or physical activity. However, the information provided by quantity and frequency measures (QF estimates) is limited because respondents must base their estimates on a large variety of experiences. QF estimates often reflect less drinking and tend to misclassify drinkers compared to daily diary or timeline reports. They also provide lower absolute food intake estimates than a longer, interviewer-administered diet history.

In rare occasions, *physiological methods* can be used, which are most accurate for measuring alcohol consumption (via blood or urine sampling), drug consumption (via immunoassay, hair or sweat bioassay procedures), habitual dietary intakes (via biochemical markers), or physical activity (via doubly labelled water). However, such bioassay methods are only required when a high level of accuracy about recent health behaviour is needed (e.g. for workplace drug testing). They can also be used in addition to self-report data in order to confirm or falsify self-report information (e.g. about recent drug use). However, in some circumstances it may only be necessary to lead respondents to believe that there is an objective way to identify their behaviours via physiological measures, which is done to reduce misreporting. Another direct method is *behavioural observation*, used to assess physical activity among children or a driver's speed, for example.

Unstructured or semistructured *interviews* are qualitative techniques for research on understanding individuals' cognitive and conceptual models of health behaviours and the frames of reference used to organize these behaviours. Therefore, qualitative methods are mainly concerned with exploration and analysis of health behaviour because they

Table 1. Dimensions of health behaviours

<ul style="list-style-type: none"> <li>● Voluntary; consciously undertaken by the individual</li> <li>● Avoidance of harmful activities</li> <li>● Undertaken without medical assistance</li> <li>● Vital</li> <li>● Occasional; unstable</li> <li>● Simple</li> </ul>	<ul style="list-style-type: none"> <li>● Involuntary; unconsciously undertaken by the individual</li> <li>● Engagement in protective activities</li> <li>● Needs professional medical assistance</li> <li>● Non-vital</li> <li>● Habitual; stable</li> <li>● Complex, multifaceted</li> </ul>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

allow the interviewee to address the issues that are relevant to the topics raised by the investigator. One major disadvantage of qualitative methods is that generality is, by definition, not quantifiable. Furthermore, since anonymity is not given, self-reports may be affected by social desirability biases, which lead to overreporting of socially desirable behaviours as well as underreporting of socially undesirable behaviours.

Stone and Shiffman (1994) have labelled strategies for collecting self-reports of respondents' momentary or current state as *Ecological Momentary Assessment* (EMA). EMA studies usually consist of repeated assessment of participants' momentary state as they go about the tasks of daily living in their natural environment. Interval-contingent assessments require assessment at regular intervals. One example is the method of interactive voice response where alcoholics are asked to call in on a regular basis to report their drinking status to the interviewers. Another way is asking respondents to record every episode of smoking, eating, or another behaviour of interest. This event-contingent approach may not lead to a representative sample of the participant's general state, and it requires a clear definition of the triggering event. In contrast, signal-contingent sampling supplies participants with an external signal cue that is usually timed to be emitted at random to prompt them to complete a written assessment or an electronic diary. Signal device beepers, electronic watches, and palmtop computers can be used. EMA is a method that precisely assesses recent health behaviours. Its major advantage is that it minimizes deviations due to recall from memory by relying on respondents' reports of their experience at the very moment of inquiry.

A *diary log* is a data collection strategy that gathers information as time passes. The distinctive feature of this method is that it yields information that is temporally ordered. It shows the sequence of events and the profile of actions across time. Diary techniques can be particularly useful when data from the same person are required over a considerable period of time and/or very frequently, such as assessing smoking behaviour, alcohol consumption, or dietary habits, in order to provide a general estimate of the amounts consumed. For example, alcohol consumption diaries often include questions

about the frequency of drinking, the type of drink, and the typical quantity consumed on each occasion. In comparison to questionnaires, the diary log format minimizes recall biases associated with retrospective reporting, but daily reporting may be more reactive. In addition, diaries could be valuable for getting access to so-called 'intimate' information (e.g. sexual behaviour).

*Timeline Followback Method Reports* (TLFB), developed by Sobell and Sobell (1995), provide a detailed insight into health behaviours (smoking, taking drugs, or drinking, etc.) over a designated time period. Participants are asked to provide retrospective estimates of their daily behaviour by using a calendar over a certain time period, ranging up to 12 months prior to the interview. With this method, the pattern, variability, and level of drinking or smoking can be profiled, which is especially useful when precise estimates are needed or when researchers wish to evaluate specific changes in health behaviours before, during, and after interventions. However, this is a rather time-consuming method.

## BIASES IN SELF-REPORTS

Some problems shared by all surveys relying on self-reports could seriously decrease internal and external validity (Schwarz & Strack, 1991). Short-term fluctuations, such as in substance use, produced by environmental (e.g. social settings) and psychological (e.g. mood or stress) variables, may affect the psychometric properties of usage measures. For example, there is a tendency for students to become increasingly exuberant as their high school graduation approaches. Increased party activity during the spring months contributes significantly to the actual level of drug use. Therefore, seasonal effects and short-term fluctuations may lead to superficial behavioural changes that could be misinterpreted by researchers as being genuine changes.

Questions about past behaviours assume accurate memory of events as well as willingness to report them to a researcher. However, respondents might not recall the actual events, employing instead various cognitive heuristics (rules of thumb) to estimate frequencies. This could result in certain

biases. Individuals use different strategies to answer frequency questions over different time spans. Episodic enumeration (recalling and counting individual incidents) is more likely to be used with shorter time spans in frequency reports, whereas rate-based estimation (projecting the typical rate over the length of the recall period) is more likely to be used when longer time spans are involved. Reported behavioural frequencies for a year are generally lower than 12 times the equivalent frequencies for a month. People probably forget more behavioural instances over the time span of a year than over a month. Therefore, behavioural reports over a month are the more accurate of the two. The use of different time spans across or within studies may lead to inconsistent or even misleading results.

Accurate and reliable measurements of health behaviours, especially drug use and sexual activity, have proven to be difficult because of social desirability influences. People underreport smoking and underestimate alcohol consumption. Self-reports of alcohol consumption can account for as little as half the amount obtained from sales figures. Likewise, the total number of cigarettes sold or otherwise estimated to be consumed is substantially higher than the estimate calculated from smokers' self-reports. In addition, studies that focus on behavioural frequencies consistently yield illusory superiority: respondents report a lower frequency of unhealthy behaviours and higher frequency of healthy behaviours for themselves than for an average peer. Illicit problem behaviours, such as drug or alcohol use, may elicit stronger self-serving biases than more mundane health-threatening behaviours in adolescents (for details, see Renner, in press).

## References

- Conner, M. & Norman, P. (Eds.) (1996). *Predicting Health Behaviour: Research and Practice With Social Cognition Models*. Buckingham, England: Open University Press.
- Renner, B. (2001, in press). Assessment of health behaviours. In Smelser, N.J. & Baltes, P.B. (Eds.), *The International Encyclopedia of the Social and Behavioural Sciences*. Oxford, England: Elsevier.
- Schwarzer, R. & Gutiérrez-Doña, B. (2000). Health Psychology. In Pawlik, K. & Rosenzweig, M.R. *International Handbook of Psychology* (pp. 452–465). London: Sage.
- Schwarzer, R. & Renner, B. (2000). Social-cognitive predictors of health behaviour: action self-efficacy and coping self-efficacy. *Health Psychology, 19*(5), 487–495.
- Schwarz, N. & Strack, F. (1991). Context effects in attitude surveys: applying cognitive theory to social research. In Stroebe, W. & Hewstone, M. (Eds.), *European Review of Social Psychology, Vol. 2*, (pp. 31–50). Chichester, England: Wiley.
- Sobell, L.C. & Sobell, M.B. (1995). Alcohol consumption measures. In Allen, J.P. & Columbus, M. (Eds.), *Assessing Alcohol Problems* (pp. 55–73). NIAAA Treatment Handbook Series 4. Bethesda, MD: NIH.
- Stone, A.A. & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioural medicine. *Annals of Behavioural Medicine, 16*, 199–202.

Britta Renner and Ralf Schwarzer

## Related Entries

HEALTH, QUALITY OF LIFE, INTERVIEW IN BEHAVIOURAL CLINICAL AND HEALTH SETTINGS, BRAIN ACTIVITY MEASUREMENT, GOAL ATTAINMENT SCALING, PSYCHOPHYSIOLOGICAL EQUIPMENT AND MEASUREMENTS, OUTCOME ASSESSMENT/OUTCOME TREATMENT.

# A APPLIED FIELDS: NEUROPSYCHOLOGY

## INTRODUCTION

Neuropsychological assessment as a formal procedure is a relatively recent development. Its evolution has paralleled advances, in the past

fifty years, in the areas of neuroscience in general, and cognitive neuroscience in particular. It has also been influenced by developments in applied clinical disciplines such as neurology, neuroradiology, rehabilitation medicine, special



education, geriatrics, developmental psychology, etc. In this section, we review the historical trajectory of this aspect of clinical neuropsychology, and present the current state of the field.

## HISTORICAL ANTECEDENTS

Neuropsychological assessment did not come of age until after the Second World War. In the second half of the 19th century, there had been a flurry of clinical studies that correlated brain structures and cognitive activity. The work of Broca, Déjerine, Jastrowitz, Korsakoff, Lichteim, Liepmann, Oppenheimer, Ribot, Wernicke, and many others in the latter part of the 19th century described the neurological substrates of disorders such as the aphasias, apraxias, amnesia, and frontal disinhibition (Walsh, 1978; Benton, 2000). However, these advances in localization of function lay dormant (except in the USSR) for over half a century. This approach regained its popularity in the 1950s and 1960s, in part as a result of the work of Brenda Milner and her colleagues in Montreal, who described the pivotal role of the hippocampus in memory (Scoville & Milner, 1957), and in part due to the work of Benton, Zangwill, Hécaen, Ajurriaguera, and Goodglass. Sperry's work and the seminal case study of a human disconnection syndrome (Geschwind & Kaplan, 1962) lent further impetus to the belief that higher cognitive functions could be compartmentalized and subjected to analysis via objective techniques. Interest in the pioneering 19th century studies and their potential contribution to the study of brain-behaviour relationships was revived by Norman Geschwind in Boston at approximately the same time (Geschwind, 1997).

## PARADIGMS IN NEUROPSYCHOLOGICAL ASSESSMENT

### Global Measures of Brain Damage

At the outset, the primary goal of the neuropsychological evaluation in the United States was to assist in differentiating behavioural disorders of 'organic' (i.e. structural) nature, from those

of 'functional' (i.e. psychological) origin. This focus can be attributed to the influence of psychoanalytic thinking, which postulated that psychiatric disturbance could result from intrapsychic (moral and psychological) and disturbed inter-personal relationships (Hill, 1978: vii). Further, clinicians in the USA and Britain were formed in a positivist, psychometric culture, which has more readily trusted an actuarial, mechanistic approach to data gathering, and statistically driven decision-making algorithms (Meehl, 1954), while being less comfortable with the methodology of single-case studies. Thus Ward Halstead's purpose in designing tests was to determine whether a person had sustained brain damage or not; asking, 'more practically, can convenient indices be found which, like blood pressure, accurately reflect the normal and pathological range of variance for the individual? Is there a pathology of biological intelligence which is of significance to psychiatry and to our understanding of normal behaviour?' (Halstead, 1947: 7). He noted accurately that the tests developed by Binet and standardized by Terman (for the purpose of identifying 'sub-normal' children who required remediation in school) were completely insensitive to the effects of brain damage. Citing the work of Hebb and Penfield (1940) he wrote, 'Evidence is now on record to the effect that surgical removal of one or both prefrontal lobes – that is, a mass of brain substance constituting about one-fourth of the total cerebrum – may not significantly alter the I.Q.' (Halstead, 1947: 7).

### Fixed and Flexible Batteries

The Halstead-Reitan Battery (Reitan & Davison, 1974; Reitan & Wolfson, 1993) and extensions of it (e.g. Heaton, Grant & Mathews, 1991) gained widespread recognition in the USA from the 1950s as the best practice in neuropsychological assessment, since it provided a means of summarizing an array of observations into numerical values that can be compared across patients and situations and, which provide reliable predictions (Boll, 1981; Russell, 1986). This battery (the Halstead-Reitan Battery; Reitan & Davison, 1974) began as a selection of seven tests chosen for their ability to best discriminate patients with frontal versus non-frontal or non-injured controls. Currently,

five of the original seven tests are typically administered to derive an Impairment Index (the proportion of scores in the impaired range), together with the Wechsler Intelligence scales, memory tests, and other tests of specific functions (Lezak, 1995: 709). The five tests include the Categories Test, the Tactile Perceptual Test, the Seashore Rhythm Test, the Finger Oscillation Test, and the Speech-Sounds Perception Test.

Halstead was fully aware of the view that prevailed in the 1930's (and well into the 1960's) that brain dysfunction is unitary (i.e. the notion of equipotentiality). Other tests sensitive to 'brain damage' were also available at that time. A well-known example is the Visual Motor Gestalt Test (Bender, 1938), now commonly referred to as the Bender-Gestalt test. Piotrowski might be credited with developing the first 'impairment index' when he stated (in reference to interpretation of responses to the Rorschach ink blot test) that, 'No single sign alone points to abnormality in the psychiatric sense, to say nothing of organic involvement of the brain. It is the accumulation of abnormal signs in the record that points to abnormality' (Piotrowski, 1937, cited in Lezak, 1995: 773). He considered five signs (out of the ten that he proposed) to be the minimal number needed to support an inference of brain damage, and noted that the number of signs increased with age. Halstead insisted on 'blind' administration of tests by trained technicians to ensure objectivity of results, although his qualitative observations were based on an impressive variety of sources. The use of cut-off scores (usually one and a half or two standard deviations from the mean, indicating impairment) and an Impairment Index (the number or proportion of tests on which the patient's score equals or exceeds the cut-off) as applied to the Halstead-Reitan battery (Reitan & Davison, 1974) attests to the influence of then prevalent theories of brain function on neuropsychological test interpretation. Nonetheless, both Halstead and later Reitan rejected the notion that brain function is unitary, based on the fact that patients with lesions in different areas produced different patterns of performance on the tests (Halstead, 1947; Reitan & Davison, 1974). Over time, there was recognition that identifiable neurological syndromes exist, and rather than apply a fixed battery of tests to everyone, regardless of the diagnosis, a flexible

battery approach, espoused by Benton, in which standardized tests are selected to assess the functions most likely to be affected by the presenting conditions, has come to be preferred by the majority of clinicians in the United States.

### Alternatives to the Psychometric Approach

The psychometric approach has not gone unchallenged. One of the pillars in the area of assessment in the USA, Anne Anastasi, expressed early concerns about the indiscriminate use of standardized assessment with diverse populations (Anastasi & Cordova, 1953). Further, the essential tenet of this approach is that 'the final solution to a problem, arrived at within a given time, is an objective measure of an underlying cognitive mechanism' (Kaplan, 1988: 129). A number of people have taken issue with such a premise, pointing to the multifactorial nature of the tasks used for assessment, and the various routes that an individual can take to reach a solution (e.g. Luria, 1966; Walsh, 1978; Kaplan, 1988). The score-based approach to assessment is quite different from an attempt to understand brain-behaviour relationships in terms of the way in which the organism or person interacts with the environment to attain a goal, regardless of the integrity of the nervous system. As early as the mid-1920s, Luria and his mentor Vygotsky in the USSR had decided that the best approach to understanding higher cognitive functions was two-pronged: to study their normal development on the one hand, and their 'decomposition' in brain-damaged individuals on the other. Vygotsky felt that the earlier work of the 19th century neurologists was limited by the absence of an adequate theory of psychology (Luria, 1979). Luria and his followers emphasized an analysis of performance based on the belief that behaviours are the result of functional brain systems that interact with each other. Thus, a function could be subserved by various sub-systems, and difficulty in performing a task could be the result of a breakdown in any of those mechanisms. Conversely, compensatory routes engaging alternate subsystems can sometimes be utilized to achieve the same goal. This approach was particularly relevant to the rehabilitation of individuals who sustained

brain damage during World War II. Analysis of the compensatory strategies that are or can be brought into play to reach a goal (that is, an analysis of the different circumstances that elicit or inhibit a given behaviour) provides a basis for intervention that can enhance the individual's success. Largely for this reason, Luria's approach to neuropsychological assessment has been widely adopted in rehabilitation centres throughout the world (e.g. Caetano & Christensen, 1997). His work has had a wide-ranging impact in neuropsychological practices and assessment in many countries.

### **Evolving Procedures and Roles for the Neuropsychologist**

Christensen (1979) attempted to systematize Luria's approach to testing in order to make his procedures more accessible to a wider audience and to present stimuli in a format and sequence consistent with Luria's conceptualization of cortical functions. In the United States, this approach was assimilated within a quantitative scoring framework by Golden and his colleagues, and is now known as the Luria-Nebraska Neuropsychological Battery (Golden, Purisch, & Hammeke, 1981). This battery is rarely in use today, as it has been widely criticized on a number of both conceptual and methodological grounds (Lezak, 1995). The publication in 1976 of Lezak's 'Neuropsychological Assessment' (now in its 3rd edition), which describes and reviews many tests, as well as syndromes, provided an important resource to the field. One of the legacies of Luria's conceptualization of a hierarchy of cognitive abilities has been the need to separate the impact of primary on secondary functions (e.g. the need to assess activation and attention as they relate to memory and other higher mental processes). An important distinction must be made, especially in clinical practice, between psychometric testing (which in many clinics is performed by technicians) and neuropsychological assessment (which involves the interpretation and integration of information regarding the patient). A comprehensive neuropsychological evaluation will, at a minimum, address basic attentional, linguistic, visuo-perceptual and visuo-constructional, motor, learning and memory, calculations, sequencing, executive and emotional functions, social interactions, and

problem-solving abilities. The importance of reviewing the records, obtaining a comprehensive history, family interviews, and an analysis of the person's goals and behaviour across different settings and over time, provide a more contextualized understanding of the individual as a whole, and better insights into how recommendations can be realistically formulated (Armengol, Kaplan, & Moes, 2001). Attention to the role and possible impact on testing of medication, pain, physical limitations, and mental status (including neurovegetative functions such as sleep, appetite, sensorimotor changes, and mood) is essential.

Technological breakthroughs in the field of neuroimaging, specifically the advent of the CT scan in the early 1970s, and more recently with technologies that allow visualization of areas of brain activation (such as functional Magnetic Resonance Imaging (fMRI) and PET/SPECT scanning), along with the availability of more sophisticated neuropsychological evaluation procedures in clinical settings, has gradually changed the focus and role of neuropsychological assessments. No longer is lesion localization the primary aim; rather, it has shifted in the direction of describing and understanding the functional consequences and rehabilitation implications of brain dysfunction. An important exception to this in the USA has been the area of forensic neuropsychology, where the focus continues to be on establishing the presence of structural brain damage following injury, with its functional and prognostic implications. This is particularly a concern in cases of minor head injury, where neuroimaging is likely to be unhelpful and where the potential for malingering is inevitably raised. This has led to interest in measures designed to detect deception (if only to be able to preemptively refute the assertion of malingering in the majority of cases), as well as an appreciation for the need to take into account the baseline incidence in the normal population of symptoms and patterns of test scores, in order to be able to establish the presence or absence of pathology.

In light of relatively new standards for presenting evidence in courtrooms (i.e. the Daubert rule of 1993), clinical neuropsychologists have had to rely on standardized instruments (rather than clinical experimental techniques) to document changes in functioning.

Over the years, within the experimental tradition of cognitive psychology, investigations of selective deficits in individuals with brain lesions led to the identification of discrete components of complex functions, as well as the development of ingenious and elegant laboratory procedures to demonstrate disconnections, levels of processing, and double dissociation of functions (e.g. Warrington, 1982; Shallice, 1988; McCarthy & Warrington, 1990; Gazzaniga, 1995). Experimental paradigms that have been used with lesioned non-human animals have also been applied in research and clinical settings to see if brain-behaviour relationships established for other species can be successfully applied to the study of humans. A good example is the use of delayed object alternation tasks with individuals who have sustained prefrontal damage (e.g. Oscar-Berman, McNamara, & Freedman, 1991).

### Current Trends

Edith Kaplan, who was trained by the developmental psychologist Heinz Werner, has formulated and championed a process approach to neuropsychological assessment. 'For Werner (1956) every cognitive act involves "microgenesis" (i.e. an "unfolding process over time"). Thus close observation and careful monitoring of behaviour en route to a solution (process) is more likely to provide more useful information than can be obtained from right or wrong scoring of final products (achievement)' (Kaplan, 1988). The Boston Process Approach, as it is known, attempts to bridge the case study method (grounded in an understanding of neuropsychological syndromes) developed by Luria on the one hand, and the focus on the need for replicable, empirical, and normatively standardized data, on the other. This has been pursued in several ways. Following up on developments in cognitive neuroscience, new tests such as the California Verbal Learning Test (Delis et al., 1987) and the Delis-Kaplan Executive Function System (Delis, Kaplan & Kramer, 2001), were developed to better assess aspects of learning and executive function which are found to differ among patients with different neuropsychological disorders. This approach has also included (a) the addition of standardized procedures to existing tests to assist in clarifying the process underlying a patient's response (e.g. the Wechsler

Adult Intelligence Scale as a Neuropsychological Instrument or WAIS-RNI, and the Wechsler Intelligence Scale for Children as a Process Instrument or WISC-III PI); (b) the addition of new indices to score existing data that allow for better capture of relevant process variables (e.g. new methods to score the Rey-Osterrieth Complex Figure drawings, as developed by Stern et al., 1995); and (c) a conceptual reanalysis of performance on existing tests based on alternative theoretical models (see Poreh 2000, for examples of this last approach). Poreh (2000) refers to this new trend as the 'Quantified Process Approach'. One of the potential advantages of computerized approaches to assessment is the ability to capture sequential qualitative aspects of performance, although this potential remains largely unfulfilled at this time.

### FUTURE PERSPECTIVES AND CONCLUSIONS

Neuropsychological assessment is central to attempts to understand the biological bases of behaviour. Even as our technology becomes more sophisticated and we unravel genetic codes, behavioural functions must be mapped, and behavioural and cognitive markers for particular syndromes and disorders become more relevant. Structural and functional in vivo neuroimaging techniques provide exciting opportunities to examine patterns of brain activation during the performance of tests and induced psychological states. Neuropsychological assessment must keep pace with the new demands imposed by technological advances and limitations. Tests adapted for presentation during fMRI are good examples of the latter (e.g. Whalen et al., 1996). In the immediate future, the greater use of computerized technologies will open possibilities for more naturalistic assessment, the evaluation of more complex behaviours, and the ability to collect a wide sample of measures, including the incorporation of physiological measures, concomitantly with performance of various activities. One area with particular promise for assessment and rehabilitation is the developing field of virtual reality (Riva, 1997). Neuroimaging has also permitted an analysis of brain functioning in individuals who differ in terms of the ecological

demands placed upon them, such as illiterates and bilingual subjects (e.g. Castro-Caldas et al., 1998; Perani et al., 1998). The finding that structural and functional differences emerge under different environmental circumstances reinforces the need to take into account issues relating to ecological validity. That is, tests that have been developed for one population may have limited validity when administered to a different population (this certainly applies to populations in different stages or trajectories of development). Similarly, results that are obtained under one set of circumstances (e.g. the clinic or research laboratory) may not generalize to other, more typical daily tasks and situations. Clearly there is much work to be done in this area.

