

# COMPUTER-BASED TESTING

## INTRODUCTION

Modern test tradition began with the French psychologist Alfred Binet who created the first intelligence test in 1905. This test was created to identify children who were likely to perform badly at school (Schultz & Schultz, 1992). The test was a so-called individual test, because of the individual contact between the psychologist and the child: The psychologist asked a question and the child gave an answer, just like in an interview or in a conversation between two people. Nowadays this test is still of importance because of its basic design. This design was that Binet changed the order of the questions depending on the child's answers. If a child gave incorrect answers he asked easier questions until the child gave a minimum number of correct answers in a row. Then Binet moved on to a more difficult level of questions until the child gave a certain number of incorrect answers in a row. In other words, Binet adapted the order of questions to the child's competency level.

Individual tests take a long time to administer. During the First and the Second World War there was a need for procedures to select and assign a lot of army workers in a very fast but accurate way. A new method of group-administered, or collective testing, came into use. Now, one proctor alone, that is a person who is in charge of the test session, could give instructions to many test takers at the same time.

Well-known group-administered intelligence tests used during World War I are the Army  $\alpha$  (read alpha), a verbal test, and the Army  $\beta$  (read beta), a non-verbal test for people who could not read or write. During World War II more than 