## References

- Anastasi, A. & Cordova (1953). Some effects of bilingualism upon the intelligence test performance of Puerto Rican children in New York. *Journal of Educational Psychology*, 44, 1–19.
- Armengol, C.G., Kaplan, E. & Moes, E.J. (2001). *The Consumer-Oriented Neuropsychological Report*. Odessa, FL: Psychological Assessment Resources.
- Benton, A. (2000). Historical aspects of cerebral localization. In Riva, D. & Benton, A. (Eds.), *Localization of Brain Lesions and Developmental Functions* (pp. 1–14). London, England: John Libbey.
- Bender, L. (1938). A visual motor gestalt test and its clinical use. American Orthopsychiatric Association, Research Monographs, No. 3.
- Boll, T. (1991). The Halstead-Reitan Neuropsychological Battery. In Filskov, S.B. & Boll, T.J. (Eds.), *Handbook of Clinical Neuropsychology*. New York: Wiley-Interscience.
- Caetano, C. & Christensen, A.L. (1997). The design of neuropsychological rehabilitation: the role of neuropsychological assessment. In Leon-Carillon, J. (Ed.), *Neuropsychological Rehabilitation: Fundamentals, Innovations, and Directions*. Delray Beach, FL.: St. Lucie Press.
- Castro-Caldas, A., Petersson, K.M., Reis, A., Stone-Elander, S. & Ingvar, M. (1998). The illiterate brain. Learning to read and write during childhood influences the functional organization of the adult brain. *Brain*, 121, 1053–1063.
- Christensen, A.-L. (1978). *Luria's neuropsychological investigation* (2nd ed.). Copenhagen: Munksgaard.
- Delis, D.C., Kramer, J.H., Kaplan, E. & Ober, B.A. (1987). *The California Verbal Learning Test Manual*. San Antonio: Psychological Corporation.
- Delis, D.C., Kaplan, E. & Kramer, J.H. (2001). *Delis-Kaplan Executive Function System*. San Antonio, TX: The Psychological Corporation.
- Gazzaniga, M.S. (1995). *The Cognitive Neurosciences*. Cambridge: Massachusetts Institute of Technology.
- Geschwind, N. (1997). Selected Writings. In Devinsky, D.O. & Schachter, S.C. (Eds.), *Norman Geschwind: Selected Publications on Language, Epilepsy, and Behaviour*. Boston: Butterworth-Heinemann.
- Geschwind, N. & Kaplan, E. (1962). A human cerebral disconnection syndrome. *Neurology*, 12, 675–685.
- Golden, C.J., Purisch, A.D. & Hammeke, T.A. (1985). *Luria-Nebraska Neuropsychological Battery: Forms I and II*. Los Angeles, CA: Western Psychological Services.
- Halstead, W.C. (1947). *Brain and Intelligence: A Quantitative Study of the Frontal Lobes*. Chicago, IL: The University of Chicago Press.
- Heaton, R.K., Grant, I. & Mathews, C.G. (1991). *Comprehensive Norms for an Expanded Halstead-Reitan Battery: Demographic Corrections, Research Findings, and Clinical Applications*. Odessa, FL: Psychological Assessment Resources.
- Hill, D. (1978) *Forward to the First Edition of Lishman, W.I, Organic Psychiatry; The Psychological Consequences of Cerebral Disorder*. Oxford, England: Blackwell Scientific Publications.
- Kaplan, E. (1988). A process approach to neuropsychological assessment. In Boll, T. & Bryant, B.K. (Eds.), *Clinical Neuropsychology and Brain Function: Research, Measurement, and Practice* (pp. 129–167). Washington, DC: American Psychological Association.
- Lezak, M.D. (1995). *Neuropsychological Assessment* (3rd ed.). New York: Oxford University Press.
- Lishman, W.A. (1977). *Organic psychiatry*. Oxford, England: Blackwell Science.
- Luria, A.R. (1966). *Higher Cortical Functions in Man*. New York: Basic Books.
- Luria, A.R. (1979). The making of mind. In Cole, M. & Cole S. (Eds.), *The Making of Mind: A Personal Account of Soviet Psychology*. Cambridge: MIT Press.
- McCarthy, R.A. & Warrington, E.K. (1990). *Cognitive Neuropsychology: A Clinical Introduction*. San Diego, CA: Academic Press.
- Meehl, P.E. (1954). *Clinical versus Statistical Prediction*. Minneapolis: University of Minnesota Press.
- Oscar-Berman, M., McNamara, P. & Freedman, M. (1991). Delayed response tasks: parallels between experimental ablation studies and findings in patients with frontal lesions. In Levin, H.S., Eisenberg, H.M. & Benton, A.L. (Eds.), *Frontal Lobe Function and Dysfunction*. New York: Oxford University Press.
- Perani, D., Paulesu, E., Sebastian Galles, N., Dupoux, E., Dehaene, S., Battinardi, V., Cappa, S., Fazio, F. & Mehler, J. (1998). *Brain*, 121, 1841–1852.
- Poreh, A. (2000). The quantified process approach: an emerging methodology to neuropsychological assessment. *The Clinical Neuropsychologist*, 14, 212–222.
- Reitan R.M. & Davison, L.A. (1974). *Clinical Neuropsychology: Current Status and Applications*. Washington, D.C.: V.H. Winston & Sons, Inc.
- Reitan R.M. & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and*

- Clinical Interpretation*. Tucson, AZ: Neuropsychology Press.
- Riva, G. (1997). Virtual reality in neuro-psychophysiology: cognitive, clinical and methodological issues in assessment and treatment. *Studies in Health Technology and Informatics*, Vol. 44. Amsterdam: IOS Press.
- Russel, E.W. (1986). The psychometric foundation of clinical neuropsychology. In Filskov, S.B. & Boll, T.J. (Eds.), *Handbook of Clinical Neuropsychology*. New York: John Wiley & Sons.
- Shallice, T. (1988). *From Neuropsychology to Mental Structure*. New York: Cambridge University Press.
- Scoville, W.B. & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20, 11–21.
- Stern, R.A., Singer, E.A., Duke, L.M., Singer, N.G., Morey, C.E., Daugherty, E.W. & Kaplan, E. (1995). The Boston qualitative scoring system for the Rey-Osterrieth complex figure: description and inter-rater reliability. *Clinical Neuropsychologist*, 3, 309–322.
- Walsh, K. (1978). *Neuropsychological Assessment: A Clinical Approach*. New York: Churchill Livingstone.
- Warrington, E., (1982). The fractionation of arithmetical skills: a single case study. *Quarterly Journal of Experimental Psychology*, 34A, 31–51.
- Whalen, P.J., Bush, G., McNally R.J., Sabine, W., McInerney S.C., Jenike, M.E. & Rauch, S.L. (1998). The emotional counting stroop paradigm: a functional magnetic resonance imaging probe of the anterior cingulate affective division. *Biological Psychiatry*, 44(12), 1219–1228.

Carmen Armengol de la Miyar,  
Elisabeth J. Moes and Edith Kaplan

### Related Entries

MEMORY DISORDERS, VISUO-PERCEPTUAL IMPAIRMENTS, EXECUTIVE FUNCTIONS, VOLUNTARY MOVEMENTS, TRAUMATIC BRAIN INJURY, EQUIPMENT FOR ASSESSING BASIC PROCESSES, NEUROPSYCHOLOGICAL TEST BATTERIES, OUTCOME MEASURES IN NEUROPSYCHOLOGICAL AND BRAIN INJURY REHABILITATION.

## A APPLIED FIELDS: ORGANIZATIONS

### INTRODUCTION

Psychologists interested in describing, diagnosing or changing organizational behaviour are compelled to assess psychological properties of organizations at some stage of their work. It is for this reason that, as in other applied fields, multiple approaches and techniques concerning psychological assessment have been developed and used in organizations. The present article aims to describe a multilevel psychological assessment, adopting a social systems perspective. To this end, we define psychological assessment of organizations, analyse how it is implemented at different levels, and present future perspectives.

### CONCEPT AND OBJECTIVES OF ORGANIZATIONAL ASSESSMENT

Psychological assessment of organizations refers to the measure of human behaviour in organizations

using scientific instruments. The primary objective of this assessment is to describe the organization as an individual and collective behaviour system accurately. However, psychological measures should be also relevant in terms of practical implications, serving the purpose of helping managers and other members of the organizations to make decisions.

Traditionally, psychological assessment in organizations has been restricted to the measure of individual differences, implicitly assuming that organizational effectiveness is the result of the aggregation of the psychological characteristics of individuals. This individual level of analyses, however, is limited and the measurement of the work group and the organization as a whole offer a complementary and more comprehensive assessment. Psychological properties exist at different levels of analyses and all of them contribute to the effectiveness. Thus, a multilevel assessment is needed in obtaining a deeper description of the organizations.

## MAIN TOPICS IN PSYCHOLOGICAL ASSESSMENT AT DIFFERENT LEVELS OF ANALYSES

### The Individual Level

There is a persistent interest in the study of individual experiences in organizations and continuously there are emerging topics and controversies (Nord & Fox, 1996). Personality, cognitive, affective, and behavioural variables have been assessed during decades. With this in mind, the most relevant issues currently associated with the measurement of individuals in organizations are summarized in this section.

#### *Personality*

Individuals can be characterized by a number of enduring dispositional properties, which help to understand people's behaviours in Organizations. One of the most popular methods of assessing personality is derived from the big five theory. Through self-report inventories, five dimensions of personality are measured: (1) extraversion; (2) emotional stability; (3) agreeableness; (4) conscientiousness; and (5) openness to experience. Several authors prefer the use of a composite of several big five constructs, labelled integrity test, because this broader measure can be more reliable in predicting overall job performance. However, narrower trait constructs can show better prediction of specific job performance criteria within specific occupations (Gatewood, Perloff, & Perloff, 2000).

#### *Knowledge, Skill, and Abilities (KASs)*

KASs are defined, respectively, as the amount of factual information known by an individual, his/her conduct of job specific activities, and his/her conduct of generalized job activities. With respect to the abilities, different goals are associated with the measure of general mental ability or 'g' versus specific abilities. Although there is some consensus about the predictive efficiency of the 'g' factor, measures of specific abilities tend to be more useful when the goals are understanding people's behaviours or their classification. Given that abilities, as they are measured by aptitude tests, refer to a wide and general range of human experiences, more circumscribed measures of skills and knowledge

have been developed in order to improve the validity of measures. This is the case of interpersonal skills, which are especially critical in customer service jobs, work groups, and leadership. Also, job knowledge and tacit-knowledge measures are closely related to specific job performance criteria. For instance, subjects can be exposed to a job-related situation, and their capabilities to solve problem situations can be measured through assessment centre procedures.

#### *Individual Performance*

Production (e.g. quantity) and other employee behaviour records (e.g. absenteeism) are used as objective direct or indirect measures of individual performance. Also, subjective evaluations from individuals familiar with the work of the focal person are considered (e.g. 360° feedback). These performance indicators are the result of task and contextual performance. The first is defined as the proficiency with which subjects perform core technical activities of well-defined jobs. Thus, cognitive abilities are relevant for predicting task performance. In contrast, contextual performance is defined as extra-task proficiency that contributes more to the organizational goals, including aspects such as enthusiasm and volunteering to make duties not formally part of one's job. It is assumed that personality variables are critical for predicting contextual performance criteria (Arvey & Murphy, 1998).

#### *Work Attitudes*

Work attitudes are defined as positive or negative evaluations about aspects of one's work environment (O'Reilly, 1991). The most common constructs measured by attitude instruments are job satisfaction, commitment, involvement, and stress. Satisfaction refers to a emotional state resulting from job experiences. The questionnaires used to measure job satisfaction can be classified in two groups: measures of overall satisfaction and measures of satisfaction with specific aspects of the job (Peiró & Prieto, 1996). The most frequently measured facets are satisfaction with pay, promotion, supervision, and job content (Gatewood et al., 2000). With regard to commitment, there is no generally accepted definition and measurement. While affective commitment measures include aspects such as loyalty towards the organization,

the effort to achieve organizational goals, and the acceptance of organization's values, continuance commitment measures are related to the personal sacrifice associated with leaving the organization and the perceived employment alternatives. Finally, another measure of work attitudes refers to the degree to which the job experiences are perceived as stressful. However, caution is needed because self-report measures of stress may be easily inflated by the person's disposition toward negative affectivity (O'Reilly, 1991).

### The Group Level

The work group consists of individuals who see themselves and who are seen by others as a social entity, who are interdependent because of the tasks they perform as members of a group, who are embedded in one or more larger social systems (e.g. organization), and who perform tasks that affect others (Guzzo & Dickson, 1996). Psychological assessment at group level is primarily focused on three aspects: design, processes, and performance.

### Group Design

Although a good group design cannot guarantee a satisfactory group functioning, it is necessary to facilitate competent group behaviours. It is for this reason that group design should be measured and controlled. Of the different facets of group design (structure of task, group composition, and establishment of norms), composition of group has received increasing attention, especially heterogeneity (Guzzo & Dickson, 1996). Group heterogeneity refers to the mix of abilities, personalities, gender, attitudes, background, and demographic characteristics. In order to work effectively, a 'right mix' of group members is needed. Efforts have been devoted to assess the right mix of members in terms of abilities and personality (West & Allen, 1997). It is the case of 'skill mix', particularly popular within teams in health service setting, which refers to the efficient balance between trained and untrained, qualified and unqualified, and supervisory and operative staff. Also, personality compatibility can be measured. For instance, according to the Schutz's theory of fundamental interpersonal relations orientations (FIRO) there are three basic needs expressed in group interaction: needs for inclusion, control, and affection. A compatible balance of initiators and receivers of control,

inclusion, and affection characterize effective groups.

### Group Process

It is generally assumed that, in addition to group design, the process of interaction among group members affects the effectiveness of the group as a whole. As Hackman (1987) pointed out, assessing group process can pursue different goals. A trained observer can focus on the interpersonal transactions that reflect conscious and unconscious social and emotional forces (e.g. who is talking with whom). Group process assessment can also be focused on the issues of interaction directly related to work of group on its task (e.g. the degree to which knowledge and skill members are used). Group interaction can result in 'synergy'; that is, outcomes that are different from those that would be obtained by simply adding up the contributions of individuals (Hackman, 1987). Synergy can be positive (e.g. a very creative solution of a job-related problem) or negative (e.g. a severe failure of coordination). In general, different methods can be used to assess group process. It is the case of some assessment centre techniques (e.g. simulation), where real job tasks are represented and a group of individuals is assessed by a group of judges.

### Group Performance

Three criteria are typically used to measure group performance: (1) group-produced outputs, (2) the influences of group for its members, and (3) the state of the group as a performing unit (see Guzzo & Dickson, 1996; Hackman, 1987). Although some objective indicators of group outputs can be measured (e.g. quantity), objective criteria are only available for a restricted number of work groups in organizations. In general, the assessment from others (e.g. a manager) is more critically associated with the consequences for a group and its members than objective measures. It is for this reason that there is a tendency to assess outputs in terms of satisfaction of the standards of the people ('clients') who receive and/or review the output. The second measure is related to the impact of group on individuals. It is assumed that the cost of generating group outputs is high if its members are dissatisfied. Accordingly, the degree to which the group



experiences satisfy the needs of group members should be also assessed. The third measure reflects the probability that a group performs effectively in the future. Although the present outputs of a group can be satisfactory, it is possible that the social processes by which these outputs are obtained hamper the group as a performing unit, and its members are not willing anymore to work together on future tasks.

### The Organization Level

Individuals and work groups are embedded in a more general organizational system that can be measured itself. Psychological properties of organizations as whole, such as culture, climate, and performance can also be assessed.

#### *Culture and Climate*

Although culture and climate have been sometimes used as synonyms, they refer to different concepts. As Schneider (1985) pointed out, culture is a deeper construct than climate has been. While organizational climate is defined as the shared perceptions of employees related to the practices, procedures, and behaviours that are rewarded and supported in an organization, culture refers to the beliefs, norms, and values underlying the policies and activities, as well as the manner in which the norms and values systems are communicated and transmitted. Consequently, the modes by which culture and climate are assessed are also different. Culture is usually measured by using qualitative and case study methodologies. In contrast, the survey approach is the dominant method in measuring climate (Schneider, 1985).

#### *Organizational Performance*

Financial performance and productivity are considered as the typical measures of organizational performance as a whole. In addition, other measures associated with customer responses of satisfaction and perceived quality have received increasing attention. While economic measures of performance reflect quantity of outputs, psychological measures of customer evaluations refer to quality of outputs as they are perceived by the customer (Fornell, 1992). Psychological measures offer information that is not included in current-term financial measures (Aaker & Jacobson,

1994). In the absence of alternatives, short-term financial gains are usually used as indicators of long-term prospects. However, the strategies devoted to increase long-term performance often diminish short-term earnings. The myopic management style, focused on short-term gains, can be corrected by considering non-financial measures. In fact, the measurement of customer perceptions of product quality is able to predict information concerning long-term competitiveness that is not captured by short-term financial measures (see Aaker & Jacobson, 1994).

### FUTURE PERSPECTIVES

#### **An Integrated Assessment of Organizations**

In the preceding discussion, we have analysed how the psychological assessment is implemented at different levels of organizations. However, a more integrated perspective can be considered where the different levels of analyses are interrelated showing complex interactions. Herriot and Anderson (1997) proposed that the relationships between measures at individual, group, and organizational levels of analyses could show three kinds of patterns: complementary, neutral, and contradictory. The complementary interaction is observed when a high score at one level of analysis is desirable in combination with a high score at another level (e.g. when high interpersonal skills are required for both individual work and group working). The neutral interaction occurs when a high score on a construct is desired at one level and, simultaneously, it is not applicable at another level of analysis (e.g. when interpersonal skills are required for group working, but they are not related to individual performance). Finally, the contradictory interaction is observed when a high score at one level of analyses is desirable in combination with a low score at another level (e.g. when extraversion is desirable for team working, but introversion is positively related to individual performance). Because of its relevance to research and management, future efforts are needed in developing and testing these kinds of approaches. An integrated assessment is able to describe an organization more accurately, given that it serves to diagnose their complex and contradictory character.

### Links Between the Context of Organizations and Psychological Assessment

It is generally assumed by managerial orientations that organizations are free in order to design and implement practices and policies (see Morishima, 1995). However, the external context of organization impacts on the organizational choices, including the type of procedures and techniques used in the psychological assessment. For instance, Rousseau and Tinsley (1997) suggested that the culture of a country (e.g. in terms of individualism vs. collectivism) can be related to the appropriateness of individual versus group measures of performance, as well as to the emphasis on individual-job versus individual-organization or work-group fit measures. Also, Herriot and Anderson (1997) indicated that organizations are now subjected to an environment that changes with an increasing speed and unpredictability. In this context, organizations emphasize the psychological assessment related to employee flexibility, personality, and potential to innovate. Additionally, it is also likely to expect that, in some circumstances, organizations impact on their external context. For instance, organizations can demand an education system in which certifications are highly job-related, given that this type of education can facilitate the measurement and the managerial decisions (e.g. in a selection process). Thus, reciprocal influences between organisations and their contexts can be studied in the future. A contingency approach can be proposed where the psychological assessment depends on the characteristics of external contexts and the nature of the relations between these contexts and organizations.

### The Political Face of Psychological Assessment in Organizations

Research and practice in organizations espouses a rational perspective in understanding psychological assessment. Organizations are often defined as rational and efficiency-seeking systems, and managers use psychological assessment in order to achieve valued organizational outcomes. However, their political 'face' should also be considered. Following this perspective, the organization is seen as a political system with competing groups and interests, each with its own perceptions of

organizational realities. The political face is not everything, but it serves to understand some events related to psychological assessment. Additionally, the ignorance of power in organizations can result in managerial failures and incomplete assessment at different levels of analyses. For instance, there is not only a dominant culture in organizations but also 'countercultures' that reflect alternative values. Usually, individuals and work groups that have values and perceptions congruent with those of organizations, especially with the top-management group, also have more power and influence (Friedlander, 1987). Accordingly, it is reasonably to expect that divergent thinking will not be reflected in the measurement of culture. Also, psychological assessment is likely to be used to reinforce and justify the values and perceived tasks of the dominant coalition. Powerful coalitions act within their own reality, which is not necessarily better than other realities constructed within the organization as a whole. Alternative cultures can reflect adaptive values in terms of initiative and creativity. The ignorance of these cultures has contributed to long-term disasters in many companies (see Dachler, 1989). Thus, more effort is needed in order to include the diversity of organizational 'cultures' in psychological assessment, as well as in studying the impact of power forces and power games on measurement decisions at different levels of analyses.

### CONCLUSIONS

A multilevel psychological assessment has important potential benefits. Using this perspective, the great complexity of organizations is diagnosed, given that the organization is considered as an open social system with different measurable subsystems. Psychologists can focus their psychological assessment at different levels of analyses. Thus, this perspective serves to consider both the micro domain's focus on individuals and groups and the macro domain's focus on the organization as a whole.

Additionally, the multilevel psychological assessment is enriched if three complementary perspectives are also incorporated in the future. First, a more integrated assessment can be considered, assuming that constructs measured at different levels of analyses can show complex, even contradictory, relationships. Secondly, there is a need to

study the reciprocal influences, in terms of psychological assessment, between organizations and their external contexts. Finally, the political face of organizations should be measured and analysed in order to obtain a richer portrait of psychological assessment in organizations.

## References

- Aaker, D.A. & Jacobson, R. (1994). The financial information content of perceived quality. *Journal of Marketing Research*, 31, 191–201.
- Arvey, R.D., & Murphy, K.R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, 49, 141–168.
- Dachler, H.P. (1989). Selection and the Organisational context. In Herriot, P. (Ed.), *Handbook of assessment in Organisations* (pp. 45–69). Chichester: John Wiley & Sons.
- Fornell, C. (1992). A national customer satisfaction barometer: the Swedish experience. *Journal of Marketing*, 56, 6–21.
- Friedlander, F. (1987). The ecology of work groups. In Lorsch, J.W. (Ed.), *Handbook of Organisational Behaviour* (pp. 301–314). Englewood Cliffs: Prentice-Hall, Inc.
- Gatewood, R.D., Perloff, R. & Perloff, E. (2000). Testing and industrial application. In Goldstein, G. & Hersen, M. (Eds.), *Handbook of Psychological Assessment* (pp. 505–525). Oxford: Elsevier Science Ltd.
- Guzzo, R.A. & Dickson, M.W. (1996). Teams in Organisations: recent research on performance and effectiveness. *Annual Review of Psychology*, 47, 307–338.
- Hackman, J.R. (1987). The design of work teams. In Lorsch, J.W. (Ed.), *Handbook of Organisational Behaviour* (pp. 315–342). Englewood Cliffs: Prentice-Hall, Inc.
- Herriot, P. & Anderson, N. (1997). Selecting for change: How will personnel and selection psychology survive? In Anderson, N. & Herriot, P. (Eds.), *International Handbook of Selection and Assessment* (pp. 1–34). Chichester: John Wiley & Sons.
- Morishima, M. (1995). Embedding HRM in a social context. *British Journal of Industrial Relations*, 33, 617–640.
- Nord, W.R. & Fox, S. (1996). The individual in organisational studies: the great disappearing act?. In Clegg, S.R., Cynthia, C. & Nord, W.R. (Eds.), *Handbook of Organisation Studies* (pp. 148–174). London: Sage Publications.
- O'Reilly, C.A. (1991). Organisational behaviour: where we've been, where we're going. *Annual Review of Psychology*, 42, 427–458.
- Peiró, J.M. (1985). Psychological Assessment in Organisations. *Evaluación Psicológica-Psychological Assessment*, 1, 189–239.
- Peiró, J.M. & Prieto, F. (Eds.), (1996). *Tratado de Psicología del Trabajo* (2 vols.). Madrid: Síntesis.
- Rousseau, D.M. & Tinsley, C. (1997). Human resources are local: society and social contracts in a global economy. In Anderson, N. & Herriot, P. (Eds.), *International Handbook of Selection and Assessment* (pp. 39–61). Chichester: John Wiley & Sons.
- Schneider, B. (1985). Organisational Behaviour. *Annual Review of Psychology*, 36, 573–611.
- West, M.A. & Allen, N.J. (1997). Selecting for teamwork. In Anderson, N., & Herriot, P. (Eds.), *International Handbook of Selection and Assessment* (pp. 492–506). Chichester: John Wiley & Sons.

José Maria Peiró and Vicente Martínez-Tur

## Related Entries

CULTURE (ORGANIZATIONAL), LEADERSHIP IN ORGANIZATIONAL SETTINGS, OBSERVATIONAL TECHNIQUES IN WORK AND ORGANIZATIONAL SETTINGS, RISK & PREVENTION IN WORK & ORGANIZATIONAL SETTINGS, SELF REPORTS IN WORK AND ORGANIZATIONAL SETTINGS

# A APPLIED FIELDS: PSYCHOPHYSIOLOGY

## INTRODUCTION

The major focus of this entry will be to provide a clear rationale for the application of psychophysiological approaches and methods to areas of applied psychology. We will examine the reasons

for their application, the psychological constructs and processes to be assessed, the methods employed, and specific issues concerning applied uses of these techniques. Specific guidance on psychophysiological recording has been dealt with elsewhere, together with entries on brain

activity and ambulatory monitoring. For background reading and a general reference source, Cacioppo, Tassinary & Bernstron's *Handbook of Psychophysiology*, 2nd Edition (2000) is recommended. Other useful introductory texts include Cacioppo and Tassinary (1990), Hugdahl (1995) and Stern, Ray and Davis (1980).

## DEFINITIONS AND CONSTRUCTS

Psychophysiology can be loosely defined as the study of psychological constructs and processes using non-invasive physiological measures (see Cacioppo, Tassinary & Bernstron, 2000; Turpin, 1989). Traditionally it is distinguished from physiological psychology by emphasizing the importance of studying the intact and conscious organism, usually in the absence of invasive techniques, which might disrupt and limit consciousness or behaviour. As such, the usual domain of psychophysiology has been the measurement of peripheral autonomic and central cortical measures within human participants studied whilst engaged in psychologically relevant tasks or natural situations. In contrast, physiological psychology has tended to use animal subjects and to measure invasively, usually directly from the nervous system, using implanted electrodes, and frequently employing invasive manipulations such as lesioning, infusion of pharmacological agents, direct stimulation etc. More recently, these boundaries have become less distinct since physiological psychology has been incorporated within the greater multidisciplinary arena of neuroscience, and psychophysiology has been extended by more direct but still non-invasive measures of brain activity and structure such as functional imaging, dense array electroencephalography and magnetography (see Cacioppo et al., 2000). Nevertheless the cardinal features of psychophysiology as being the study of psychological processes, largely from human participants and using non-invasive physiological measures, are central to the successful application of the discipline to more applied areas of study.

## APPLIED PSYCHOPHYSIOLOGY

Psychophysiology has always been essentially an applied discipline since its identity has been very

much to do with the measures employed and their various applications. Recently, Cacioppo et al. (2000) described this as systemic psychophysiology, which refers to the study of the various physiological systems (i.e. electrodermal, cardiovascular, cortical etc.) with respect to measurement, quantification and their relationships to psychological processes and paradigms. Much psychophysiological research has been methodologically focused in validating either specific physiological measures or their use as indices of psychological constructs. Subsequently, these measures have then been applied to theoretical questions derived from other branches of psychology including both fundamental and applied research. Traditional areas of application have included psychopathology research and the search for physiological markers of psychological disorder, as well as the development of clinical assessment and outcome measures (Keller, Hicks & Miller, 2000; Stoney & Lentino, 2000; Turpin, 1989). The measurement of stress and cognitive performance using psychophysiological parameters has also meant that these techniques have been used extensively within human factors and ergonomic research (Kraemer, & Weber, 2000). Other applied areas where psychophysiological approaches have been adopted have included attitude measurement, applied developmental psychology, environmental and specific polygraphy (i.e. lie detection) applications (Cacioppo et al., 2000).

What are the benefits of using psychophysiological approaches? The answer lies in the range of psychological constructs and paradigms for which psychophysiological indices or measures have been derived. Cacioppo et al. (2000) in addition to describing 'systemic psychophysiology', also identified 'thematic psychophysiology' which describes topical areas of psychophysiological research. They cited the following examples: cognitive psychophysiology (human information processing and physiological events); social psychophysiology (reciprocal relationships between social systems and physiology); developmental psychophysiology (developmental and ageing processes); clinical psychophysiology (study of disorders); environmental psychophysiology (person-space interactions) and applied psychophysiology (psychophysiological technologies such as biofeedback, lie detection, man-machine instruction etc.). These topics are

exhaustively covered within their handbook. Similarly, we can identify at a more detailed level a myriad of psychological processes and constructs (e.g. attention, attitudes, emotion, memory consolidation) for which there are claimed to be psychophysiological indices or correlates (see Hugdahl, 1995). For example, a class of evoked potential measures of brain activity called the 'P300' is said to be associated with a variety of psychological processes surrounding stimulus evaluation, categorization and context updating (Donchin & Coles, 1988). Similarly, evoked potential Mis-Match-Negativity (Näätänen, 1992), cardiac deceleration (Graham, 1979) and the electrodermal response (Siddle, 1983) have all been associated with the detection of mismatches due to changes in stimulus novelty or significance.

It is apparent that psychophysiological correlates exist for a wide range of psychological constructs. The question, therefore, arises as to what advantages psychophysiological assessments present with respect to performance or self-report measures? It is claimed that psychophysiological measures have the following advantages: they are objective and free of either subjective or observer bias, they are continuous and unobtrusive measures, they can accurately indicate the timing of psychological events, and they may indicate the nature of mechanisms underlying the brain-behaviour relationships under study. Within an applied setting, many of these advantages become even more important. The ability to obtain objective and continuous measures which do not require either self-report or observation means that physiological measures indicating psychological changes in either state or processes may be studied in difficult or inaccessible environments. These could range from space flight to studying arousal processes in married couples during naturalistic social interaction (Gottman & Levenson, 1992). The emphasis on objective versus subjective report also means that data maybe obtained from individuals with communication difficulties either due to cognitive impairment or age and temperament. Indeed, with respect to many psychological processes, it is argued that a comprehensive understanding is not possible without recourse to physiological measurement. Lang's classical work (Lang, 1968; Turpin, 1991) on the measurement of anxiety and the three systems approach which utilized

behavioural, cognitive and physiological approaches is a prime example of this argument. Moreover, there may be situations where systematic biases might be introduced with respect to self-report (i.e. forensic settings) where the assessment of 'truth and honesty' (i.e. lie detection) or the presentation of certain disorders (e.g. Post-Traumatic Stress Disorder) are claimed to be more accurately assessed using psychophysiological techniques. This raises the interesting question as to how objective psychophysiological indices truly are and whether they themselves can be subject to conscious manipulation and bias (Iocano, 2000).

Doubts concerning objectivity are not the only disadvantages to be considered when adopting psychophysiological techniques. Whether claimed psychophysiological indices of putative psychological constructs are either reliable or valid may also be subject to challenge. With respect to reliability, psychophysiological measures might be heavily influenced by the setting and situation in which they are obtained. This may give rise to problems of generalizability, if care is not taken to carefully standardize methods, settings, paradigms and materials. Reported test-retest reliabilities vary considerably across different psychophysiological indices (Strube, 2000). Similarly, due to the practical constraints of assessing large numbers of individuals, standardised norms for psychophysiological measures are few and far between. This provides very definite psychometric limits to the application of psychophysiology to the single case.

Specific psychophysiological theories are also limited and measures are usually interpreted within the context of other theoretical frameworks from cognitive psychology and elsewhere. Sometimes this results in psychophysiological measures having particular interpretations, which are assumed rather than empirically based. An example being whether cardiac deceleration, a common psychophysiological response, should be interpreted as an index of the orienting response, the detection of stimulus novelty or merely just stimulus registration (Ohman, Hamm & Hugdahl, 2000). Similarly, psychophysiological constructs can persist even though their empirical basis may be either insubstantial or even contradictory. Perhaps the best example and one, which is commonly used within applied settings, is the notion of arousal.

Arousal is still used as a major explanatory concept in many applied social and clinical settings despite much psychophysiological research, which has been deeply critical of the construct (Gardener, Gabriel & Diekman, 2000; Turpin & Heap, 1998). This can lead to major problems regarding interpretation and construct validity.

Finally there are issues to do practical utility. Psychophysiological measures usually require complex electronic machinery for physiological measurement, sophisticated computer software for data acquisition and analysis, laboratory environments and trained technicians. These resources are expensive and may not be widely available. Furthermore, the reliance on laboratory settings may also preclude many applied settings. Consequently, many recent applications have relied on the development of ambulatory methods.

### **Applied Constructs and Uses**

As discussed above, there are a wide range of potential applications for psychophysiological measures and approaches. Within the space limitations of this entry it is impossible to present even an overview of different types of applications. However, we will describe some recent examples. Before doing so, a distinction perhaps needs to be made between applied research and research in applied settings. Much psychophysiological research is geared to applied questions relating to psychological understanding of important issues such as health and disease. However, this tends to be laboratory-based experimental research and is directed at using psychophysiological measures to seek answers to fundamentally theoretically relevant questions but with consequences for applied areas. For example, there has been an impressive growth in studies employing the potentiated startle paradigm as a method of assessing emotional valence, and anxiety in particular (Lang, Bradley & Cuthbert, 1990). At a theoretical level, this research has increased understanding of how fear cues are processed at both conscious and pre-attentive levels, and the possible neural substrates underlying some of these mechanisms (Lang, Davis & Ohman, 2000). The question arises, therefore, whether these techniques can be transferred into an applied setting and used for more practical purposes? Could measures of

potentiated startle be used to discriminate between different diagnostic groups of anxiety disorders, could they accurately track response to treatment and indicate therapeutic outcomes and gains? Unfortunately, there are in reality few areas of psychophysiology, which are used routinely in professional psychology practice. Perhaps the only real examples are biofeedback treatments and polygraphy. Nevertheless, major areas of psychophysiological endeavour such as evoked potential research influence practical applications in other areas such as clinical neurology or audiometry.

Common clinical research applications of psychophysiological measures have been as measures of attention within schizophrenia: these have included electrodermal measures of orienting, P300 type event-related potentials (EP) and early sensory gating EPs (Miller et al., 2000). Recent applications of dense array EEG to look at lateral distribution of brain activity, especially over prefrontal cortex and its relationship to affective processing and depression (Davidson, 1992). Anxiety disorders research has focused on the potentiated startle paradigm (Lang et al., 1990), as described above, together with studies of autonomic balance within Generalized Anxiety Disorders (Thayer & Lane, 2000). Therapeutic applications of psychophysiology continue in the form of studies of relaxation and meditation (Turpin & Heap, 1998) and biofeedback (Schwartz, 1995).

Psychophysiological studies within the discipline of health psychology continue to examine mechanisms underlying cardiovascular disease (Stoney & Lentino, 2000). Studies aimed at assessing cardiovascular reactivity to psychologically challenging events continue to be performed (e.g. Fredrickson & Matthews, 1990). A particular focus is the relationship between laboratory-based studies and ambulatory-monitoring based studies of reactivity. Psychophysiological measures have been particularly adopted to assess the role of stress in contributing to the aetiology and maintenance of common physical conditions. In addition to the usual autonomic measures such as heart rate and blood pressure reactivity, many studies examining 'stress' exploit techniques from psycho-immunology and endocrinology: using biochemical assays of immune or hormonal status (Uchino, Kiecolt-Glaser & Glaser, 2000).

Human factors psychophysiology has traditionally examined problems such as assessing alertness and sleep quality, mental workload and performance, and man-machine interactions. A full range of measures have been employed including endocrinological assays (Lovallo & Thomas, 2000) to evoked potential applications to man-machine interactions. Spectral analysis of physiological parameters over extended periods of time or different activities is a technique frequently employed in ergonomic applications. Mulder (1992), in particular, has exploited measures of heart rate variability to assess attentional and workload factors.

## CONCLUSION

Psychophysiology has a long tradition as being used within applied settings. Advances in technology have broadened the range of settings in which psychophysiological measures can be obtained. Developments in neuro-imaging (e.g. Reiman, Lane, Van Petten & Bandettini, 2000) also mean that psychophysiological techniques can now address exciting questions of functional brain-behaviour relationships. Hopefully, these techniques will be extended so as to include more applied questions and applications.

## References.

- Cacioppo, J.T. & Tassinary, L.G. (Eds.) (1990). *Principles of Psychophysiology: Physical, Social, and Inferential Elements*. Cambridge: Cambridge University Press.
- Cacioppo, J.T., Tassinary, L.G., & Berntson, G.G. (Eds.) (2000). *Handbook of Psychophysiology*. Cambridge: Cambridge University Press.
- Coles, M.G.H., Donchin, E. & Porges, S.W. (Eds.) (1986). *Psychophysiology: Systems, Processes and Applications*. New York: Guilford.
- Davidson, R.J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20, 125–151.
- Donchin, E. & Coles, M.G.H. (1988). Is the P300 component a manifestation of context updating? *Behavioural and Brain Sciences*, 11, 354–356.
- Fredrickson, M. & Matthews, K.A. (1990). Cardiovascular responses to behavioural stress and hypertension: a meta-analytic review. *Annals of Behavioural Medicine*, 12, 30–359.
- Gardener, W.L., Gabriel, S. & Diekman, A.B. (2000). Interpersonal processes. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology*. (pp. 643–664). Cambridge: Cambridge University Press.
- Gottman, J.M. & Levenson, R.W. (1992). Marital processes predictive of later dissolution: behaviour, physiology, and health. *Journal of Personality and Social Psychology*, 63, 221–233.
- Graham, F.K. (1979). Distinguishing among orienting, defense and startle reflexes. In Kimmel, H.D., Van Olst, E.H. & Orlebeke, J.F. (Eds.), *The Orienting Reflex in Humans*. (pp. 137–167). Hillsdale NJ: Erlbaum.
- Hugdahl, K. (1995). *Psychophysiology: The Mind-Body Perspective*. Cambridge: Harvard University Press.
- Iocano, W.G. (2000). The detection of deception. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology*. (pp. 772–793). Cambridge: Cambridge University Press.
- Keller, J., Hicks, B.D. & Miller, G.A. (2000). Psychophysiology in the study of psychopathology. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology* (pp. 719–50). Cambridge: Cambridge University Press.
- Kramer, A.F. & Weber, T. (2000). Applications of psychophysiology to human factors. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology*. (pp. 794–813). Cambridge: Cambridge University Press.
- Lang, P.J. (1968). Fear reduction & fear behaviour: problems in treating a construct. In Shlien, J.M. (Ed.), *Research in Psychotherapy*. Washington, DC: American Psychological Association.
- Lang, P.J., Bradley, M.M. & Cuthbert, B.N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, 97, 377–98.
- Lang, P.J., Davis, M. & Ohman, A. (2000). Fear and anxiety: animal models and human cognitive psychophysiology. *Journal of Affective Disorders*, 61, 137–59.
- Lovallo, W.R. & Thomas, T.L. (2000). Stress hormones in psychophysiological research. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology* (pp. 342–367). Cambridge: Cambridge University Press.
- Mulder, L.J.M. (1992). Measurement and analysis of heart rate and respiration for use in applied environments. *Biological Psychology*, 34, 205–236.
- Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ: Erlbaum.
- Ohman, A., Hamm, A. & Hugdahl, K. (2000). Cognition and the Autonomic Nervous System. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology* (pp. 533–75). Cambridge: Cambridge University Press.
- Reiman, E.R., Lane, R.D., Van Petten, C. & Bandettini, P.A. (2000). Positron emission tomography and functional magnetic resonance imaging. In Cacioppo, J.T. Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology* (pp. 85–114). Cambridge: Cambridge University Press.

- Schwartz, M.H. (1995). *Biofeedback – A Practitioner’s Guide* (2nd ed.). New York: Guilford Press.
- Siddle, D.A.T. (1983). *Orienting and Habituation: Perspectives in Human Research*. Chichester, UK: Wiley.
- Stern, R.M., Ray, W.J. & Davis, C.M. (1980). *Psychophysiological Recording*. New York: Oxford University Press.
- Stoney, C.M. & Lentino, L.M. (2000). Psychophysiological applications in clinical health psychology. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology*. (pp. 751–771). Cambridge: Cambridge University Press.
- Strube, M.J. (2000). Psychometrics. In Cacioppo, J.T. Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology*. (pp. 849–869). Cambridge: Cambridge University Press.
- Thayer, J.F. & Lane, R.D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, 1, 000–000.
- Tassinary, L.G. & Cacioppo, J.T. (2000). The skeleto-motor system: surface electromyography. In Cacioppo, J.T. Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology* (pp. 163–199). Cambridge: Cambridge University Press.
- Turpin, G. (Ed.) (1989). *Handbook of Clinical Psychophysiology*. Chichester: Wiley.
- Turpin, G. (1991). The psychophysiological assessment of anxiety disorders: three-systems measurement and beyond. *Psychological Assessment*, 3, 366–75.
- Turpin, G. & Heap, M. (1998). Arousal reduction methods: relaxation, biofeedback, meditation and hypnosis. In Hersen, M & Bellack, A. (Eds.), *Comprehensive Handbook of Clinical Psychology. Adults: Clinical Formulation and Treatment*, Vol. 6. (pp. 203–227) London: Elsevier.
- Uchino, B.N., Kiecolt-Glaser, J.K. & Glaser, R. (2000). Psychophysiological modulation of cellular immunity. In Cacioppo, J.T., Tassinary, L.S. & Berntson, G.G. (Eds.), *Handbook of Psychophysiology* (pp. 397–424). Cambridge: Cambridge University Press.

Graham Turpin

### Related Entries

AMBULATORY ASSESSMENT, ANXIETY, ANXIETY DISORDERS, PSYCHOPHYSIOLOGICAL EQUIPMENT & MEASUREMENTS, BRAIN ACTIVITY MEASUREMENT, EQUIPMENT FOR ASSESSING BASIC PROCESSES

## APPLIED FIELDS: WORK AND INDUSTRY

### INTRODUCTION

Very broadly one might say, that wherever people are busy there is a chance and a need for psychological assessment. However, it is impossible to name all fields in work and industry which are open for psychological assessment. The psychological assessor just has to look at the world of work and industry around him in order to find out what he might contribute. This may be done in terms of theories and constructs which allow to evaluate work and industriousness. This may be done by instruments which operationalize constructs and measures that are reliable and valid. This may be done in terms of methods, designs, and results to present to a customer or a team of experts.

One approach to systematize assessment in applied fields in general, and of work and

organization in particular, is to take an Individual, Group, or Organizational perspective.

### INDIVIDUAL PERSPECTIVE

Starting with assessing the Individual within a company or an organization one might question ‘what, when, what for’: Of course, psychological assessment is of interest in order to learn more about the individual’s strengths and weaknesses, about his attitudes and beliefs, and about his competencies and potentials. Here, methods like in mental tests, reaction time studies, occupational personality scales (Ones & Viswesvaran, 2001), motivation scales, and opinion questionnaires are called for. The aim is to describe a person as fully as is needed to evaluate of how she or he will do (well) on a prospective job. Thus data at job entry are



used to forecast the 'zone of proximal development' of an applicant. One has to recognize (see Furnham, 2001) that an individual:

- *chooses* a job based on pay, location, job security, and training based on his personality traits, attitudes, and values.
- *adapts* to a job out of necessity, insight, motivation
- *changes* a job by altering the physical and social environments.
- *evolves* with new technology, markets and global requirements according to what he understands are necessary requirements in the future.

All this is open for assessment. But assessment of an individual does not stop at job entry. Any job confronts incumbents with a variety of minor and major challenges. One of it is to learn to function well at a certain position. Thus learning gains or developing several competencies are of interest to assessors. Assessment results lead to improve the interaction with the individual and the work place by considering human factors for improved functioning, by motivating the individual, by designing up to date remuneration schedules, by considering aptitude treatment interactions in designing effective training programs, by monitoring communication and coordination with others, by communication and coordination programs to name but a few.

A new aspect for assessment emphasizes licensing professionals as an aspect of overall quality assurance in production and service. Companies may want coworkers who have knowledge, skills, and competencies to deal with their products within the company itself but, even more important, at all customer sites. The service person for a database product of a regional bank may create quite a loss if a new program release is not handled with care. This is part of the liability movement in modern societies which assures that products and services do not do any harm to others. Here, with each new product and each new service there has to be a model of proper use and a contingent assessment of its components. So assessment takes place in regard to accreditation and licensing.

During a person's professional life there are numerous occasions to assess what an incumbent's profile of competencies is like, or to find out about the set of strengths and weaknesses in

order to assign someone to a proper position for the sake of himself and the benefit of the organization. Placement decisions should be based on sound assessment data.

Even at the end of a career, assessment may help to find a new position outside the organization by means of outplacement or early retirement plans. One might also have to look at the loss or weakening of competencies and skills over time and find means and measures to decide about rehabilitation programs. Here, it is of interest what residual competencies are available, to which degree, and how they should be built upon in a rehabilitation training.

Seen as such, psychological assessment is a work life long companion activity which serves the individual and the organization in order to fruitfully monitor the interaction between both of them. The psychological well being of the individual is a target as also the reasonable use of his forces at work. Assessment emphasizes prerequisites to job demands, trainings, and personality developments. However, it also emphasizes effects of all the aforementioned after a new job was assigned, a training was accomplished, and a personal challenge was taken. Assessment data are vital to human resource management and thus have to be valid, reliable, and objective in the first place to sustain all personnel decisions that are taken.

(Related entries of this volume: Achievement motivation ass., achievement testing, affect ass., biographical measures in work ass., burn our ass., Assessment centres, cognitive abilities or mental aptitudes ass., creativity ass., emotional intelligence ass., ass. of intelligence, interest ass., interview, locus of control ass., motivation ass., observational methods, personality ass., ass. in personnel selection, portfolio ass., practical intelligence ass., problem solving ass., ass. of reasoning, self efficacy ass., self reports, temperament ass., ass. of personal constructs, personality ass., vocational ass.)

## GROUP PERSPECTIVE

Assessment at the group level is mainly oriented towards productivity, conflict resolution, good communication, and coordination. One may want to look at the social functioning of a team by means of a sociogram (Moreno, 1951; see

Sociometric Methods this volume), by means of interaction analysis (Bales, 1950, SYMLOG), by means of a questionnaire about role ambiguity (Rizzo et al. 1970), or by observation studies in an obtrusive or non-obtrusive manner (Putnam and Jones, 1982).

Some assessments are status oriented and should allow to judge what are the prevailing attitudes or obstacles in group life in order to go from here to improve it. Actions may involve changes in group memberships, group trainings, or re-groupings at large.

More of a process-oriented approach is called for if monitoring of actions is of interest. Longitudinal assessment data are needed to describe what changes take place in a group and explain why these changes occur. Cross sectional data reveal how different groups develop independently from each other. Harrison and Shirom (1998: 161) present some key group factors: (1) Group Composition, Structure and Technology like nationality mix, divergence of professions, decision procedures, control procedures like evaluation, comprehensiveness of controls, and (2) Group Behaviour, Processes, and Culture like relations among members, reward types, direction of information flows, openness, decision making, supervisory behaviour (supportiveness, participation, goal setting, performance expectation, conflict management).

Topics may range from modern shift systems, remuneration schedules, new production techniques, new forms of cooperation and coordination, integrating minorities, client centredness of work, quality assurance at each production step, self-organization of the team, group cohesion, role conflicts/clarity, mobbing propensity, co-worker-supervisor relationship, learning needs. This list is by far not complete but it displays minor and major topics which may be subject to an in-depth assessment. Practical problems are closely linked to some kind of sometimes political action on behalf of the management and the labour union representatives.

(Related entries of this volume: applied behavioural analysis, attitudes ass., coping style ass., ass. of couples, dynamic ass., goal attainment ass., ass. of groups, irrational beliefs ass., leadership ass., learning strategies ass., ass. of minorities, personal constructs ass., observational methods, self reports, time orientation ass., ass. of wisdom, work performance ass.)

## ORGANIZATIONAL PERSPECTIVE

Organizational assessment is by far more macroscopic than the foregoing two approaches (see Harrison and Shirom (1998). First, it has to be defined, what is the organization under scrutiny. Some of them are small shops in a small region and others are global players operating on quite diverse markets. Second, the perspective may change if one considers an organization from within, its inner dynamics, its members, in contrast to considering clients, suppliers, and organization members at the same time.

In order to assess, i.e. describe, an organization's climate, for example, quite different actions have to be taken. One has to look at what attracts people to an organization, what keeps them within, and what are the typical characteristics of those who are there for a given time (Schneider, 1987). So even for personnel marketing and in recruitment campaigns one may want to use self-assessment instruments (see Self Assessment in this volume). Pritchard and Karasick (1973) provide a scale with eleven dimensions like Autonomy, Conflict vs. Cooperation, Social Relations, Structure to name but a few. Based on this and other research, James and James (1989) provided a model which emphasizes (1) role stress and lack of harmony, (2) challenge at work and autonomy, (3) facilitation by leadership and support, (4) cooperation, friendliness, and warmth in a team. As Weinert (1998) points out these factors are related to roles, leadership, and teams.

Organizational culture (Schein, 1985; see Assessment of Organizational Culture this volume), an adjacent construct, emphasizes common shared values, norms, goals, beliefs, and perspectives. Thus here the scope is on meaning, intentions, purpose of work and tasks, as well as on methods to achieve organizational essentials and underlying norms and values in all what members do. Artifacts and behaviour patterns are by far more visible than beliefs, cognitions, and basic assumptions within a company. Sackman (1992) referred to cognitive orientations as part of organizational culture and identified four forms:

- dictionary knowledge – definitions of labels and definitions
- directory knowledge – assumptions of how common practices work and what are presumably causal relations

- recipe knowledge – prescriptions for improving remedy processes or urgencies
- axiomatic knowledge – about nature of things and why events occur.

The reasons for an assessment are manifold, too. There may be a constant interest in changes of the organization, a need to assess the organization prior or as a consequence of a re-organization, an in-depth view of what merging with another company had as effect, to assure that new products and production techniques are adopted by the workforce, to find out how new markets would affect the members, and what challenges are perceived in the light of new clients. The scope is always to find out something about the organization as a whole. Most of this will be assessed by means of questionnaires, but some is discernible by interviews, observation, or unobtrusively browsing through documents, self-reports, marketing material, and guidelines. More qualitative than quantitative results are likely with the latter.

An investigation may be launched at the beginning of a change in organizational behaviour or the end of a campaign. In particular, many questionnaire-based actions are meant to shed light on aspects the management wants to emphasize. So the questions are one means to convey to the workforce what is considered essential to the organization. The questions altogether convey a message as such, and subsequent results tell everyone the degrees to which essentials are shared. If, for example, there are several questions about cooperation formats then the responders are geared to particularly perceive this construct and evaluate his momentary reflections on this. Thus the questionnaire is highlighting a concept which may be on the organization's agenda.

Scaffoldings of how to organize an assessment are given by the Open-Systems Analysis (Harrison & Shirom, 1998), Six Box Model (Weisbord, 1976), Stream Analysis (Porras, 1987) just to name a few.

## ASSESSMENT INSTRUMENTS

There are published instruments which allow even standardized interpretation. But their drawback may be that they do not address the present

problem and thus do not answer the question raised fully. In the case of assessment of an individual's behaviour there are numerous instruments from Differential Psychology. But if one addresses group and organizational problems one finds less and less formalized and standardized instruments. One help may be 'instruments' shared among psychological assessors who worked out a scale, evaluated it at one site or within one company, but made it available to others. At least some kind of documentation about intended scope, design, and small scale results and evaluations are available (Drasgow and Schmitt, 2001).

So ad hoc instruments are created by internal staff or outside consultants. Often a sound explication and elaboration of constructs is missing. Some instruments lack a theory-based pre-evaluation of questions to be asked. This is sacrificed to immediate results because market forces drive the management to deciding. One may definitely wish that even a 'simple' questionnaire is considered and valued as a measuring instrument in itself. It provides sound data only if it had been designed and developed according to goals, established theory, constructs, and empirical results. In 'rapid practice' questions are ambiguous and so are results. Often the questionnaire falls short of a sound coverage of facets and so data are incomplete or highly one-sided.

There are, of course, good guidelines (Fleishman and Quaintance, 1984) as to how to construct a good measure. Many instruments ought to be based on sound job descriptions to pre-define relevant target behaviours, task-related competencies, and job-related social skills (see Job Characteristics Assessment in this volume). Also (item and/or person) sampling techniques (Shoemaker, 1973) allow to save costs at the expense of not asking everyone that should be invested in instrument design and evaluation.

As was mentioned above, apart from questionnaires there may be used interviews, observations, survey-feedback approaches, simulations, grid-techniques (Jenkins, 1998), and scenario techniques for example. However, the less standardized they are, the more the assessment errors that may be committed. In general, any instrument should be closely designed for that purpose it has to serve. Ad hoc instruments should be avoided, but instruments with some empirical underpinning should be preferred. The former only allow

an assessment per fiat and the latter an assessment per fact.

## ASSESSMENT DESIGNS

Designs of how to conduct an evaluation study (Cook & Campbell, 1979; Sanders, 1994) are available since long. But in regard to sound assessment of effects of introduced changes at the person, group, or organizational level there ought to be more than one measure of an effect, and even a pre-measure should be available as a standard against which one may judge any changes. Restructuring of an organizational unit is quite an investment, and it is desirable to trace back to a prior measure what and how much has changed. Often enough an effect is ascertained but vanishes over time. So more than one post measure is advocated. Designs may be borrowed from Educational Psychology (Campbell & Stanley, 1963) to assure that assessed changes are true changes and not just valid for a short time.

Not only is it possible to sample individuals, but content areas can be sampled as well (Shoemaker, 1973; Hornke, 1978) in order to have a sound picture. It is not necessary to ask everyone the same questions, and have many duplicated answer pattern. Good design of individuals and content samples yield sufficient reliable and valid data and will help to save costs quite a bit. It just demands a bit of prior construct knowledge, some speculation about possible effects, and a kind of intelligent logistic in regard to data collection. Not any all-embracing survey is worth its efforts and investments. Sometimes a sound less is much more!

## FUTURE PERSPECTIVES AND CONCLUSIONS

The initial and implicit question, of what the fields of psychological assessment are in regard to work and organization, can only be answered at a surface level. It is left to the ingenious assessor and his efforts, interests, and creativity to sense what the fields of assessment activities are. No one assigns them to him and even a contract allows for sound science-based assessments the contractor himself might not have had in mind. Applied fields in this sense are all those fields

which help to improve an individual's, a group's, and an organization's life. The latter is for the benefit for all of them.

## References

- Bales, R.F. (1950). A set of categories for the analyses of small group interaction. *American Sociological Review*, 15, 146–159.
- Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In Gage, N.L. (Ed.), *Handbook of research on teaching*. Chicago, IL: Rand McNally.
- Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation-Design & Analysis Issues for Field Settings*. Chicago, IL: Rand McNally.
- Drasgow, F. & Schmitt, N. (Eds.) (2001). *Measuring and Analyzing Behaviours in Organizations*. San Francisco: Jossey Bass Publishers.
- Fleishman, E.A. & Quaintance, M.K. (1984). *Taxonomies of Human Performance: The Description of Human Tasks*. Orlando, FL: Academic.
- Furnham, A. (2001). Personality and individual differences in the workplace-Person-Organization-Fit. In Roberts, B.W. & Hogan, R. (Eds.), *Personality in the Workplace*. (pp. 223–252). Washington DC: American Psychological Association.
- Harrison, M.I. & Shirom, A. (1998). *Organizational Diagnosis and Assessment: Bridging Theory and Practice*. Thousand Oaks, CA: Sage Publications.
- Hornke, L.F. (1978). Personen-Aufgaben-Stichproben. In: Klauer, K.J. (Ed.), *Handbuch der pädagogischen Diagnostik*, Band 1. Düsseldorf: Schwann.
- James, L.A. & James, L.R. (1989). Integrating work environment perceptions. Explorations into the measurement of meaning. *Journal of Applied Psychology*, 74, 739–751.
- Jenkins, M. (1998). *The Theory and Practice of Comparing Causal Maps*. London: Sage.
- Moreno, J.L. (1951). *Sociometry, Experimental Method and the Science of Society*. New York: Beacon House.
- Ones, D.S. & Viswesvaran, C. (2001). Personality at work: criterion-focused occupational personality scales used in personnel selection. In: Roberts, B.W. & Hogan, R. (2001). *Personality in the Workplace*. Washington, (pp. 63–92). DC: American Psychological Association.
- Porras, J.I. (1987). *Stream Analysis*. Reading, MA: Addison-Wesley.
- Pritchard, R.D. & Karasick, B. (1973). The effects of organizational climate on managerial job performance and job satisfaction. *Organizational Behaviour and Human Performance*, 9, 110–119.
- Putnam, L.L. & Jones, T. (1982). Reciprocity in negotiations: an analysis of bargaining interaction. *Communication Monographs*, 49, 171–191.
- Rizzo, J.R., House, R.J. & Lirtzman, S.I. (1970). Role conflict and ambiguity in complex organisations. *Administrative Science Quarterly*, 15, 150–153.

- Roberts, B.W. & Hogan, R. (2001). *Personality in the Workplace*. Washington, DC: American Psychological Association.
- Sackman, S. (1992). Cultures and subcultures: an analysis of organizational knowledge. *Administrative Science Quarterly*, 37, 140–161.
- Sanders, J.R. (1994), (Ed.). *The Program Evaluation Standards* (2nd ed.). Thousand Oaks, N.J.: Sage Publications
- Schein, E.H. (1985). *Organizational Culture and Leadership*. San Francisco, CA: Jossey-Bass.
- Schneider, B. (1987). The people make the place. *Personnel Psychology*, 28, 447–479.
- Shoemaker, D.M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, Mas: xxx
- Weinert (1998). *Organisationspsychologie*. (4th ed.), Weinheim: Psychologie Verlagsunion.
- Weisbord, M.R. (1976). Diagnosing your organizations: six places to look for trouble with or without

theory. *Group & Organizational Studies*, 1, 430–447.

Lutz F. Hornke

### Related Entries

COGNITIVE/MENTAL ABILITIES IN WORK & ORGANIZATIONAL SETTINGS, INTERVIEW IN WORK & ORGANIZATIONAL SETTINGS, MOTOR SKILLS IN WORK SETTINGS, OBSERVATIONAL TECHNIQUES IN WORK AND ORGANIZATIONAL SETTING, PERFORMANCE IN WORK SETTINGS, PHYSICAL ABILITIES IN WORK SETTINGS, RISK & PREVENTION IN WORK & ORGANIZATIONAL SETTINGS, SELF REPORTS IN WORK AND ORGANIZATIONAL SETTINGS.

# A

---

## ASSESSMENT PROCESS

### INTRODUCTION

In solving daily life problems, we automatically execute a lot of judgement and decision making. We also gather information or consult others in order to make well-informed decisions and judgements. The assessment process in the field of psychology is about the gathering and processing of information by a professional in order to get well-informed judgements and decisions concerning a specific request made by a person or an organization. The client is either a person or an organization that made the request; the subject is the person or organization who is the target of the assessment. Psychological assessment refers to the judgements and decisions made by the professional psychologist. Assessment process refers to how these judgement and decisions came about and how these judgement and decisions are communicated to the client.

Contrary to the layperson, the professional has the obligation to process his judgements and decisions according to three sets of standards: ethical standards, social standards, and methodological standards. Ethical and social standards apply to all fields of professional psychology. It is

with respect to the methodological standards that the professional gets his or her identity as an academically educated expert in a particular field. Most methodological standards in the field of assessment published in the standards of the professional organizations are related to the methods and procedures the psychologist uses in collecting information. Standards or guidelines with respect to the assessment process are not so well articulated. Actually, it is only recently that the European Association of Psychological Assessment installed a Task Force to formulate Guidelines for the Assessment Process (GAP) (Fernández-Ballesteros, 1998).

This entry contains four sections. The first section highlights the distinction between assessment and testing. The second section analyses the assessment process. The third section mentions some of the biases that may disturb the intrinsic validity of the process and mentions some remedies proposed in the literature. The fourth section points to developments in the field that try to model the assessment process. The last section pays attention to the most recent contribution to the field, which is the production of professional guidelines for the assessment process.

**ASSESSMENT AND TESTING**

The relatively late attention to the quality of the assessment process might partly be explained by the dominant position of the psychometric approach in assessment. Psychometrics is the discipline that deals with formal statistical foundation of measuring and validating individual differences. In the field of applied psychometrics this tradition focuses on two issues: the development of psychometrically sound tests and the validation of these tests with respect to external criteria. A test is psychometrically sound when it proves to be an objective, quantitative, and reliable measure of individual differences. It is psychometrically valid when its scores predict the position of the examinee on some other criterion or characteristic. In order to be accepted as a test, the instrument must be constructed and validated according to the prescriptions of the existing psychometric theories (Allen & Yen, 1979, and Nunnally, 1978, for further documentation). The psychometric tradition has proven very valuable and both test theory and testing are integrated in the academic education of assessors. Moreover, the tradition has witnessed distinguished scholars who published fine books on testing and test use (Anastasi & Urbina, 1997; Cronbach, 1990).

Assessment is a summary term which refers to all the activities the assessor, performs in producing an answer to the client's request. These activities may include testing among other activities, such as analysing the client's problem, generating hypotheses about its causes and searching for the appropriate intervention. It is the analytical and constructive quality of the assessment process that distinguishes assessment from mere testing.

**THE PROCESS**

The assessor has to analyse the request and to integrate his results in a case formulation, which takes into account the available knowledge in the field. In doing so he has to follow the same kind of logic any scientific researcher follows in deductively inferring hypothesis, in testing these hypotheses, and in formulating conclusions in the framework of the available knowledge. However, although the assessment process follows the same kind of logic as in any scientific search process the

context differs basically from that of the scientific process (De Groot, 1969; Sloves et al., 1979).

For the scientific researcher in psychology the context of his or her work is the body of knowledge at a particular domain and the researcher is focused upon phenomena not yet explained within that particular domain. The goal of the scientific researcher is to find descriptions and explanations that generalize across persons and situations. It is not the concrete person who is the subject, but general phenomena such as perception, motivation, or personality dimensions. The assessor, however, focuses on the person with his or her particular problems in his or her past as well as present situation. The primary goal of the process is to contribute to the solution of a person's problems. The more the person's problems can be described and understood as representative for problems shared by other persons, the more the assessor can rely on a common body of knowledge and apply procedures and protocols developed for specific groups of clients. However, in many cases the assessor cannot just apply already established knowledge. Instead, he or she has to rely on his or her methodological and professional experience in using the state of the art in the field to design a client-tailored procedure and to make an educated interpretation of the outcome.

When talking about the client's problem, it is important to make a distinction between adjustment problems and clinical problems. By problem is meant a psychological state of uncertainty for which neither the client nor his or her social network sees a preferred course of action. Adjustment problems are problems all people encounter in their daily life, and for which they may want to seek professional advice. Examples of such problems are marital conflict, study choice, and career planning. Clinical problems are problems that have dysfunctional effects on the psychological and social well being of the client. In assessing adjustment problems, the assessor uses instruments and applies knowledge that belongs to the domain of general psychology. In assessing clinical problems, the assessor uses tools and knowledge that pertain to the domain of clinical psychology.

An important part of the assessment process is the explanation to the client of why and how the assessment tools are applied and how strong the evidence is, which may be the outcome of the process. The kind of assessment tools and

knowledge involved are triggered by the requests the assessor has to answer. The simplest format to describe such requests is that of a question as if phrased by the client. Examples of such questions within the non-clinical domain are: 'Am I suited for this type of job?', 'Which qualities do I have to develop in order to be eligible for this particular education program?', 'What conditions at the workplace are responsible for getting the high rate of job turnover?'. Examples of questions in the clinical domain are: 'How serious are my anxieties?', 'Why does it happen to a person like me to be burn out', 'Do I need psychological treatment to master my feelings of self-worthlessness?'.

Concrete requests and related questions automatically specify the kind of assessment activities the assessor should perform in order to answer the questions. For instance, in order to answer the question 'How serious are my anxieties?', the assessor first has to describe the anxieties and, secondly he or she has to evaluate the anxieties against some standard or norm of severity. In answering the question 'What conditions are responsible for the labor turnover', the assessor first has to check whether the turnover is unusually high (again evaluation against a standard). Secondly – when the latter is the case – he or she has to hypothesize about conditions and, thirdly, to test these hypotheses by collecting data and evaluating the outcome.

Whatever the steps taken in the process, the process ends in an advice to the client. The oral and written report of the course and outcome of the assessment process must give the client a fair and evidence-based account of the given advice. The assessor should be careful in conveying the probabilistic and conditional nature of his or her statements.

## FLAWS AND BIASES

The assessment process contains many instances in which the assessor, alone or in dialogue with the client, determines the course of action. The assessor should be aware of and protect him- or herself against the flaws and biases of clinical judgement. Clinical judgement refers to informal and subjective thinking and decision making. There is ample evidence that the professional psychologist who is not armed by proper decision

aids is as weak a decision maker as the less-trained professional or lay-person.

The studies which demonstrate the fallibility of the clinical judgement and decision making belong to three different research streams which can be labelled as the psychometric, cognitive and social-psychological tradition. The *psychometric* tradition offers evidence for the fact that clinical prediction is nearly always less accurate than a prediction made by standardized formal predictions. Meehl already drew this conclusion in 1954, and ever since he was supported by many other reviews, the most recent one was presented by Grove et al. (2000). If one wants to predict a person's state of mind or behaviour in the future, the best thing to do is to base the prediction on the outcomes of empirical studies of the relationship between predictor (present state) and criterion (future state).

The *cognitive* research tradition presents evidence that cognitive heuristics which allow people to operate rather well in their daily lives nevertheless may have distorting effects in dealing with restricted and probabilistic information. Since the seminal work of Tversky and Kahneman (1974) the distorting effect of cognitive heuristics have been demonstrated in all kinds of choice and decision situations and with all kinds of people, professionals as well as laypersons (see Baron, 1994, and Goldstein & Hogarth, 1987, for a review). Of special critical interest for the assessment process are the heuristics people use in the generation and testing of hypotheses. One of the most famous heuristics in this respect has been called the confirmatory test strategy. People have the strong tendency to test hypotheses by searching the information that confirms the hypothesis and to neglect searching information that would disconfirm the hypothesis.

The *social-psychological* tradition presents evidence that in meeting the client the clinician is inclined to select and interpret information from the perspective of his causal attributions, stereotypes and characteristics. Of specific interest to the assessment process is the actor-observer bias hypothesized by Jones and Nisbett (1971) and empirically demonstrated in several studies (see Turk & Salovey (1988) for a review). In explaining their behaviour actors tend to attribute it to situational factors while observers tend to attribute this behaviour to internal causes like traits and motives.

Studies of flaws and biases have automatically led to the question how these flaws and biases could be avoided or at least restricted. Several proposals have been made, ranging from further standardization of data-collection and empirical validation of prediction procedures involved up to debiasing reasoning techniques and computerized decision aids. Lists of such proposals are given in Garb (1998), Haynes and O'Brien (2000), and Turk and Salovey (1988).

## MODELLING THE PROCESS

In many fields of professional psychology, one always has been well aware of the intricacies and fallacies of an assessment process that is not protected somehow against the flaws and biases of clinical judgement. Considerable progress has been made in standardizing the way in which information can be gathered by using reliable and valid tests by which a client's response can be compared with that of others. However, not only the data collection and statistical interpretation should proceed properly, the same should apply to the comprehensive assessment process, which starts with the client's requests and ends in the assessor's advice to the client.

In non-clinical domains, such as job and curriculum selection, the client's requests relate to the client's strengths and weaknesses with respect to a certain job or study curriculum. Here the relevant empirical body of knowledge is the relationship between the client's characteristics and the success or satisfaction in the job or curriculum at hand. What emanates from this empirical approach is – technically speaking – a multiple regression equation in which the scores on a standardized battery of tests are weighted according to their relationship with the criterion, and combined in such way that the prediction of the criterion is as accurate as possible. The assessment process is modelled after a *statistical prediction model*. The assessment process reaches a level of standardization that equals the level of standardization of each of its components.

Uncertainty about which job or study to engage in most often presents a problem of choice. Not only the probability of success in each of the choice options is at stake, but also the value each of these options have for the client. The value of having success in a particular career is not restricted to

financial profits, but also depends upon more personal values such as social recognition, social identity, and emotional and intellectual satisfaction. The assessment process should result in advice in which probability of success is weighted by the value of that success. In the *utility model* the assessment process is formalized as the combination of probability and values that apply to each of the choice options (Baron, 1994; Von Winterfeld & Edwards, 1986).

Neither the statistical model nor the utility model are developed to model the full assessment process which starts with the client's request and results in an advice to the client. Nor do these models formalize the specific decision rules the diagnostician should follow in going through the main phases of the process. Westmeyer (1975) proposed an *algorithmic model*. In this formal model decision algorithms are supposed to work on an adequate empirical knowledge base which contains complete sets of conditional probabilities for a specified type of both problem and client.

All three models presented so far are *normative* in the sense that they process information according to statistical or decision rules. Strict normative models set formal conditions that usually cannot be met in psychological practice nor in the knowledge base this practice is supposed to work with (see Westmeyer & Hageböck 1992, for a discussion). Therefore, many students of the assessment process have tried to model the process according to more heuristic principles that could guide the process. Most of these *heuristic models* have been restricted to a diagrammatic presentation of the assessment process (Maloney & Ward, 1976) while some others (De Bruyn, 1992; Haynes & O'Brien, 2000) have led to elaborations which show how the assessor can proceed if he wants to follow the logical decision flow depicted in the model.

## FUTURE DIRECTIONS: GUIDELINES FOR THE ASSESSMENT PROCESS

Despite the growing interest in the quality of the assessment process, a comprehensive set of *heuristic guidelines* that could support the assessor in executing the process is still lacking. This is in contrast to the related fields of testing (American Psychological Association, 1999) and program evaluation (Joint Committee on Standards for Educational Evaluation, 1994) which eventually



have succeeded in the formulation of standards that monitor professional work. It is only recently (Fernández-Ballesteros, 1998) that a task force consisting of psychologists from different fields in psychology started to think of formulating guidelines to cover all phases of the assessment process. The task force formulated a set of guidelines to cover the phases of analysing the case, organizing and reporting results, planning the intervention, and evaluation and follow-up. (Fernández-Ballesteros et al., 2001). Instead of being rigid rules, fixed forever, these guidelines represent recommendations for professional behaviour.

As already demonstrated in the fields of testing and evaluation, such guidelines highly contribute to the development of the profession. Therefore, as stated by Fernández-Ballesteros et al.: 'We hope that the efforts made in developing and disseminating these Guidelines stimulate the discussion among interested scientific and professional audiences and, in the long run, will contribute to improve the practice of psychological assessment as well as the education and training of psychological assessors.' (2001: 185).

## References

- Allen, Mary J. & Yen, Wendy, M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole.
- American Psychological Association (1999). *Standards for Educational and Psychological Tests*. Washington, DC: American Psychological Association.
- Anastasi, Anne & Urbina, Susana (1997). *Psychological Testing*. Upper Saddle River, NJ: Prentice Hall.
- Baron, Jonathan (1994). *Thinking and Decision* (2nd ed.). Cambridge: Cambridge University Press. (1st ed., 1988).
- Cronbach, Lee.J. (1990). *Essentials of Psychological Testing* (5th ed.). New York: Harper & Row (1st ed. 1949).
- De Bruyn, E.E.J. (1992). A normative-prescriptive view on clinical psychodiagnostic decision making. *European Journal of Psychological Assessment*, 8(3), 163–171.
- De Groot, Adriaan D. (1969). *Methodology: Foundations of Inference and Research in the Behavioural Sciences*. The Hague: Mouton.
- Fernández-Ballesteros, R. (1998). Task force for the development of guidelines for the assessment process (GAP). *Newsletter of the European Association of Psychological Assessment*, 1(1), 2–7.
- Fernández-Ballesteros, R., De Bruyn, E.E.J., Godoy, A., Hornke, L.F., Ter Laak, J., Vizcarro, C., Westhoff, W., Westmeyer, H. & Zaccagnini, J.L. (2001). Guidelines for the assessment process (GAP): a proposal for discussion. *European Journal of Psychological Assessment*, 17(3), 178–191.
- Garb, Howard J. (1998). *Studying the Clinician: Judgment Research and Psychological Assessment*. Washington, D.C.: American Psychological Association.
- Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E. & Nelson, C. (2000) Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment*, 12(1), 19–30.
- Haynes, Stephen, N. & O'Brien, William (2000). *Principles and Practice of Behavioural Assessment*. New York: Kluwer Academic.
- Hogarth, Robin, M. (1987). *Judgement and Choice: The Psychology of Decision* (2nd ed.). Chichester: Wiley (1st ed., 1980).
- Joint Committee on Standards for Educational Evaluation (1994) *The Program Evaluation Standards* (2nd. ed.) Thousand Oaks, CA: Sage (1st ed., 1981).
- Jones, Edward E. & Nisbett, Richard E. (1971). The actor and the observer: divergent perceptions of the causes of behaviour. In Jones, E.E., Kanouse, D.H., Kelley, H.H., Nisbett, R.E., Valins, S. & Weiner, B. (Eds.), *Attribution: Perceiving the Causes of Behaviour* (pp. 79–94). Morristown, NJ: General Learning Press.
- Maloney, M.P. & Ward, M.P. (1976). *Psychological Assessment: A Conceptual Approach*. New York: Oxford University Press.
- Meehl, Paul E. (1954). *Statistical versus Clinical Prediction*. Minneapolis, MN: University of Minnesota Press.
- Nunnally, Jim C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill (1st ed., 1967).
- Sloves, R.E., Doherty, E.M. & Schneider, K.C. (1979). A scientific problem-solving model of psychological assessment. *Professional Psychology*, 1(1), 28–35.
- Turk, D. & Salovey, P. (Ed.). (1988). *Reasoning, Inference and Judgment in Clinical Psychology*. New York: Free Press.
- Tversky, A. & Kahneman, D. (1974) Judgment under uncertainty: heuristics and biases. *Science*, 185(50), 1124–1131.
- Von Winterfeld, Detlof & Edwards, Ward (1986). *Decision Analysis and Behavioural Research*. Cambridge: Cambridge University Press.
- Westmeyer, H. (1975) The diagnostic process as a statistical-causal analysis. *Theory and Decision*, 6(1), 57–86.
- Westmeyer, H., & Hageböck, J. (1992). Computer-assisted assessment: a normative approach. *European Journal of Psychological Assessment*, 8(1), 1–16.

Eric E.J. de Bruyn

## Related Entries

ASSESSOR'S BIAS, CLINICAL JUDGEMENT, CASE FORMULATION, ETHICS.

# A

---

## ASSESSOR'S BIAS

### INTRODUCTION

Psychological assessment is subject to various errors of measurement. While some are random, as assumed in classical test theory, others are systematic and lead to consistent distortions of the true value of a characteristic. These latter errors may be partially due to assessor's biases. This term does not refer to elementary professional mistakes such as implementing test instructions incorrectly, but to systematic tendencies in case-related information processing that reduce the validity of data. Although these biases normally impair objectivity and reliability, they remain undetected when they are consistent across individuals and time. In addition, a low interrater agreement is not necessarily a sign of assessor's bias but may be due to valid differences between settings and informants (Lösel, 2001).

Not all types of assessment information are equally susceptible to assessor's biases. Whereas standardized tests or biographical inventories are less affected, their impact may be strong in unstructured interviews, behaviour observations, or trait ratings. For example, some studies on the judgement of job performance have shown that more than half of the variance is due to differences in the assessors (Scullen et al., 2000). In a meta-analysis, approximately 37% of the variance in ratings was attributed to them (Hytot & Kerns, 1999).

This article concentrates on biases in assessments by other persons (e.g. psychologists, psychiatrists, teachers, or lay informants). Although these biases are similar to the numerous response sets and distortions in self-reports, some seem less important (e.g. lying, simulation, dissimulation, social desirability, or positive self-presentation) and others more relevant (e.g. halo, leniency, stringency, or contrast effects). In the following, we will first describe several of these errors and afterwards address factors that differentiate and moderate

these distortions. Finally, we will take a brief look at approaches for detecting and reducing assessor's biases.

### EXAMPLES OF ASSESSOR'S BIASES

#### Halo and Logical Error

In psychological assessment, a *halo effect* refers to an overgeneralization from one prominent characteristic of a person to other judgements on this individual. Most typically, it is an overestimation derived from a general impression. For example, if a person is judged to be good in general, he or she will be judged more positively on any specific dimension. Halo errors may arise particularly when there is insufficient information for a detailed assessment or when traits are not well defined. In these cases, the general impression is used to fill information gaps (Saal et al., 1980). A related bias is the *logical error*. Here, assessors are likely to give similar ratings to traits that seem logically related in their minds (Guilford, 1954). Whereas the halo effect derives from a perceived coherence of characteristics in an individual, the logical error refers to a more explicit and abstract coherence of variables or traits. The latter is often anchored in the assessor's subjective personality theory.

Both biases produce the same outcome, namely spurious and inflated correlations (Murphy et al., 1993). The underlying mechanisms are also related. Occasionally, a halo effect can have some advantage because it accentuates differences between individuals (Murphy et al., 1993). This is the case when a quick decision has to be made and the core determinants of the halo effect are empirically valid. Then, one can simply follow the useful decision rule 'take the best, ignore the rest' (Gigerenzer & Selten, 2001).

## Position Effects

Whereas halo effects result from the psychological or logical closeness of the rated characteristics, their sequence or position may have a similar effect. One such distortion is the *proximity error*. Judgements that are close to each other in time or space contain a higher risk of mutual influence. A related error is the *primacy effect* in which the first impression of a person overshadows the assessment of their further behaviour. Its opposite is the *recency effect*: the last information on a person influences the evaluation of previous data. Early stereotyping (primacy) or easy remembering (recency) are among the mechanisms that underlie these modes of information processing. Although some experiments suggest that recency is more influential than primacy (e.g. Betz et al., 1992), it is questionable whether such findings can be generalized to real-life assessments.

## Leniency and Stringency

These errors refer to the tendency to make relatively positive (leniency) or negative (stringency, severity) assessments. For example, somebody who is rather intelligent would be judged to be even more intelligent by a lenient rater but less intelligent by a stringent one. Leniency seems to be more frequent than stringency (Guilford, 1954). It may partially reflect tendencies toward social desirability, harmony, or other dispositions. Assessors who score high in self-monitoring tend to deliver more lenient ratings. Similarly, leniency correlates negatively with conscientiousness and positively with agreeableness (Bernardin et al., 2000). Nonetheless, other studies question the view that leniency is primarily due to personality dispositions. Situational and relational factors must also be taken into account.

## Central Tendency

Leniency and stringency go along with polarizations between extreme judgements. In contrast, other assessors tend to produce scores in the middle range. Sometimes, this may express a lack of differentiated information on a person. In other cases, it involves indifferent perceptions or a general ambivalency or insecurity in the assessor.

## Contrast and Projection Effects

Biases may also result from comparisons between a person's behaviour and the assessor's own dispositions. A *contrast error* is when the assessor attributes characteristics at the opposite pole to his self-perception; a *projection effect* when he evaluates a person as being similar to himself. Both tendencies relate to self-awareness and self-presentation processes in the assessor. For example, persons with behavioural problems may rate others higher on the same dimensions. Whereas projection errors can contribute to self-worth by reinforcing social comparability, contrast effects can serve a similar function by protecting the assessor's individuality.

## Interactional Biases

Assessors' biases not only influence their own information processing but also how persons behave in the assessment situation. Although the assessor's age, gender, ethnicity, role, status, or institutional affiliation may have such effects (Hagenaars & Heinen, 1982), these should not be viewed as biases. Interactional biases refer to influences that derive from the assessor's information processing. One example is the self-fulfilling prophecy of positive expectations, although the typical Rosenthal effect has not been replicated sufficiently (Elashoff & Snow, 1972). In the practice of psychological assessment, even minor biases can have an effect (e.g. slightly nodding the head or providing other nonverbal reinforcements based on halo or leniency effects). Unstructured interviews are particularly vulnerable to biases derived from assessor's attitudes and expectations. Hyman et al. (1975) distinguishes three forms: (a) *Attitude-structure expectations* refer to the belief that the attitudes of the respondent are unified. They resemble the halo effect and may reinforce uniform reactions. (b) *Role expectations* relate to the respondent's membership of a certain group. These stereotypes can result in assessor behaviour that triggers prototypical reactions in the respondent. (c) *Probability expectations* refer to the base rates of diagnostic characteristics in the respective population. They can lead to assessor behaviour that tries to confirm these specific assumptions. Other interactive biases contribute to *missing data*. For example, projection or contrast effects may lead an interviewer to evaluate a question as being

extremely difficult or intimate. This can reduce emphasis and thus lead to more incomplete or 'don't know' answers. On the other hand, a very stringent assessor may elicit similar effects through behaviour that reduces the respondent's willingness to cooperate.

Overall, the impact of such interviewer biases seems to be small or not well-investigated (Hagenaars & Heinen, 1982). Probably, the more an assessor complies with professional standards and is not socially involved, the fewer biases will occur (Hyman, 1975).

## DIFFERENTIAL ISSUES

### Rater- versus Dyad-Specific Biases

Assessor's biases contain both rater-specific and dyad-specific components (Hyo & Kerns, 1999). In the former, the error variance is attributable to the assessor alone (e.g. a rater who generally tends to leniency when judging coworkers). In the latter, it is due to a specific relation between the assessor and the assessee (e.g. a teacher who judges a difficult student more negatively than he should). Rater-specific biases are a minor problem when only one assessor compares individuals on one dimension, because the error is the same across all judgements. It becomes more problematic when there are several assessors with different biases. The same holds for complex assessments by a single assessor who confounds specific information due to a halo effect.

Dyad-specific biases seem to be more powerful. Because they are less general, they are also more difficult to detect and correct. Neither rater-specific nor dyad-specific biases need to be stable. They may fluctuate over time and situations according to current influences such as emotional state, task involvement, or organizational factors.

### Moderating Factors

The magnitude of biases also depends on what information is gathered. Their impact is relatively small (4% of variance) when ratings are based on explicit and objective criteria such as behaviour frequencies (Hyo & Kerns, 1999). However, it is much stronger (47% of variance) when assessors rate global trait characteristics. Training of assessors is another important moderator. When

they are well-trained, less than 10% of variance is attributable to assessor's biases, but with minimal training, these sources may account for over 50% (Hyo & Kerns, 1999). Furthermore, rater agreement varies according to the observed behaviour samples. If assessors refer to different samples, they will agree less. However, as mentioned before, such interrater differences may indicate true variance rather than biases (reliability-validity trade-off; Scullen et al., 2000). For example, employees behave differently with their bosses than with their colleagues.

## DETECTING AND REDUCING ASSESSOR'S BIASES

The valid assessment of an assessor's biases is a prerequisite for intervention. Unfortunately, there is little systematic and practice-oriented research on this issue.

One strategy is to reconstruct the errors from the assessor's judgements. If he rated specific dimensions in various persons and other assessors did the same, inter- and intrarater comparisons are possible. Different frequency distributions, means, variances, and correlations between variables may indicate halo, leniency, extremity, or other errors. However, as mentioned above, this is only possible when assessors work on the same samples of data. Another strategy is to compare the individual judgements with objective data structures. Brunswik's lense model can be used to compare regression weights between the respective data and both the assessor's judgement and an objective criterion. For example, a teacher may place too much weight on verbal intelligence in predicting student achievement. Similarly, configurational analyses can be used to detect biases in nonlinear data structures.

Such reconstructions require a great deal of analogue data and judgements. If these are not available, one can try to assess directly what goes on in the assessor's mind (e.g. by the method of thinking aloud or analysing subjective theories by using structure-placing or repertory grid techniques). However, it is questionable how far these approaches can detect automatized and unconscious mental processes. Verbal ambiguities and social desirability effects must also be expected.

Assessor's biases may further be reduced through supervision by neutral experts or team feedback

sessions. These approaches are most common in clinical contexts but can also be applied in other fields of psychological assessment.

Last, but not least, assessor's biases can be reduced by a systematic organization and quality management of the whole assessment process. This includes, for example, standardized procedures, detailed behavioural indicators of categories, intensive training of assessors, random-routine check of assessment quality, re-analysable data registration (e.g. video recordings), adequate time-spacing of judgements, techniques that enhance systematic comparisons (e.g. in pairs vs. ratings), the clear distinction between data description and interpretation, and explicit rules for data integration.

## CONCLUSION AND FUTURE PERSPECTIVES

Assessor's biases are important sources of error variance. Although these biases cannot be eliminated completely in the human process of assessment, they can be reduced substantially. For example, this is possible by following the Guidelines for the Assessment Process recently proposed by a Task Force of the European Association of Psychological Assessment (Fernández-Ballesteros et al., 2001).

## References

- Betz, A.L., Gannon, K.M. & Skowronski, J.J. (1992). The moment of tenure and the moment of truth: when it pays to be aware of recency effects in social judgements. *Social Cognition*, 10(4), 397–413.
- Bernardin, H.J., Cooke, D.K. & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, 85(2), 232–236.
- Elashoff, J.D. & Snow, E. (1971). *A Case Study in Statistical Inference: Reconsideration of the Rosenthal-Jacobson Data on Teacher Expectancy*. Stanford: Stanford University Press.
- Fernandez-Ballesteros, R., De Bruyn, E.E.J., Godoy, A., Hornke, L.F., Ter Laak, J., Vizcarro, C., Westhoff, K., Westmeyer, H. & Zaccagnini, J.L. (2001). Guidelines for the Assessment Process (GAP): a proposal for discussion. *European Journal of Psychological Assessment*, 17(3), 187–200.
- Gigerenzer, G. & Selten, R. (Eds.) (2001). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.
- Guilford, J.P. (1954). *Psychometric Methods* (2nd ed.). New York: McGraw-Hill.
- Hagenaars, J.A. & Heinen, T.G. (1982). Effects of role-independent interviewer characteristics on responses. In Dijkstra, W. & van der Zouwen, J. (Eds.), *Response Behaviour in the Survey – Interview* (pp. 91–130). London: Academic Press.
- Hyman, H.H., Cobb, W.J., Feldman, J.J., Hart, C.W. & Stember, C.H. (1975). *Interviewing in Social Research*. Chicago: University of Chicago Press.
- Hyot, W.T. & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: a meta-analysis. *Psychological Methods*, 4(4), 403–424.
- Lösel, F. (2001) Risk/need assessment and prevention of antisocial development in young people. In Corrado, R., Roesch, R., Hart, S.D. & Gierowski, J.K. (Eds.), *Multiproblem Violent Youth*. Amsterdam: IOS Press (in press).
- Murphy, K.R., Jako, R.A. & Anhalt, R.L. (1993). Nature and consequences of halo error: a critical analysis. *Journal of Applied Psychology*, 78(2), 218–225.
- Saal, F.E., Downey, R.G. & Lahey, M.A. (1980). Rating the ratings: assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413–428.
- Scullen, S.E., Mount, M.K. & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6), 956–970.

Friedrich Lösel and Martin Schmucker

## Related Entries

ITEM BIAS, CLINICAL JUDGEMENT, PROCESS (THE ASSESSMENT PROCESS)



Children are attached, if they tend to seek proximity to and contact with a specific caregiver in times of stress arising from factors such as

distress, illness, or tiredness (Bowlby, 1984). Attachment is a major developmental milestone in the child's life, and it remains an important

issue throughout the lifespan. In adulthood, attachment representations shape the way adults feel about the strains and stresses of intimate relationships, including parent–child relationships, and the way in which the self in relation to important others is evaluated. Attachment theory is a special branch of Darwinian evolution theory, and the need to become attached to a protective conspecific is considered one of the primary needs in the human species. Attachment theory is built upon the assumption that children come to this world with an inborn inclination to show attachment behaviour – and this inclination would have had survival value, or better: would increase ‘inclusive fitness’ – in the environment in which human evolution originally took place. Because of its ethological basis, assessment of attachment implies careful and systematic observations of verbal and non-verbal behaviour.

### ASSESSMENT OF ATTACHMENT IN INFANTS

Attachment to a protective caregiver helps the infant to regulate his or her negative emotions in times of stress and distress, and to be able to explore the environment even if it is somewhat frightening. The idea that children seek a balance between the need for proximity to an attachment figure and the need to explore the wider environment is fundamental to the various attachment measures, such as the *Strange Situation* procedure (SSP; Ainsworth et al. 1978) and the *Attachment Q-Set* (AQS; Vaughn & Waters, 1990) (see Table 1). Ainsworth and her colleagues observed one-year-old infants with their mothers in a standardized stressful separation procedure, and used the reactions of the infants to their reunion with the caregiver after a brief separation to assess the amount of trust the children had in the accessibility of their attachment figure.

Table 1. Attachment measures

Attachment measure	12–24 months	24–48 months	12 years and older
Strange Situation	X		
Attachment Q Sort	X	X	
Adult Attachment Interview			X

The SSP consists of eight episodes, of which the last seven ideally take three minutes. Each episode can however be curtailed when the infant starts crying. *Episode One* begins when the experimenter leads caregiver and child into an unfamiliar playroom. *Episode Two* is spent by the caregiver together with the child in the playroom. In *Episode Three* an unfamiliar adult (the ‘stranger’) enters the room and after a while starts to play with the infant. *Episode Four* starts when the caregiver departs, and the infant is left with the stranger. In *Episode Five* the caregiver returns, and the stranger unobtrusively leaves the room immediately after reunion. *Episode Six* starts when the caregiver leaves again: the infant is alone in the room. In *Episode Seven* the stranger returns. In *Episode Eight* the caregiver and the infant are reunited once again, and the stranger leaves unobtrusively immediately after reunion.

The Strange Situation procedure has been used with mothers, fathers, and other caregivers. Infants usually are between 12 and 24 months of age. For pre-schoolers, the same SSP is used, but the rating system for classifying the children is different and still is in the process of validation (Cassidy & Marvin, 1992). On the basis of infants’ reactions to the reunion with the caregiver, three patterns of attachment can be distinguished. Infants who actively seek proximity to their caregivers upon reunion, communicate their feelings of stress and distress openly, and then readily return to exploration are classified as secure (B) in their attachment to that caregiver. Infants who seem not distressed, and ignore or avoid the caregiver following reunion are classified as insecure-avoidant (A). Infants who combine strong proximity seeking and contact maintaining with contact resistance, or remain inconsolable, without being able to return to play and explore the environment, are classified insecure-ambivalent (C).

An overview of all American studies with non-clinical samples (21 samples with a total of 1584 infants, studies conducted in the years 1977–1990) shows that about 67% of the infants are classified secure, 21% are classified as insecure-avoidant, and 12% are classified insecure-ambivalent (Van IJzendoorn, Goldberg, Kroonenberg, & Frenkel, 1992). The Strange Situation classifications have been demonstrated to be valid. For example, secure infants have

more sensitive parents than insecure infants (in 66 studies with more than 4000 infants, DeWolff & van IJzendoorn, 1997). Furthermore, secure infants have more satisfactory peer relations, and they develop better language skills (Cassidy & Shaver, 1999). The SSP also shows discriminant validity in comparison with temperament. One of the most powerful demonstrations of the absence of a causal link between attachment and temperament is the lack of correspondence between a child's attachment relationship to his or her mother, and the same child's relationship to his or her father.

The concept of 'disorganized' attachment emerged from the systematic inspection of about 200 cases from various samples that were difficult to classify in one of the three organized attachment categories (Main & Solomon, 1986). In particular in studies on maltreated infants, the limits of the traditional Ainsworth et al. (1978) coding system became apparent because many children with an established background of abuse or neglect nevertheless had to be forced into the secure category. Common denominator of the anomalous cases appeared to be the (sometimes momentary) absence of an organized strategy to deal with the stress of the SSP. Disorganized attachment can be described as the breakdown of an otherwise consistent and organized strategy of emotion regulation. Whether secure or insecure, every child may show disorganization of attachment depending on the earlier child-rearing experiences. Maltreating parents are supposed to create disorganized attachment in their infants because they confront their infants with a pervasive paradox: they are potentially the only source of comfort for their children, whereas at the same time they frighten their children through their unpredictable abusive behaviour. Disorganization of attachment occurs in about 15% of non-clinical cases, where associations with parental unresolved loss have been found, and it is considered a major risk factor in the development of child psychopathology.

## ATTACHMENT IN TODDLERS AND PRESCHOOLERS

Although the SSP has become remarkably popular and successful, it has been a drawback that attachment research was almost exclusively

dependent on a single procedure for the measurement of attachment. Waters and his co-workers introduced another method for assessing attachment security in infants and toddlers, i.e. the *Attachment Q-Sort* (AQS). The AQS consists of 90 cards. On each card a specific behavioural characteristic of children between 12 and 48 months of age is described. The cards can be used as a standard vocabulary to describe the behaviour of a child in the natural home-setting, with special emphasis on secure-base behaviour (Vaughn & Waters, 1990). After several hours of observation the observer ranks the cards into nine piles from 'most descriptive of the subject' to 'least descriptive of the subject'. The number of cards that can be put in each pile is fixed, i.e. 10 cards in each pile. By comparing the resulting Q-sort with the behavioural profile of a 'prototypically secure' child as provided by several experts in the field of attachment theory, a score for attachment security can be derived.

The AQS has some advantages over the SSP. First, it can be used for a broader age range (12–48 months) than the SSP. Moreover, AQS scores for attachment security are based on observation of the child's secure-base behaviour in the home and may therefore have higher ecological validity. Furthermore, because the application of the AQS does not require the artificial induction of stress used in the SSP, the method can be applied in cultures and populations in which standard application of the SSP has proved to be somewhat complicated. Because the AQS is less intrusive than the SSP, it may be used more frequently with the same child, for example in repeated measures designs, in interventions studies, and in studies on children's attachment networks. Lastly, the application of the AQS in divergent cultures or populations may be attuned to the specific prototypical secure-base behaviour of the children from those backgrounds.

When the AQS is sorted by a trained observer it shows an impressive predictive validity. In particular, the observer AQS is strongly correlated with sensitive responsiveness. At the same time, it should be noted that the association between observer AQS security and SSP security is rather modest (Van IJzendoorn, Vereijken, & Riksen-Walraven, in press). The AQS and the SSP may therefore not measure the same construct, or they may be indexing different dimensions of the

same construct. Support for the validity of the AQS as sorted by the mother is less convincing. The association between the mother AQS and the SSP is disappointingly weak, and the instrument surprisingly shows a stronger association with temperament (van IJzendoorn et al., in press). Mothers of insecure children may lack the observational skills that are necessary for an unbiased registration of secure-base behaviours in their children.

In this contribution three assessment procedures are discussed that play a central role in attachment theory and research. The *Strange Situation Procedure* (SSP; Ainsworth et al., 1978) has been developed to assess attachment security of infants with their parents or other caregivers in a laboratory playroom. The *Attachment Q Sort* (AQS; Vaughn & Waters, 1990) is an instrument to observe secure-base behaviour and attachment security in children from 12–48 months at home. The *Adult Attachment Interview* (AAI; Main, Kaplan, & Cassidy, 1985) is a semi-structured interview with a coding system (Main & Goldwyn, 1994) to assess adolescent and adult mental representations of attachment. We start with a brief discussion of the theoretical background of these assessment tools.

## ASSESSMENT OF ATTACHMENT IN ADOLESCENCE AND ADULTHOOD

Attachment experiences are supposed to become crystallized into an internal working model or representation of attachment (Bowlby, 1984), which Main, Kaplan, and Cassidy defined as ‘a set of rules for the organisation of information relevant to attachment and for obtaining or limiting access to that information’ (1985, pp. 66–67). They developed an interview-based method of classifying a parent’s mental representation of attachment, the *Adult Attachment Interview* (AAI). The AAI is a semi-structured interview that probes alternately for general descriptions of attachment relationships, specific supportive or contradicting memories, and descriptions of the current relationship with one’s parents. The interview can be administered to parents, professional caregivers, and older adolescents, and stimulates respondents to both retrieve attachment-related autobiographical memories and evaluate these memories from

their current perspective. For example, subjects are asked, which five adjectives describe their childhood relationship with each parent, and what concrete memories or experiences led them to choose each adjective.

The AAI lasts about an hour and is transcribed verbatim. Interview transcripts are rated for security of attachment as derived from the subjects’ present discussion of their attachment biographies (Hesse, 1999). The coding of the interviews is not based primarily on reported events in childhood, but rather on the coherency with which the adult is able to describe and evaluate these childhood experiences and their effects. The interview, therefore, does not assess the actual quality of childhood attachment relationships, and a secure representation of attachment is not incompatible with an insecure attachment history throughout childhood. This is a major difference with questionnaires that ask for descriptions of the relationship with parents or parent’s parenting, in which descriptions of childhood experiences are decisive and taken for granted. Instead, the AAI takes into account that retrospection is not necessarily reliable, and that repression and idealization do take place. Hesse (1999) has suggested that the central task presented to the subject is that of producing and reflecting upon attachment-related memories while *simultaneously* maintaining coherent discourse with the interviewer.

The coding system of the AAI (Main & Goldwyn, 1994) includes scales for inferred childhood experiences with parents (e.g. loving, rejecting, role-reversing) and scales for state of mind with respect to attachment (e.g. anger, idealization, insistence on lack of recall, coherency). The scale scores for state of mind are of overriding importance when it comes to classification of an interview, in one out of three main categories. *Autonomous* or *secure* adults are able to describe their attachment-related experiences coherently, whether these experiences were negative (e.g. parental rejection or overinvolvement) or positive. They tend to value attachment relationships and to consider them important for their own personality. *Dismissing* adults tend to devalue the importance of attachment experiences for their own lives or to idealize their parents without being able to illustrate their positive evaluations with concrete events demonstrating secure interaction. They often appeal to lack of memory of childhood



experiences. *Preoccupied* adults are still very much involved and preoccupied with their past attachment experiences and are therefore not able to describe them coherently. They may express anger or passivity when discussing current relationships with their parents. Dismissing and preoccupied adults are both considered insecure. Some adults indicate through their incoherent discussion of experiences of trauma (such as maltreatment, or the loss of an attachment figure) that they have not yet completed the process of mourning. They receive the additional classification *Unresolved*, which is superimposed on their main classification. In a meta-analysis on 33 studies, the distribution of non-clinical mothers was as follows: 24% dismissing, 58% autonomous, and 18% preoccupied mothers (Van IJzendoorn & Bakermans-Kranenburg, 1996). About 19% of the mothers were additionally classified as unresolved. Fathers and adolescents showed about the same distribution of AAI classifications. Clinical respondents, however, showed highly deviating distributions, with a strong overrepresentation of insecure attachment representations. Systematic relations between clinical diagnosis and type of insecurity could not be established.

The test-retest reliability of the AAI has been established in several studies, and the same is true of the AAI's discriminant validity. AAI classifications turned out to be independent of respondents' IQ, social desirability, temperament, and general autobiographical memory abilities (for a review, see Hesse, 1999). The predictive validity of the AAI has been thoroughly tested in a large number of studies in different countries, and the results can best be described by meta-analytic findings. First, the AAI appears to be predictive of parent's sensitive responsiveness. Autonomous parents are more responsive to their child's attachment signals and needs than insecure parents (Van IJzendoorn, 1995). Second, in several (cross-sectional as well as longitudinal) studies parents' representations of attachment were related to the security of the parent-child attachment relationship as measured through the Strange Situation procedure. Autonomous parents tended to have secure children, dismissing parents had insecure-avoidant children, preoccupied parents had insecure-ambivalent children, and parents with unresolved loss or other trauma more often had disorganized children (Van IJzendoorn,

1995). In longitudinal studies covering the first 15 to 20 years of life, the infant SSP classifications have been found to predict the later AAI classifications when major changes in life circumstances were absent (Waters, Hamilton, & Weinfield, 2000).

## CONCLUSIONS AND FUTURE PERSPECTIVES.

We conclude that the Strange Situation Procedure, the Attachment Q Sort, and the Adult Attachment Interview have proven to be invaluable tools for testing empirical hypotheses. They have helped to advance attachment theory far beyond Bowlby's first draft some thirty years ago. During the past ten years or so, several other attachment measures have been developed, mostly based on the same construction principles that guided the development of the SSP, AQS, and AAI (Cassidy & Shaver, 1999). Some measures mirror the SSP and focus on attachment in pre-schoolers (The Preschool Assessment of Attachment), others involve projective techniques for preschoolers and older children, such as the SAT, drawings or photographs, or doll play. Other measures are adaptations of the AAI and cover younger (adolescent) age ranges or different representational dimensions (working model of the child; working model of caregiving). Self-report paper-and-pencil measures have been proposed for assessment of attachment in adolescence or adulthood, as well as interview measures for partner relationships. These alternative attachment measures are still in the process of validation, and do not yet present the psychometric qualities that SSP, AQS, and AAI have shown to possess (Cassidy & Shaver, 1999). In the near future, more data will become available on the reliability and validity of these promising measures. They may help to investigate attachment across the lifespan, in various contexts, populations, and cultures.

## References

- Ainsworth, M.D.S., Blehar, M.C., Waters, E. & Wall, S. (1978). *Patterns of Attachment*. Hillsdale, NJ: Lawrence Erlbaum.
- Bowlby, J. (1984). *Attachment and Loss. Attachment*, Vol. 1 (2nd ed.). London: Penguin.

- Cassidy, J., Marvin, R.S. & MacArthur Working Group on Attachment. (1992). Attachment organization in three- and four-year-olds: procedures and coding manual. Unpublished coding manual. Pennsylvania State University.
- Cassidy, J. & Shaver, P.R. (1999). *Handbook of Attachment. Theory, Research, and Clinical Applications*. New York: Guilford.
- DeWolff, M.S. & Van IJzendoorn, M.H. (1997). Sensitivity and attachment: a meta-analysis on parental antecedents of infant-attachment. *Child Development* 68, 571–591.
- Hesse, E. (1999). The Adult Attachment Interview: historical and current perspectives. In Cassidy, J. & Shaver, P.R. (Eds.). *Handbook of Attachment. Theory, Research, and Clinical Applications* (pp. 395–433). New York: Guilford.
- Main, M. & Goldwyn, R. (1994). *Adult Attachment Classification System*. Department of Psychology, University of California at Berkeley. Unpublished manuscript.
- Main, M., Kaplan, N. & Cassidy, J. (1985). Security in infancy, childhood, and adulthood: a move to the level of representation. In Bretherton, I. & Waters, E. (Eds.), *Growing Points of Attachment Theory and Research* (pp. 66–104). Society for Research in Child Development.
- Main, M. & Solomon, J. (1986). Discovery of an insecure-disorganized/disoriented attachment pattern. In Brazelton, T.B. & Yogman, M.W. (Eds.), *Affective Development in Infancy* (pp. 95–124). Norwood, NJ: Ablex.
- Van IJzendoorn, M.H. (1995). Adult attachment representations, parental responsiveness, and infant attachment. A meta-analysis on the predictive validity of the Adult Attachment Interview. *Psychological Bulletin*, 117, 387–403.
- Van IJzendoorn, M.H. & Bakermans-Kranenburg, M. J. (1996). Attachment representations in mothers, fathers, adolescents, and clinical groups: a meta-analytic search for normative data. *Journal of Consulting and Clinical Psychology*, 64, 8–21.
- Van IJzendoorn, M.H., Goldberg, S., Kroonenberg, P.M. & Frenkel, O.J. (1992). The relative effects of maternal and child problems on the quality of attachment: a meta-analysis of attachment in clinical-samples. *Child Development*, 63, 840–858.
- Van IJzendoorn, M.H., Vereijken, C.M.J.L. & Riksen-Walraven, J.M.A. (1996). Is the Attachment Q-Sort a valid measure of attachment security in young children? In Vaughn, B., Waters, E. & Posada, D. (Eds.), *Patterns of Secure Base Behaviour: Q-sort Perspectives on Attachment and Caregiving in Infancy and Childhood*. Hillsdale, NJ: Erlbaum.
- Vaughn, B.E. & Waters, E. (1990). Attachment behaviour at home and in the laboratory: Q-sort observations and Strange Situation classifications of one-year-olds. *Child Development*, 61, 1965–1973.
- Waters, E., Hamilton, C.E. & Weinfield, N.S. (2000). The stability of attachment security from infancy to adolescence and early adulthood: general introduction. *Child Development*, 71, 678–683.

Marinus van IJzendoorn and  
Marian J. Bakermans-Kranenburg

### Related Entries

PERSONALITY (GENERAL), EMOTIONS, MOTIVATION

# A

---

## ATTENTION

### INTRODUCTION

Attention involves being in a state of alertness, focusing on aspects of the environment that are deemed important for the task at hand, and shutting out irrelevant information. As the task demands change, attention involves the ability to flexibly shift focus to another target. Originally, attention was considered a unitary construct but currently it is conceptualized as a complex process involving (a) distributed neural systems, (b) perceptual, emotional, motivational and motor systems, as well as (c) links to multiple sources of environmental information.

Some commonly studied processes of attention include selecting, sustaining, and shifting. Selection refers to the ability to narrow the field of stimuli to which one attends for the purpose of enhanced processing. Sustained attention refers to the ability to maintain focus and alertness over time. Shifting refers to the ability to change focus of attention to suit one's goals and needs.

Research has focused on visual or auditory attention, although environmental stimuli are perceived through other modalities as well (i.e. touch, smell, taste). In addition, research has focused on attention to the external environment rather than to the internal environment (thoughts

and emotions) since the internal environment is less amenable to objective and reliable methods of assessment.

## WHY IS IT IMPORTANT TO ASSESS ATTENTION?

Attention is central to the ability to function perceptually, cognitively and socially. For that reason it is important to have basic scientific understanding of attention processes and the psychological and environmental conditions that govern the development of attention and its deployment under specific circumstances. With such knowledge in hand, one can design environments that promote optimal attention to important characteristics in those settings.

In addition, it is important to assess attention so as to map out individual differences in the development and use of attention. These differences are mostly in the normal range but may also include deficits that are quite marked as seen in children diagnosed with Attention Deficit Disorder or in adults diagnosed with schizophrenia, depression or substance abuse problems. The assessment of attention is important for parents and teachers who detect difficulties in a child's ability to focus attention and wish to have the child evaluated. Similarly, attention problems may be presented in adults who have suffered head injuries or stroke, and who would need to be evaluated to determine the seriousness of the deficits involved. Diagnosing such deficits is dependent on information about individual differences in attention and on the availability of appropriate assessment tools.

## ASSESSMENT METHODS

Methods have been developed for the assessment of specific aspects of attention, including selective attention, sustained attention, and shifting attention. These methods include performance tests, mapping brain activity during performance of tasks and finally, rating scales. Table 1, lists commonly used performance tasks, the aspects of attention they assess and the contexts in which they are used (clinical or research). Additional information can be found in Barkley (1994). Other tests include Trenerry, Crosson and

DeBoe's Visual Search and Attention Test (VSAT), Miller's California Computerized Assessment Package (CalCAP), Arthur, Barrett and Doverspike's Auditory Selective Attention Test (ASAT), and The Gordon Diagnostic System. Table 2 lists commonly used scales for rating attention.

## FUTURE DIRECTIONS

### Deal with Issues Pertaining to Assessment for the Purpose of Increasing Knowledge about Specific Processes

There is a need to understand to what extent the processes outlined above are really independent rather than different manifestations of the same core. This calls for a more integrated understanding of attention and for the development of a basic assessment battery that could be used when people are referred with problems in attention.

### Checking Ecological Validity

To what extent are the assessments telling something about functioning under some specific environmental conditions but don't generalize to these processes as they operate in every-day, out of the lab environments? Questions remain about the extent to which it is possible to do well on all laboratory assessments but have problems in the every day context. Similarly, is it possible to function well in the everyday environment and yet have problems on laboratory assessments.

### Developing an Attention Battery

The battery would need to be based on normative data and would need to have specified cut off lines between the normal range and problem range. Children would benefit from a routine assessment using such a battery in the same way that they benefit from routine examination of their hearing and vision. Systematically evaluating how children perform in terms of their attention is important since children may have deficits that they mask through idiosyncratic cognitive strategies or by working harder than what would normally be required.

Table 1. Commonly used performance tests

Process Assessment Name		Short Description	Assessed Behaviour	Contexts of Use
I. Selective Attention	Children's checking Task	Symbol cancellation	Number targets identified; Number targets missed; Incorrect identifications	Research
	Digit Symbol/Coding	Wechsler scales subtest	Timed task of correctly indicating which symbol corresponds to a number	Clinical Research
	Stroop Colour-World Interference Test	Naming the ink colour of words that spell a colour different from the ink colour	Time to complete each portion; Number of correct responses	Research Clinical
	The Trail Making Test	Connecting letters and numbers placed randomly on a page	Time to complete each part; Number of errors	Research Clinical
	Children's Embedded Figures Test	Identifying a target figure embedded among non-targets	Mean time to respond; Number of correct responses	Clinical Research
II. Sustained Attention	Posner's Visual-Spatial Selective Attention Test	Responding to targets presented to the left/right visual fields	Difference in reaction time in the presence of valid and invalid cues	Research
	Reaction Time Task(s)	Responding to simple target visual stimuli	Mean reaction time; Variability of response time	Research
	Continuous Performance Test (CPT)	Responding to target stimuli and inhibiting response to on-target stimuli	Response time; Number correct responses; Errors of omission; Errors of commission	Research Clinical
	KABC Hand Movements	Imitating progressively longer sequences of skilled hand movements	Standard score of successful number of sequences	Research Clinical

III. Shifting Attention	Wisconsin Card Sorting Task	Sorting 128 cards containing sets of geometric designs – varying colour, form, number	% of correct; Number of categories achieved; Perseverative errors; Perseverative responses; Non-perseverative responses	Research Clinical
	Halstead-Reitan Neuropsych. Test Battery – Categories Test	Choosing from 1 of 4 choices from a projected stimuli based on a principle	Number of correct responses; Same behaviours as above	Clinical
IV. Numerical Mnemonic Attention	Digit Span	Wechsler scales subtest	Accurate memory for a specific string of numerical stimuli (forward & backward)	Research Clinical
	Arithmetic	Wechsler scales subtest	Correct solutions provided verbally	Research Clinical
V. Physiological Heart Rate processes		Electrodes placed on chest record the electrocardiogram (EKG)	Decrements in heart rate reflect attention	Research
	Cortical Electrophysiology Cerebral blood flow	Electrodes placed on scalp record electroencephalograph (EEG) Blood flow to brain regions is mapped by positron emission tomography (PET)	Large, slow waves indicate lapses in attention during sustained attention task Denser distribution indicates more active metabolism	Research Research

Table 2. Commonly used scales for rating attention

Rating scales
<i>Title</i>
ADD-H Comprehensive Teacher Rating Scale
ADHD Rating Scale
Attention Deficit Disorders Evaluation Scale
Behaviour Assessment System for Children
Child Attention Profile by Edelbrock
Child Behaviour Checklist
Conners' Parent and Teacher Rating Scale – Revised
Hyperactive Behaviour Code

## CONCLUSIONS

Attention is central to cognitive and social functioning and has been the subject of scientific research for decades. It is regulated by neural, perceptual, emotional, motivational and motor systems and influenced by both internal and external stimuli. Because of its central and complex role in behaviour, there are many methods for assessing its various aspects. Despite the long history of interest in the topic, scientists are still working to achieve greater

understanding of attention processes and on developing new assessment tools.

## Further Readings

- Barkley, R.A. (1994). The assessment of attention in children. In Lyon, G.R. (Ed.), *Frames of Reference for the Assessment of Learning Disabilities*. Baltimore: Paul H. Brookes Publishing.
- Ruff, H.A. & Rothbart, M.K. (1996). *Attention in Early Development: Themes and Variations*. New York: Oxford University Press.
- Underwood, G. (Ed.), (1993). *The Psychology of Attention*, Vols. I and II. New York: New York University Press.

Sarah Friedman and Anita Konachoff

## Related Entries

THEORETICAL PERSPECTIVES: COGNITIVE, INTELLIGENCE (GENERAL), AMBULATORY ASSESSMENT, BRAIN ACTIVITY MEASUREMENT, EQUIPMENT FOR ASSESSING BASIC PROCESSES

# A | ATTITUDES

## INTRODUCTION

Evaluation is a fundamental reaction to any object of psychological significance (Jarvis & Petty, 1996; Osgood, Suci, & Tannenbaum, 1957). The present entry reviews some of the major techniques that have been developed to assess these evaluative reactions, or attitudes. A discussion of methods based on explicit evaluative responses – direct and inferred – is followed by a consideration of disguised and implicit assessment techniques. Emphasis is placed on questions of reliability, validity, and practicality.

## EXPLICIT MEASURES OF ATTITUDE

Virtually any response can serve as an indicator of attitude toward an object so long as it is

reliably associated with the respondent's tendency to evaluate the object in question. In contrast to implicit responses, which cannot be easily controlled, explicit evaluative responses are under the conscious control of the respondent. Most explicit attitude measures either rely on direct attitudinal inquiries or infer the respondents' evaluations from their expressions of beliefs about the attitude object.

## Direct Evaluations

*Single-item direct measures.* Laboratory experiments and attitude surveys frequently use single items to obtain direct evaluations of the attitude object. Confronted with the item, 'Do you approve of the way the President is doing his job?' respondents may be asked to express their degree of approval on a five-point scale that

ranges from 'approve very much' to 'disapprove very much'. Such single items can be remarkably good indicators, especially for well-formed attitudes toward familiar objects. They are sometimes found to have quite high levels of reliability and to correlate well with external criteria. For example, the single item, 'I have high self-esteem' (attitude toward the self), assessed on a five-point scale ranging from 'not very true of me' to 'very true of me', was found to have a test-retest reliability of 0.75 over a four-year period, compared to a reliability of 0.88 for the multi-item Rosenberg Self-Esteem Scale (Robins, Hendin, & Trzesniewski, 2001, Study 1). Moreover, the single- and multi-item measures correlated highly with each other, and they had comparable correlations with various external criteria (e.g. self-evaluation of physical attractiveness, extraversion, optimism, life satisfaction).

However, single items do not always exhibit such favourable psychometric properties. They often have low reliabilities and can suffer from limited construct validity. Many attitude objects are multidimensional and a single item can be ambiguous with respect to the intended dimension (e.g. 'religion as an institution' vs. 'religious faith'). Furthermore, single items contain nuances of meaning that may inadvertently affect responses to attitudinal inquiries. An item inquiring whether the United States should *allow* public speech against democracy leads to different conclusions than one asking whether the United States should *forbid* such speech (see Schuman & Presser, 1981). In addition to such framing effects, research has revealed strong context effects in attitudinal surveys. Respondents tend to interpret a given item in light of the context created by previous questions. Thus, responses to questions about satisfaction with life in general and satisfaction with specific aspects of one's life, such as one's work or romantic relationship, are found to be influenced by the order in which these questions are asked (Schwarz, Strack, & Mai, 1991).

*Multi-item direct measures.* It is possible to raise the reliability of a direct attitude measure by increasing the number of questions asked. The Rosenberg Self-Esteem Scale (Rosenberg, 1965), for example, contains 10 items, each a direct inquiry into self-esteem (e.g. 'I feel that I am a person of worth, at least on an equal basis with

others'; 'All in all, I am inclined to feel that I am a failure'). Coefficients of internal consistency and test-retest reliability for this measure are typically quite high (see Robinson, Shaver, & Wrightsman, 1991).

The most frequently employed multi-item direct measure of attitude, however, is the evaluative semantic differential (Osgood et al., 1957). Using large sets of seven-point bipolar adjective scales, Osgood and his associates discovered that evaluative reactions (i.e. attitudes) capture the most important dimension of any object's connotative meaning. Consequently, it is possible to obtain a measure of attitude by asking respondents to rate any construct on a set of bipolar evaluative adjective scales, such as *good-bad*, *harmful-beneficial*, *desirable-undesirable*, *pleasant-unpleasant*, and *useful-useless*. When a sufficient number of such scales is used, the evaluative semantic differential is found to have very high internal consistency and temporal stability. One caveat with respect to the semantic differential has to do with possible 'construct-scale interactions'. Although certain adjective pairs generally indicate evaluation, these adjectives can take on more specific denotative meaning in relation to particular attitude objects. Thus, the adjective pair *sick-healthy* usually reflects evaluation when rating people, but it may be a poor measure of evaluation when respondents are asked to judge the construct 'mental patients'.

### Inferred Evaluations

Although multi-item direct attitude measures exhibit high degrees of reliability, they do not address the problems raised by the multidimensionality of attitude objects, or by framing and context effects, problems that jeopardize the validity of direct evaluations. Several standard attitude-scaling methods, such as Thurstone and Likert scaling, avoid these difficulties by sampling a broad range of responses relevant to the attitude object and then inferring the common underlying evaluation. Whereas responses to items on a Thurstone scale are required to have a curvilinear relation to the overall attitude, the more common Likert method requires that item operation characteristics have a linear or at least monotonic shape (Green, 1954). In practice, an investigator using Likert's method of summated

ratings (Likert, 1932) begins by constructing a large set of items, usually statements of belief, that are intuitively relevant for the attitude object. To illustrate, the following items are part of a Likert scale that was designed to assess attitudes toward illegal immigrants (Ommundsen & Larsen, 1997).

- Illegal aliens should not benefit from my tax dollars.
- There is enough room in this country for everyone.
- Illegal aliens are a nuisance to society.
- Illegal aliens should be eligible for welfare.
- Illegal aliens provide the United States with a valuable human resource.
- We should protect our country from illegal aliens as we would our own homes.

The investigators initially constructed 80 items of this kind. Selection of items that had high correlations with the total score yielded a final 30-item scale. Most Likert scales ask respondents to indicate their degree of agreement with each statement on a five-point scale (*strongly agree, agree, undecided, disagree, strongly disagree*). Responses to negative items are reverse scored and the sum across all items constitutes the measure of attitude. The respondents' attitudes are thus *inferred* from their beliefs about the attitude object (see Fishbein & Ajzen, 1975).

By covering a broad range of issues relevant to the attitude object, multi-item belief-based scales can do justice to the multidimensional nature of the issue under consideration, avoiding the potential ambiguity of direct measures. Furthermore, by including many differently worded questions that appear in unsystematic order, they also avoid idiosyncratic framing and context effects. As a result, standard multi-item attitude scales tend to have high reliability and, in many applications, exhibit high degrees of predictive and construct validity (Ajzen, 1982). Collections of scales designed to assess social and political attitudes can be found in Robinson, Shaver, and Wrightsman (1991, 1999). The obvious disadvantage in comparison to direct attitude assessment lies in the increased time and effort required to develop multi-item inferred attitude scales and in the fact that such scales may not be suitable for large-scale telephone surveys.

## DISGUISED ATTITUDE MEASURES

Notwithstanding the psychometric advantages of inferred attitude measures over direct assessment techniques, all explicit measures – direct and inferred – are subject to response biases that may jeopardize their validity. The most serious of these biases is the tendency to respond to attitudinal inquiries in a socially desirable manner (Paulhus, 1991). This tendency is a particularly severe threat to validity when dealing with such socially sensitive issues as racism and sexism, or with potentially embarrassing topics, such as sexual behaviour or tax evasion. Various methods have been developed in attempts to overcome or at least alleviate social desirability responding.

One approach assumes that individuals differ in their tendency to provide socially desirable responses. Scales are available to assess a person's general tendency to respond in a socially desirable manner (see Paulhus, 1991), and these scales can be used to select attitude items that are relatively free of general social desirability influences or to statistically remove variance due to individual differences in social desirability responding. Unfortunately, this approach fails to identify socially desirable responses that are not part of a general tendency but rather are unique to a given topic or assessment context.

The problem of social desirability responding arises because the purpose of explicit attitude measures is readily apparent. Other approaches to this problem therefore attempt to reduce the measure's transparency or completely disguise its purpose. In measures of whites' attitudes toward African Americans, for example, item wording has changed over the years to accommodate the changing social climate. The ethnocentrism scale (Adorno, Frenkel-Brunskwik, Levinson, & Snaford, 1950), used in the 1950s, contained such blatantly racist statements as, 'Manual labor and unskilled jobs seem to fit the Negro mentality and ability better than more skilled or responsible work'. About 15 years later, the Multifactor Racial Attitude Inventory (Woodmansee & Cook, 1967) employed more mildly worded items, such as, 'I would not take a Negro to eat with me in a restaurant where I was well known'. The most popular explicit attitude scale used today, the Modern Racism Scale



(McConahay, Hardee, & Batts, 1981), is an attempt at a relatively nonreactive measure that captures the ambivalence many people experience with respect to African Americans: negative feelings that contrast with a desire to live up to ideals of equality and fairness. Among the items on this scale are, 'It is easy to understand the anger of black people in America' and 'Blacks are getting too demanding in their push for equal rights'.

Although less blatant than earlier measures, the Modern Racism Scale is still quite transparent in its attempt to assess attitudes toward African Americans and is thus potentially subject to social desirability responding. The error-choice method (Hammond, 1948) was an early attempt to avoid social desirability responding by disguising the purpose of the measurement and exploiting the tendency of attitudes to bias responses without a person's awareness. Respondents are asked to choose which of two apparently factual items, equidistant from the known state of affairs, is true (e.g. '25% of African Americans attend college' versus '55% of African Americans attend college'). Choice of the low estimate may indicate a more negative attitude, but because the survey is presented as a fact quiz, participants will usually not be aware that their attitudes toward African Americans are being assessed and their responses may thus be uninfluenced by social desirability concerns.

## IMPLICIT MEASURES OF ATTITUDE

Perhaps the most effective way to avoid response biases associated with explicit attitude measures is not to obscure the test's purpose but to observe evaluative responses over which respondents have little or no control.

### Bodily Responses

A variety of physiological and other bodily responses have been considered as possible indicators of evaluation, including facial expressions, head movements, palmar sweat, heart rate, electrical skin conductance (GSR), and constriction and expansion of the pupil (see Petty & Cacioppo, 1983). By and large, measures of this kind have been found to have relatively low

reliability and to be of questionable validity as measures of attitude. The most promising bodily response measure to date is the facial electromyogram (EMG), an electrical potential accompanying the contraction of muscle fibers. Subtle contractions of facial muscles during exposure to attitude-relevant stimuli appear to reveal underlying positive or negative affective states (Petty & Cacioppo, 1983). Relatively few studies have been conducted to test the validity of this method, but even if its validity is confirmed, the facial EMG requires extensive training and complex technology. It is thus not a very practical method for conducting large-scale attitude surveys, although it may be quite useful in a laboratory context.

In a related method, electrodes are attached to various sites and an attempt is made to persuade respondents that physiological responses are being measured and that these responses provide a reliable indication of their true attitudes. Even though no physiological measures are actually taken, respondents believing that their true attitudes are being read by the machine are expected to provide truthful answers to attitudinal inquiries (Jones & Sigall, 1971). Empirical evidence suggests that the 'bogus pipeline' method can indeed help to reduce response biases due to social desirability concerns (Quigley-Fernandez & Tedeschi, 1978). This method, however, again requires a fairly complex laboratory setup.

### Response Latency

Somewhat more practical are methods that rely on response latencies to assess implicit attitudes because the time it takes to respond to an attitudinal inquiry can be assessed with relative ease. The most popular response-latency method is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) which is based on the assumption that evaluative responses or judgements can be activated automatically, outside the respondent's conscious awareness. Participants are asked to respond as quickly as possible to words that signify the attitude object and words with positive or negative valence. When measuring implicit attitudes toward African Americans, for example, the attitude object may be represented by first names recognized as belonging to white or black

Americans (e.g. 'Josh' vs. 'Jamel') and the valenced words by common positive or negative concepts (e.g. 'health' vs. 'grief'). Instructions that require highly associated categories to share a response key tend to produce faster reactions than instructions that require less associated categories to share a response key. Prejudiced individuals would therefore be expected to respond more quickly to combinations of black names with negative words than to combinations of black names with positive words, and they should show the reverse pattern for white names. The discrepancy between the response latencies for the two situations is taken as a measure of implicit acceptance of the association between an attitude object and valenced attributes, thus providing an implicit measure of attitude.

An alternative procedure relies on sequential evaluative priming (Fazio, Jackson, Dunton, & Williams, 1995). Applied to the measurement of racial attitudes, photos of black and white faces may be presented as primes, followed by positive or negative target words. The participant is asked to judge the valence of each target word as quickly as possible. As in the IAT, a low response latency is taken as an indication of a strong association between the valenced word and the category ('black' or 'white') represented by the prime. Thus, if words with negative valence are judged more quickly when they follow a 'black' prime as compared to a 'white' prime, and when the opposite is true for positive words, it is taken as evidence for a negative attitude toward African Americans.

Response-latency measures have been used mainly in attempts to assess implicit racial and sexual stereotypes and prejudice. Test-retest reliabilities of implicit measures have been found to be of moderate magnitude (0.50 to 0.60) over a time span of one hour to three weeks (Kawakami & Dovidio, 2001); they tend to be virtually uncorrelated with corresponding explicit measures (Fazio et al., 1995; Greenwald et al., 1998; Kawakami & Dovidio, 2001), indicating that they indeed tap a different type of attitude; and they tend to reveal prejudice where explicit measures reveal little or none (e.g. Greenwald et al., 1998), suggesting that implicit measures may be subject to less social desirability bias than explicit measures. However, questions have been raised with respect to the

predictive validity of implicit attitude measures. It has been suggested that low response latencies reflect commonly shared and automatically activated stereotypes, but that privately held, explicit beliefs in conflict with the implicit stereotype can override the automatic response in determining actual behaviour (Devine, 1989).

## FUTURE PERSPECTIVES AND CONCLUSIONS

The great effort that has been invested over the years in the development of attitude measurement procedures attests to the centrality of the attitude construct in the social and behavioural sciences. Table 1 summarizes the different types of measures commonly employed in attitude research. Single items are often used with considerable success to assess evaluative reactions to attitude objects, but multi-item instruments that infer attitudes from a broad range of responses to the attitude object tend to yield measures of greater reliability and validity. Implicit attitude measures hold out promise for overcoming people's tendencies to respond in socially desirable ways to explicit attitudinal inquiries, especially when dealing with sensitive issues or with domains in which attitudes are conflicted or ambivalent. However, more work is needed to establish the conditions under which implicit attitude measures are better indicators of response dispositions than are explicit measures. It appears that implicit attitudes may be predictive of actual behaviour in ambiguous contexts where the relevance of an explicit

Table 1. Common Attitude Assessment Techniques

Response type	Representative technique
Explicit – direct	
Single-item	Self-rating scale
Multi-item	Semantic differential
Explicit – infrared	Thurstone scaling, Likert scaling
Disguised	Error-choice method
Implicit	
Bodily responses	GSR, heart rate, papillary response, EMG
Response latency	Implicit association test, evaluative priming

attitude is unrecognized or can be denied, but explicit attitudes may override implicit response tendencies when the relevance of the explicit attitude is readily apparent (see Fiske, 1998 for a discussion of these issues).

## References

- Adorno, T.W., Frenkel-Brunskwik, E., Levinson, D.L. & Snaford, R.N. (1950). *The Authoritarian Personality*. New York: Harper.
- Ajzen, I. (1982). On behaving in accordance with one's attitudes. In Zanna, M.P., Higgins, E.T. & Herman, C.P. (Eds.), *Attitude Structure and Function. The Third Ohio State University Volume on Attitudes and Persuasion*, Vol. 2, (pp. 3–15). Hillsdale, NJ: Erlbaum.
- Devine, P.G. (1989). Stereotypes and prejudice: their automatic and controlled components. *Journal of Personality & Social Psychology*, 56, 5–18.
- Fazio, R.H., Jackson, J.R., Dunton, B.C., & Williams, C.J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–1027.
- Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behaviour: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.
- Fiske, S.T. (1998). Stereotyping, prejudice, and discrimination. In Gilbert, D.T., Fiske, S.T. & Gardner, L. (Eds.), *The Handbook of Social Psychology*, Vol. 2 (4th ed., pp. 357–411). Boston, MA: McGraw-Hill.
- Green, B.F. (1954). Attitude measurement. In Lindzey, G. (Ed.), *Handbook of Social Psychology*, Vol. 1 (pp. 335–369). Reading, MA: Addison-Wesley.
- Greenwald, A.G., McGhee, D.E. & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Hammond, K.R. (1948). Measuring attitudes by error choice: An indirect method. *Journal of Abnormal and Social Psychology*, 43, 38–48.
- Jarvis, W.B.G. & Petty, R.E. (1996). The need to evaluate. *Journal of Personality and Social Psychology*, 70, 172–194.
- Jones, E.E. & Sigall, H. (1971). The bogus pipeline: a new paradigm for measuring affect and attitude. *Psychological Bulletin*, 76, 349–364.
- Kawakami, K. & Dovidio, J.F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin*, 27, 212–225.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 5–53.
- McConahay, J.B., Hardee, B.B. & Batts, V. (1981). Has racism declined in America: it depends on who is asking and what is asked. *Journal of Conflict Resolution*, 25, 563–579.
- Ommundsen, R. & Larsen, K.S. (1997). Attitudes toward illegal aliens: the reliability and validity of a Likert-type scale. *The Journal of Social Psychology*, 135, 665–667.
- Osgood, C.E., Suci, G.J. & Tannenbaum, P.H. (1957). *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.
- Paulhus, D.L. (1991). Measurement and control of response bias. In Robinson, J.P., Shaver, P.R. & Wrightsman, L.S. (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). San Diego, CA: Academic Press.
- Petty, R.E. & Cacioppo, J.T. (1983). The role of bodily responses in attitude measurement and change. In Cacioppo, J.T. & Petty, R.E. (Eds.), *Social Psychophysiology: A Sourcebook* (pp. 51–101). New York: Guilford Press.
- Quigley-Fernandez, B. & Tedeschi, J.T. (1978). The bogus pipeline as lie detector: two validity studies. *Journal of Personality and Social Psychology*, 36, 247–256.
- Robins, R.W., Hendin, H.M. & Trzesniewski, K.H. (2001). Measuring global self-esteem: construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 27, 151–161.
- Robinson, J.P., Shaver, P.R. & Wrightsman, L.S. (Eds.) (1991). *Measures of Personality and Social Psychological Attitudes*. San Diego, CA: Academic Press.
- Robinson, J.P., Shaver, P.R. & Wrightsman, L.S. (Eds.) (1999). *Measures of Political Attitudes. Measures of Social Psychological Attitudes*, Vol. 2. San Diego, CA: Academic Press.
- Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press.
- Schuman, H. & Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. San Diego, CA: Academic Press.
- Schwarz, N., Strack, F. & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: a conversational logic analysis. *Public Opinion Quarterly*, 55, 3–23.
- Woodmansee, J. & Cook, S. (1967). Dimensions of racial attitudes: their identification and measurement. *Journal of Personality and Social Psychology*, 7, 240–250.

Icek Ajzen

## Related Entries

PERSONALITY (GENERAL), INTERESTS, EMOTIONS, ENVIRONMENTAL ATTITUDES AND VALUES.

# A | ATTRIBUTION STYLES

## INTRODUCTION

Shortly after research on attribution theory blossomed, measures were developed to assess attributional style – the presence of cross-situational consistency in the types of attributions people make. Two approaches to measuring attributional style are reviewed here. The first involves global measures that assume attributional style broadly applies across a variety of situations (see Table 1 for a list of the most widely used measures of attributional style). These measures were developed to test predictions from the reformulated theory of learned helplessness depression (Abramson, Seligman, & Teasdale, 1978). The second approach involves more specific measures of attributional style. This approach emerged, in part, from critiques of the cross-situational consistency of the global measures. These measures assess attributional style in more limited contexts such as work, school, and relationships.

## GLOBAL MEASURES OF ATTRIBUTIONAL STYLE

### Dimensional Measures

Dimensional measures of attributional style require respondents to generate causes for hypothetical events and then to rate them along

several attributional dimensions. The Attributional Style Questionnaire (ASQ; Peterson, Semmel, Von Baeyer, Abramson, Metalsky, & Seligman, 1982) is the most widely known. It contains 12 hypothetical events, half describing positive events ('you meet a friend who compliments you on your appearance') and half describing negative events ('you go out on a date and it goes badly'). Events are further divided into an equal number of interpersonal and achievement contexts. The perceived cause of each event is rated along the dimensions of locus (due to the person or the situation), stability (likely or unlikely to occur again), and globality (limited in its influence or widespread) using seven-point scales. Scores can be computed for each dimension within positive and negative events. Factor analyses of the ASQ have supported the presence of distinct attributional styles for negative and positive events (Xenikou, Furnham, & McCarrey, 1997) although results presented by Cutrona, Russell, & Jones, 1985, indicate that each event on the ASQ represents its own factor. However, findings suggest that attributions for negative events are most strongly related to depression (Sweeney, Anderson, & Bailey, 1986). Scores can be further analysed within interpersonal and achievement contexts, a distinction that appears to be more relevant to positive than negative events.

The ASQ has proven to be a valid predictor of depression. People who make internal, stable, and

Table 1. Widely used measures of attributional style

#### Global measures

- Attributional Style Questionnaire (ASQ; Peterson et al., 1982)
- Attributional Style Assessment Test (ASAT; Anderson & Riger, 1991).
- Children's Attributional Style Questionnaire (CASQ; Seligman et al., 1984)
- Content Analysis of Verbatim Explanations (CAVE; Peterson, 1992)

#### Intermediate measures

- Academic Attributional Style Questionnaire (AASQ; Peterson & Barrett, 1987)
- Organizational Attributional Style Questionnaire (OASQ; Kent & Martinko, 1995)
- Relationship Attribution Measure (RAM; Bradbury & Fincham, 1990)

global attributions for negative events tend to be more depressed. However, there are at least four problems with the ASQ. First, internal consistency for the ASQ ranges from adequate to low, especially for the locus dimension. A frequent solution is to combine the three dimensions into a single index to increase reliability, as the dimensions tend to correlate highly with one another. However, this creates a second problem one of interpretation. There are unique predictions for each attributional style dimension; using a composite score prevents valid tests of the model (Carver, 1989). Reivich (1995) advises researchers to analyse ASQ data in terms of both individual dimensions and composite scores. The third problem is also related; the ASQ does not assess the key attributional dimension of controllability. The few studies that included controllability consistently find that it is the most important attributional style dimension, whereas globality is the least important (e.g. Deuser & Anderson, 1995). The fourth problem concerns the affiliation versus achievement distinction; several of the 'achievement' items involve affiliative contexts. The *Expanded Attributional Style Questionnaire* (EASQ; Peterson & Villanova, 1988) uses an identical format to the ASQ and addresses the problem of low reliability by increasing the number of situations included in the measure. However, reliabilities remain modest and the other problems remain unresolved.

The third and fourth versions of the Attributional Style Assessment Test (ASAT-III and ASAT-IV) provide another dimensional assessment of attributional style (Anderson & Riger, 1991). These measures use a format similar to the ASQ but they incorporate a larger number of items (20 for the ASAT-III and 36 for the ASAT-IV), include the controllability dimension, and use success and failure items that mirror each other (e.g. 'succeeded' vs. 'failed' at coordinating an outing for a group of people...). The interpersonal versus noninterpersonal subsets of items are more clearly differentiated than the affiliation versus achievement items of the ASQ. Internal reliabilities at the subscale level tend to be weak to modest, in the 0.5–0.6 range; collapsing across situation types (e.g. ignoring the interpersonal vs. non-interpersonal distinction) yields somewhat larger alphas. These scales have successfully predicted

depression, loneliness, and shyness as well as depressive-like motivational deficits in laboratory settings. Furthermore, this body of work has demonstrated the importance of assessing attributional styles separately for interpersonal and noninterpersonal situations. Finally, this work has shown substantial correlations between attributional styles for successful events and depression (and loneliness and shyness).

Several other dimensional measures of attributional style use the same basic approach as the ASQ and ASAT. The *Balanced Attributional Style Questionnaire* (BASQ; Feather & Tiggemann, 1984) uses a format similar to the ASQ but, like the ASAT, the positive and negative items mirror one another. The scales have moderate reliabilities and correlate with depression, self-esteem, and protestant work ethic. The *Real Events Attributional Style Questionnaire* (REASQ; Norman & Antaki, 1988) requires that respondents generate the positive and negative events for which they then make attributions. This may yield a better prediction of depression, but the loss of item standardization creates other problems.

### Forced-Choice Measures

Forced-choice measures have respondents select a cause from a list of potential explanations. One benefit is that this method may more accurately mirror how people typically select a cause (i.e. without thinking about dimensions). Also, the types of causes in the list can be restricted to only those attributions of theoretical interest. Forced-choice measures also require less time to complete.

The ASAT-I and ASAT-II use this forced choice format. Respondents are provided with a number of hypothetical situations (20 for the ASAT-I and 36 for the ASAT-II). On the ASAT-I, the listed types of causes are strategy, ability, effort, personality traits, mood, and circumstances. ASAT-II includes only strategy, effort, and ability causes. The number of times a particular cause is selected is summed to create a measure of attributional style for that dimension. Kuder–Richardson (K-R 20) reliabilities for the subscales tend to be in the low to moderate range. Correlations with loneliness and depression have established the validity of these

scales in both U.S. and Mainland China college student populations (Anderson, 1999).<sup>1</sup>

### Measures For Children

The *Children's Attributional Style Questionnaire* (CASQ; Seligman et al., 1984) was developed to allow researchers to study attributional style in children ages 8–13. The CASQ includes 48 items divided equally between positive (“You get an ‘A’ on a test”) and negative events (“You break a glass”). The scale uses both a forced choice and a dimensional approach. Respondents select between two possible causes for the event, and each option represents the presence or absence of one attribution dimension (for example, an internal or external cause). Attributions for each dimension are computed by summing the number of internal, stable, or global responses. Scores similar to the ASQ can then be computed. Internal consistency of the CASQ is low to adequate and improves when the separate dimensions are combined into a single composite.

### Content Analysis Measure

The *Content Analysis of Verbatim Explanations* (CAVE; Peterson, 1992) technique assesses attributional style through a content analysis of an individual's writing. This allows analysis of ecologically valid events without requiring the participant to complete a questionnaire. The CAVE can also be applied to historical data, and it has established the stability of attributional style over a 52-year period (Burns & Seligman, 1989). Coders first extract causal explanations from a text, then rate them along the dimensions of locus, stability, and globality. Inter-rater reliability for the CAVE technique is satisfactory, and internal consistency has been reported as low to adequate. More standard questionnaire measures of attributional style may be better predictors of depression, but the CAVE technique has proven useful when written content is all that is available.

## INTERMEDIATE MEASURES OF ATTRIBUTIONAL STYLE

Global measures of attributional style assume a high degree of cross-situational consistency in the

types of attributions people make. However, several studies have questioned this assumption. Cutrona et al. (1985) found that the ASQ was a poor predictor of attributions for actual events, suggesting that situational factors may play a more important role in predicting attributions. Factor analyses by Cutrona et al. (1985) suggest that there is little cross-situational consistency in global measures of attributional style. Intermediate measures of attributional style address this problem by limiting the situations about which an explanatory style is being assessed. Increased specificity should increase the ability of such measures to predict actual attributions. The ASAT's emphasis on four situation types (success/failure by interpersonal/noninterpersonal) is one approach to increasing specificity. Other research on this issue has been mixed, however (Henry & Campbell, 1995), suggesting that further work is needed to establish the appropriate level of specificity in attributional style measures.

### Academic Settings

Two measures have been used to assess attributional style in academic settings. The *Academic Attributional Style Questionnaire* (AASQ; Peterson & Barrett, 1987) uses the same format as the ASQ and contains descriptions of 12 negative events that occur in academic settings. The measure has demonstrated high internal consistency, and findings suggest that students who make internal, stable, and global attributions for negative events tend to do more poorly in classes. Henry and Campbell (1995) also developed a measure of attributional style for academic events. Their measure contains 20 items, equally divided between positive and negative events. The measure displayed adequate to good reliability and also predicted academic performance.

### Work Settings

The *Organizational Attributional Style Questionnaire* (OASQ; Kent & Martinko, 1995) was developed to assess attributional style for negative events in a work setting. The format is similar to that of the ASQ, and the measure contains descriptions of 16 negative events that can occur in a work setting. After writing down an explanation for the event, respondents rate the

explanation along the dimensions of internal locus, external locus, stability, controllability, globality, and intentionality. The internal consistency for the scale is moderate to good.

## Relationships

Several different types of intermediate attributional style measures have been developed for measuring attributions in the context of relationships. The *Relationship Attribution Measure* (RAM; Bradbury & Fincham, 1990) assesses the types of attributions people make for a spouse's negative behaviour. Respondents read a hypothetical negative action by their partner and rate the causes of that event along six dimensions: locus, stability, globality, and responsibility (intent, selfishness, and blame). Researchers can use either a four- or eight- item version. A composite of all attributional dimensions displays high internal consistency and predicts marital satisfaction. Partners who attribute negative partner behaviour to internal, stable, and global causes are more likely to be dissatisfied with the relationship. Fincham has also developed a version of the RAM for use with children to assess attributions for parent-child interactions. The *Children's Relationship Attribution Measure* (CRAM; Fincham, Beach, Arias, & Brody, 1998) uses a format similar to the RAM, and contains descriptions of two negative events.

## CONCLUSIONS

### Future Research

Measures of attributional style have generated several issues which require additional research. The first issue involves level of specificity. Many studies question the presence of a global attributional style, and it is not clear if intermediate measures provide a satisfying solution to this problem. Additional research is needed to resolve these issues. Furthermore, attributional style measures typically suffer from poor reliability. New measures need to be developed to address this shortcoming. Finally, more research is needed on the controllability dimension of attributional style and on the unique contributions of the various attributional dimensions.

## Using Attributional Style Measures

There are numerous ways of measuring attributional style, each with particular strengths and weaknesses. In deciding which scale to use, the researcher needs to carefully consider the specific goals of the research project, and then pick the tool that best meets the needs of that project. The modest reliabilities of these scales suggests that considerable attention be paid to sample size and power.

## Notes

- 1 The various ASAT scales, as well as Chinese versions of that ASAT-I, the Beck Depression Inventory, and the Revised UCLA loneliness scales, can be downloaded from the following web site: [psych-server.iastate.edu/faculty/caa/Scales/Scales.html](http://psych-server.iastate.edu/faculty/caa/Scales/Scales.html)

## References

- Abramson, L.Y., Seligman, M.E.P. & Teasdale, J. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, 87, 49–74.
- Anderson, C.A. (1999). Attributional style, depression, and loneliness: a cross-cultural comparison of American and Chinese students. *Personality and Social Psychology Bulletin*, 25, 482–499.
- Anderson, C.A., & Riger, A.L. (1991). A controllability attributional model of problems in living: dimensional and situational interactions in the prediction of depression and loneliness. *Social Cognition*, 9, 149–181.
- Bradbury, T.N., & Fincham, F.D. (1990). Attributions in marriage: review and critique. *Psychological Bulletin*, 107, 3–33.
- Burns, M.O. & Seligman, M.E.P. (1989). Explanatory style across the life span: evidence for stability over 52 years. *Journal of Personality and Social Psychology*, 56, 471–477.
- Carver, C.S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology*, 56, 577–585.
- Cutrona, C.E., Russell, D. & Jones, R.D. (1985). Cross-situational consistency in causal attributions: does attributional style exist? *Journal of Personality and Social Psychology*, 47, 1043–1058.
- Deuser, W.E. & Anderson, C.A. (1995). Controllability attributions and learned helplessness: some methodological and conceptual problems. *Basic and Applied Social Psychology*, 16, 297–318.

- Feather, N.T. & Tiggemann, M. (1984). A balanced measure of attributional style. *Australian Journal of Psychology*, 36, 267–283.
- Fincham, F.D., Beach, S.R.H., Arias, I. & Brody, G.H. (1998). Children's attributions in the family: the children's relationship attribution measure. *Journal of Family Psychology*, 12, 481–493.
- Henry, J.W. & Campbell, C. (1995). A comparison of the validity, predictiveness, and consistency of a trait versus situational measure of attributions. In Marinko, M.J. (Ed.), *Attribution Theory: An Organizational Perspective* (pp. 35–52). Delray Beach, FL: St. Lucie Press.
- Kent, R. & Martinko, M. (1995). The development and evaluation of a scale to measure organizational attributional style. In Martinko, M. (Ed.), *Attribution Theory: An Organizational Perspective* (pp. 53–75). Delray Beach, FL: St. Lucie Press.
- Norman, P.D. & Antaki, C. (1988). Real events attributional style questionnaire. *Journal of Social and Clinical Psychology*, 7, 97–100.
- Peterson, C. (1992). Explanatory style. In Smith, Charles P. & Atkinson, John W. (Eds.), *Motivation and Personality: Handbook of Thematic Content Analysis*. (pp. 376–382). New York: Cambridge University Press.
- Peterson, C. & Barrett, L. (1987). Explanatory style and academic performance among university freshman. *Journal of Personality and Social Psychology*, 53, 603–607.
- Peterson, C., Semmel, A., Von Baeyer, C., Abramson, L., Metalsky, G.I. & Seligman, M.E.P. (1982). The attributional style questionnaire. *Cognitive Therapy and Research*, 3, 287–300.
- Peterson, C. & Villanova, P. (1988). An expanded attributional style questionnaire. *Journal of Abnormal Psychology*, 97, 87–89.
- Seligman, M.P. et al. (1984). Attributional style and depressive symptoms among children. *Journal of Abnormal Psychology*, 93, 235–238.
- Sweeney, P., Anderson, K. & Bailey, S. (1986). Attributional style in depression: a meta-analytic review. *Journal of Personality and Social Psychology*, 50, 974–991.
- Xenikou, A., Furnham, A. & McCarrey, M. (1997). Attributional style for negative events: A proposition for a more reliable and valid measure of attributional style. *British Journal of Psychology*, 88, 53–69.

Robert M. Hessling, Craig A. Anderson and Daniel W. Russell

### Related Entries

PERSONALITY (GENERAL), COGNITIVE STYLES, MOTIVATION, IRRATIONAL BELIEFS.

# AUTOBIOGRAPHY

## INTRODUCTION

Autobiography constitutes a critical resource for psychological assessment and yet a complex challenge to it. The essence of this challenge lies in the fact that autobiography can be seen as both a focus of assessment and a means of conducting it. Since autobiography does not lend itself to assessment by instruments or scales, the sections in this entry will focus on general issues associated with the defining, assessing, and researching of autobiography, as well as on future developments concerning it.

## DEFINING AUTOBIOGRAPHY

Autobiography is a narrative accounting of a person's life as interpreted or articulated by the

person him or herself. It is a self-report by which a person expresses, explains, or explores his or her subjective experience over time. It thus represents a route to what it means and feels like to *be* that person, on the inside. Such a definition distinguishes immediately between autobiography and biography (an account of a life, presumably with greater objectivity, by someone else.) An equivalent term for autobiography would be life *story*. This can in turn be distinguished from life *history*, or indeed *case history*, which is an account of a life for specific purposes by, for example, a social worker or physician.

Starting from this basic definition, autobiography can be categorized according to whether it is formal or informal. Though the distinction can be a fine one, formal autobiography means a deliberate and comparatively structured recounting of one's



life with the express intention of summing it up to date or making a public statement concerning it. While the expression may take many forms, including poetry and sculpture, obvious examples range from a published memoir to a curriculum vita. Informal autobiography includes what one reveals about oneself in less intentional ways, through one's speech, as in conversation or therapy, one's words, as in letters or diaries, or one's gestures and deeds. Behind both formal and informal autobiography lies one's autobiographical memory, or the memory one has of one's life as a whole (Rubin, 1996). However, insofar as such memory is internal to a person, assessments of its structure and possible impairments are impossible except as it is mediated by that person's actions or words. In this entry, then, 'autobiography' means any autobiographical *activity* that has some mode of external expression.

Additional distinctions by which autobiography can be categorized – and assessed – are whether it is voluntary (spontaneous, self-directed) or involuntary (requested, assigned); intended for a public audience or for private reflection; partial (concerning a particular period or theme in one's life) or complete (concerning one's life as a whole); superficial or in-depth; and whether the cue prompting it is specific or general (for example, *What was it like growing up blind?* or simply *Tell me about your life*).

## ASSESSING AUTOBIOGRAPHY

What is assessed from autobiographical activity, the method or instrument by which the assessment is carried out, and the theoretical perspective(s) in which the assessment is rooted, depend on the discipline or context that is involved.

Within the context of psychology, the most obvious example of this point is in relation to psychotherapy, and not least to the field of psychoanalysis. While the assessment and interpretation of autobiography constitute an integral source of information about an individual and about possible issues or themes on which the analysis can focus, the focus itself depends on the therapeutic perspective that is employed. Accordingly, it may be on, for example, a person's self-concept; degree of introversion–extroversion; obvious omissions from the person's self-report and their possible significance;

evidence of self-deception or of specific disorders; and/or locus of control.

Within developmental psychology, the focus may be on one's interpretation of life events; on one's life-course trajectory; on the evolution of personal identity (McAdams, 1988); on guiding personal metaphors; on the relationship between life story and values or emotions; and on changes over time in the content and form of one's self-report – or 'the development of autobiography' (Bruner, 1987). Within social psychology, sociology, and anthropology, the focus of assessment may be on the social constructedness of the self and on how 'narrative practice' (Holstein and Gubrium, 2000) concerning the self is portrayed and utilized. As conventions of self-talk and self-representation, or 'forms of self-telling' (Bruner, 1987), can vary profoundly by culture, language, gender, ethnicity, and class, they are necessarily of major concern in assessing differences in the accounts that individuals give of their lives.

Within cognitive science, the aim of assessment may be on the formation and function of one's autobiographical memory and on its completeness, reliability, and accuracy – that is, the interplay between fact and fiction within autobiographical memory (Rubin, 1996), or between 'historical truth' and 'narrative truth' (Spence, 1982).

Within a healthcare context, autobiographical activity can convey invaluable information concerning a patient's medical history, social networks and relationships, living conditions, and overall emotional and cognitive status. It can also provide a reference point for assessing differences between subjective and objective measures of physical health; and can assist in the detection and diagnosis of particular psychopathologies, including dementia.

Within the humanities, and specifically literary criticism, assessment of autobiographical activity may draw upon psychological or psychoanalytic theory to focus on the various functions, personal and social, that autobiography serves for the person who engages in it (LeJeune, 1989). In addition, it can focus on the narrative structure and integrity of particular autobiographical texts in terms of, for instance, plot, genre, theme, metaphor, point of view, and voice; on the role of language, and thus culture, in the formation and development of self-awareness and subjectivity; on the complex inter-relationships between

author, text, context, and audience (Olney, 1980); and on the philosophical and hermeneutical significance of being, at once, composer, narrator, editor, character, and reader in relation to one's own life story (Randall, 1995).

Finally, within gerontology, the study and assessment of autobiographical activity has perhaps a special significance insofar as gerontology is concerned with social and psychological development across the lifespan. Accordingly, the focus may overlap with that used in other disciplines and be on, for example, an individual's subjective experience of the ageing process, or *biographical* ageing; on the question of competence and of the relationship between person and environment (Svensson, 1996); and on the role played by autobiographical activity in relation to life review, generativity, spirituality, and preparing for death.

One particular method that uses autobiography in working with older adults – as a means not only of assessment but also of education, recreation, and (informal) therapy – is called 'guided autobiography' (Birren and Deutchman, 1991). In guided autobiography, persons write about their lives in relation to set themes – such as career, family, money, health, and love – and then share their writings with other individuals in a group setting. Such groups have been shown to be successful for those involved in increasing their sense of self-understanding and of personal integration.

In general, autobiographical activity in an advanced age can be assessed and utilized in terms of numerous functions that it can be said to serve:

- identifying and honouring key turning-points during one's life-course
- coming to grips with past resentments and negative feelings
- setting the record straight
- finding meaning amid life's struggles and challenges
- seeking answers to personal issues
- reviewing one's life to attain a sense of peace
- leaving a unique legacy of experience and wisdom

It should be noted, though, that autobiographical activity can serve many of the above functions at any point throughout the lifespan, and not only in later life.

## RESEARCHING AUTOBIOGRAPHY

From a research perspective, it would be valuable to examine the development of autobiography using qualitative methods within a longitudinal design. Of course, the very nature of autobiography leads us to treat it as 'longitudinal', since it provides a good characterization of how a person perceives his or her past in light of what life is like today and is expected to be like tomorrow, or in the future. However, such data represents not the past as it was at the time it occurred – not the 'true story' – but the past as perceived at the time it is recounted, and as portrayed to a particular audience. Of central interest in research on autobiography, then, would be how people's perception of their lives change, or remain stable, as they age, and what changes occur in both the selection of events that they recount and the angle or tone from which those events are interpreted and told.

One possible design is to ask people at age 60, for example, to tell about their lives at 60, at age 70 to tell about life at 70, and so on. This would enable an assessment of the degree of change or stability in the content of their autobiographies as they grow older. Similarly, asking people at 70 to tell about life at 60, and at 80 to tell about life at 70 (and 60), would permit an assessment of change and stability in people's perspectives on both their age and the ageing process. Finally, having people at 60 tell about their entire lifespan, at 70 the same, and so on, would provide a picture of the relative change and stability in their perspectives on the content and significance of their lives as a whole. Overall, such a design would permit a better understanding of how people perceive, represent, and interpret their lives at different stages.

## FUTURE PERSPECTIVES

In the future, due to rapid social change, there will probably be a more pronounced need and use of autobiography as a means for individuals to evaluate, understand, and integrate their lives, if not as a continuous process, then at different intervals over the lifespan. From a research perspective, there will most probably be a greater focus on using autobiographical data in longitudinal studies, especially of older persons, to

gain a sense of change and stability in their inner experiences of the ageing process.

Though it presents many issues for consideration, autobiography constitutes a valuable tool in several disciplines for assessing people's perceptions of their lives. In many ways, however, it has not yet been fully exploited as a qualitative method, especially in longitudinal research. As a complement to various tests and measures, it merits greater use in order to provide a fuller description and a richer understanding of the process of human life.

## References

- Birren, J.E. & Deutchman, D. (1991). *Guiding Autobiography Groups for Older Adults: Exploring the Fabric of Life*. Baltimore, MD: The Johns Hopkins University Press.
- Bruner, J. (1987). Life as narrative. *Social Research*, 54(1), 11–32.
- Holstein, J. & Gubrium, J. (2000). *The Self We Live By: Narrative Identity in a Postmodern World*. New York: Oxford University Press.
- LeJeune, P. (1989). *On Autobiography*. (trans. K. Leary) Minneapolis, MN: University of Minnesota Press.
- McAdams, D. (1988). *Power, Intimacy, and the Life Story: Personological Inquiries into Identity*. New York: Guilford.
- Olney, J. (Ed.) (1980). *Autobiography: Essays Theoretical and Critical*. Princeton, NJ: Princeton University Press.
- Randall, W. (1995). *The Stories We Are: An Essay on Self-Creation*. Toronto: University of Toronto Press.
- Rubin, D. (Ed.) (1996). *Remembering Our Past: Studies in Autobiographical Memory*. New York: Cambridge University Press.
- Spence, D. (1982). *Narrative Truth and Historical Truth*. New York: W.W. Norton.
- Svensson, T. (1996). Competence and quality of life: theoretical views of biography. In Birren, J.E., Kenyon, G.M., Ruth, J.-E., Schroots, J.J.F. & Svensson, T. (Eds.), *Aging and Biography: Explorations in Adult Development* (pp. 100–116). New York: Springer.

Torbjörn Svensson and William Randall

## Related Entries

QUALITATIVE METHODS, THEORETICAL PERSPECTIVE: CONSTRUCTIVIST, SELF-PRESENTATION MEASUREMENT, SUBJECTIVE METHODS, SELF, THE (GENERAL)

# A AUTOMATED TEST ASSEMBLY SYSTEMS

## INTRODUCTION

Historically, test construction in education and psychology has shown a development from: (1) the construction of standardized tests to the practice of assembling tests from item banks tailored to the test assembler's specifications; (2) the use of intuitive rules of test construction to the application of model-based algorithms; and (3) manual sorting of items on index cards to selection by a computerized system.

Test assembly can be characterized as the task of finding a combination of items from an item pool that satisfies a list of content specifications and is optimal in a statistical sense. Formally, the problem has the structure of a constrained combinatorial optimization problem in which an objective

function is maximized subject to a set of constraints, both typically modelled using 0–1 decision variables for the inclusion of the items in the test. Currently, a large variety of test assembly problems have been modelled this way and powerful algorithms for solving them are available.

## MODELLING TEST ASSEMBLY PROBLEMS

A common view underlying all attempts to automate test assembly is to see each item in the pool as a carrier of a set of *attributes* relevant to the psychological variable or the domain of knowledge or skills the pool is designed to measure. A formal distinction can be made between the following

types of attributes:

- 1 Categorical attributes, such as item content, cognitive level, format, answer key, and item author. This type of attribute implies a discrete classification of the pool, that is, a partition with classes of items containing the same attribute.
- 2 Quantitative attributes, such as item parameter estimates, expected response time, previous exposure rate, and word counts. This type of attribute is a value on a variable or parameter that, for all practical purposes, is to be considered as continuous.
- 3 Logical attributes, which imply relations among subsets of items in the pool, mostly relations of inclusion or exclusion. A relation of inclusions exists if an item has to be presented with other items in the pool because they share a stem or the description of a case. A relation of exclusion exists if items cannot be in the same test form, for instance, because some of them clue the correct answer to the others.

In addition to item attributes, it is useful to introduce the notion of test attributes. A test attribute is defined as a (function on the) distribution of item attributes (van der Linden, 2000a). Examples of test attributes are: the distribution of item content or  $p$ -values in a test, its information function, the number of items with a gender orientation, and its (classical) reliability. A test can now be defined as a set of items from a pool that meets a list of specifications with respect to its attributes.

An important distinction is between test specifications formulated as constraints and as objective functions:

- 1 A specification is a *constraint* if it requires a test attribute to meet an upper limit, lower limit, or equality.
- 2 A specification is an *objective function* if it requires a test attribute to take a minimum or maximum value.

The standard format of a test assembly problem is illustrated by the following example of a classical test assembly problem:

Maximize test reliability  
subject to

- 1 Number of items on knowledge of facts smaller than 15;

- 2 Number of items on application equal to 20;
- 3 All items four response alternatives;
- 4 Number of items with graphics at least 10;
- 5 Total number of items equal to 50;
- 6 No items with more than 150 words;
- 7 All item difficulties larger than 0.40;
- 8 All item difficulties smaller than 0.60;
- 9 All item discrimination indices larger than 0.30;
- 10 Item 73 and 98 not together in the test.

When translating test specifications into constraints, each constraint is required to have a simple form. For example, though it seems convenient to combine Constraint 7 and 8 into a 'single' constraint ('All item difficulties between 0.40–0.60'), such a step would obscure the total number of constraints actually involved in the problem. Also, for each problem only one objective function can be optimized at a time. If we have more functions, optimizing one of them automatically gives a suboptimal solution for the others. Finally, exchanging objective functions and constraints does not sometimes have too much effect. For example, we can replace the objective function in the above example by one in which the test is constrained to have reliability close to an educated guess of its optimum value and replace Constraint 7 and 8 by an objective function that minimizes the distances between the item difficulties and a target value of 0.50. In large-scale testing programs, test assembly problems in a standard format can easily have more than 200 constraints. For a more complete introduction to item and test attributes, test specifications, and rules for translating specifications into objective functions and constraints, see van der Linden (in preparation; Chapter 2)

A mathematical solution to test assembly problems becomes possible if the objective function and constraints are modelled using variables for the decision to select the items in the test. Let index  $i = 1, \dots, I$  denote the items in the pool. The most commonly used decision variables are binary variables  $x_i$ , where  $x_i = 1$  denotes the selection of item  $i$  and  $x_i = 0$  otherwise. (Other types of variables are sometimes necessary though; see (Section Some Applications).

A few examples of constraints modelled in terms of decision variables are:

- 1 Constraint 2 in the above example is a constraint with respect to a categorical attribute. If  $V_a$  denotes the set of indices of the items with the attribute Application, the constraint can be modelled as:

$$\sum_{i \in V_a} x_i = 20. \quad (1)$$

- 2 Constraint 7 is an example of a constraint with respect to a quantitative attribute. If  $p_i$  denotes the  $p$ -value of item  $i$ , it can be modelled as:

$$p_i x_i \leq 1, \quad i = 1, \dots, I. \quad (2)$$

- 3 Constraint 10 is a logical constraint. It can be modelled as:

$$x_{73} + x_{98} \leq 1. \quad (3)$$

All these constraints are linear equalities or inequalities in the decision variables. The feature holds nearly universally for all test specifications used in practice. A simple recipe to check if constraints are modelled correctly is to substitute trial values for the decision variables and determine the truth-value of the constraint. Examples of objective functions modelled in terms of decision variables are given in the section on Applications, below.

## SOLVING TEST ASSEMBLY PROBLEMS

Mathematical optimization problems with a linear objective function and linear constraints belong to the domain of Linear Programming (LP). The first to see the applicability of LP to test assembly were Feuerman and Weiss (1973) and Votaw (1952). If the decision variables are binary, the problem is known as a 0–1 LP problem. For a general introduction to these optimization techniques, see Nemhauser and Wolsey (1988) or Wagner (1972).

Once a test assembly problem has been modelled as a 0–1 LP problem, a solution can easily be found by solving the model for optimal

values of the decision variables using one of the algorithms available from the literature. Although 0–1 LP problems are known to be NP-hard, that is, to have solutions that cannot generally be found in a time bounded by a polynomial in the size of the problem, current technology has reached a level of sophistication that allows us to find exact solutions to problems with 1000–2000 variables and hundreds of constraints within seconds. Sometimes, test assembly models have the special structure of a network-flow programming problem. For such structures solutions to problems of virtually unlimited size can be calculated within a second (for examples, see Armstrong, Jones, & Wang, 1995). A very efficient general-purpose LP software package is CPLEX 6.5 (ILOG, 2000). A dedicated software package that helps test assemblers to define their problem and then translates the problem into an LP model is ConTEST (Timminga, van der Linden & Schweizer, 1996).

An alternative to model-based test assembly is test assembly based on a heuristic. Test assembly heuristics are computer algorithms that assemble a test in a sequential fashion, that is, by selecting one item at a time. They do so using an item-selection criterion designed to meet the test specifications. Because of their sequential nature, heuristics are generally fast. However, steps early in the sequential process cannot be undone later, and heuristics produce solutions that are not optimal. Another difference between the two approaches becomes manifest if a new class of test assembly problems has to be addressed. In an LP approach, the problem only has to be modelled and the model can be solved immediately by the algorithms and the software already available, whereas in a heuristic approach a new item-selection criterion and computer algorithm have to be developed and checked for the quality of their solutions. Examples of test assembly heuristics proven to be useful are given in Luecht (1998) and Swanson and Stocking (1993).

## SOME APPLICATIONS

### Target Information Function

The practice to assemble a test to meet a target for its information function was introduced in Birnbaum's (1968) pioneering work on

IRT-based test assembly. Theunissen (1985) was the first to realize that the problem can be solved using 0–1 LP, provided the information function is required to meet the target,  $T(\theta)$ , only in a series of discrete points,  $\theta_k, k = 1, \dots, K$ . Uniform approximation of the test information function to a series of target values is possible through a maximin approach (van der Linden & Boekkooi-Timminga, 1989). In this approach, test information is required to be in intervals about the target values,  $(T(\theta_k) + y, T(\theta_k) - y)$ , and the objective function minimizes the common size of the intervals. Formally, the model is

minimize  $y$  (4)  
subject to

$$\sum_{i=1}^I I_i(\theta_k)x_i - T(\theta_k) \leq y, \quad k = 1, \dots, K, \quad (5)$$

$$\sum_{i=1}^I I_i(\theta_k)x_i - T(\theta_k) \geq -y, \quad k = 1, \dots, K, \quad (6)$$

where  $y$  is a real-valued decision variable with optimal value to be calculated by the algorithm. (LP problems with both integer and real-valued variables are known as mixed integer programming problems.) Of course, these equations should be extended with a set of constraints to meet the content specifications for the test.

An empirical example for a pool of 753 items from the Law School Admission Test (LSAT) is given in Figure 1. The test length was set at 75 items. (The actual LSAT is longer because it duplicates one of its sections.) In all, a 0–1 LP model with 804 variables and 276 constraints was needed to assemble the test to deal with all specifications (including an item-set structure of some of the sections; see Section Tests with Itemsets). The test information function had to approximate the target at five values. Figure 1 shows both the information function of the test assembled and the full target.

### Multiple Test Forms

If examinees are allowed to take tests at different sessions, tests are often assembled as sets of

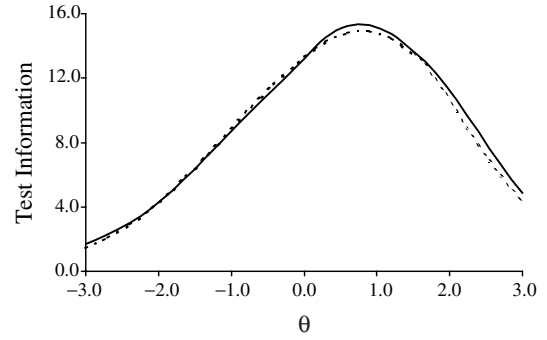


Figure 1. Information function for test form assembled from an LSAT pool (solid line represents target)

parallel forms. The best result is obtained if such sets are assembled simultaneously. If they are assembled sequentially, the value of the objective functions of each next form can be expected to be worse than those of its predecessors.

Multiple test forms can be assembled simultaneously if the following modifications are introduced:

- 1 The decision variables are replaced by variables  $x_{if}$ , with value 1 if item  $i$  is assigned to form  $f = 1, \dots, F$  and value 0 otherwise.
- 2 Constraints are added to the model to guarantee that each item is assigned to no more than one form:

$$\sum_{f=1}^F x_{if} \leq 1, \quad i = 1, \dots, I \quad (7)$$

For the same LSAT item pool, Figure 2 shows the information functions of three parallel forms assembled to meet the same target as in Figure 1. For more on this application as well as methods to deal with large multiple-form assembly problems, see van der Linden and Adema (1998).

### Tests with Item Sets

Tests with item sets are popular because they allow for the testing of knowledge or skills using the same case for more than one item. Often, the item pool has more items per set than needed in the test. Let  $s = 1, \dots, S$  denote the item sets in the pool,  $i_s = 1, \dots, I_s$  the items in set  $s$ , and  $n_s$  the

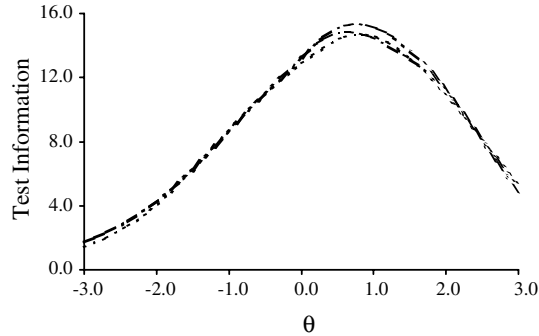


Figure 2. Information functions for three parallel test forms assembled from an LSAT pool (same target as in Figure 1).

number of items required from set  $s$  if it is selected to be in the test.

Tests with item sets can be assembled if the following modifications are introduced:

- 1 In addition to decision variable for the items, 0–1 variables  $z_s$  for the selection of set  $s$  are introduced.
- 2 Constraints are added to the model that both coordinate the selection of item and sets and guarantee the correct number of items in each selected set:

$$\sum_{i_s} x_{i_s} - n_s z_s = 0, \quad s = 1, \dots, S. \quad (8)$$

The LSAT form assembled for Figure 1 had an item-set structure for some of its sections. For other empirical examples and approaches to assembling tests with item sets, see van der Linden (2000a).

### Other Applications

The above applications illustrate only a few of the options made possible by 0–1 LP test assembly. Other options include: (1) classical test assembly, with Cronbach's alpha represented by a combination of an objective function and a constraint; (2) assembly of tests required to match a given test form item by item; (3) assembly of tests measuring a multidimensional ability; (4) assembly of multiple test forms that differ systematically, for example, a set of subtests for a multi-stage testing system or

testlets for testlet-based computerized adaptive testing (CAT); (5) assembly of tests with observed scores equated to those on a previous version of the test. A recent review of these and other applications is given in van der Linden (1998; in preparation).

### FUTURE PERSPECTIVES

Though the development of computerized adaptive testing (CAT) was mainly motivated by statistical considerations, real-life CAT systems have to meet a host of nonstatistical specifications as well. A recent development is the use 0–1 LP test assembly to introduce nonstatistical constraints in CAT (van der Linden, 2000b). The technique is applied through the assembly of a *shadow test* prior to the selection of the next item for an examinee. Shadow tests are tests that: (1) contain all items already assembled; (2) meet all constraints that have to be imposed on the adaptive test; and (3) have maximum information at the last update of the ability estimator. The item actually administered is the most informative item in the shadow test not administered to the examinee yet. Because after each update of the ability estimate the shadow test is re-assembled, the adaptive test is maximally informative. An addition, because each shadow test meets all necessary constraints, the adaptive test does.

Even though automated test assembly guarantees the best test from the pool, the result may be of low quality if the item pool is poor. In the parlance of 0–1 LP test assembly, the most important constraint imposed on the assembly of the test may be the poor composition of the item pool. It is therefore expected that an important future activity will be the development of methods to design item pools better targeted towards the tests to be assembled from them. A first attempt at optimal item pool design is given in van der Linden, Veldkamp and Reese (2000). A key notion in their approach is the one of a design space for the item pool. This space is defined as the Cartesian product of all statistical and nonstatistical item attributes involved in the specifications for the tests from the pool. (This operation may require discretization of quantitative attributes.) A point in this space identifies a possible item in the pool. The technique of integer programming is then used to calculate an optimal

blueprint of the item pool from the specifications for the tests the pool has to serve. The blueprint specifies the optimal number of items required for each point in the design space.

## CONCLUSION

Over the last decade several models and algorithms for automated test assembly have been developed. Automated assembly is now possible for almost every type of test and every set of specifications. This development seems timely because automated test assembly is the key to any form of computer-based testing and the current expectations about the improvements in the practice of testing that have become possible by the introduction of computers in testing are high.

## References

- Armstrong R.D., Jones, D.H. & Wang, Z. (1995). Network optimization in constrained standardized test construction. In Lawrence, K.D. (Ed.), *Applications of Management Science: Network Optimization Applications*, Vol. 8. (pp. 189–212). Greenwich, CT: JAI Press.
- Birnbaum, A. (1986). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Feurman, F. & Weiss, H. (1973). A mathematical programming model for test construction and scoring. *Management Science*, 19, 961–966.
- ILOG, Inc. (2000). *CPLEX 6.5* [Computer program and manual]. Incline Village, NV: Author.
- Luecht, R.M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, 22, 224–236.
- Nemhauser, G. & Wolsey, L. (1988). *Integer and Combinatorial Optimization*. New York: Wiley.
- Swanson, L. & Stocking, M.L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, 17, 151–166.
- Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika*, 50, 411–420.
- Timminga, E., van der Linden, W.J. & Schweizer, D.A. (1997). *ConTEST 2.0 Modules: A Decision Support System for Item Banking and Optimal Test Assembly* [Computer Program and Manual]. Groningen, The Netherlands: iec ProGAMMA.
- van der Linden, W.J. (1998). Optimal assembly of educational and psychological tests, with a bibliography. *Applied Psychological Measurement*, 22, 195–211.
- van der Linden, W.J. (2000a). Optimal assembling of tests with item sets. *Applied Psychological Measurement*, 24, 225–240.
- van der Linden, W.J. (2000b) Constrained adaptive testing with shadow tests. In van der Linden, W.J. & Glas, C.A.W. (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 27–52). Norwell, MA: Kluwer Academic Publishers.
- van der Linden, W.J. (in preparation). *Linear Models for Optimal Test Design*. New York: Springer-Verlag.
- van der Linden, W.J. & Adema, J.J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, 35, 185–198.
- van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika*, 54, 237–247.
- van der Linden, W.J., Veldkamp, B.P. & Reese, L.M. (2000). An integer programming approach to item pool design. *Applied Psychological Measurement*, 24, 139–150.
- Votaw, D.F. (1952). Methods of solving some personnel classification problems. *Psychometrika*, 17, 255–266.
- Wagner, H.M. (1972). *Principles of Operations Research, with Applications to Managerial Decisions*. London: Prentice-Hall.

Wim van der Linden

## Related Entries

ITEM RESPONSE THEORY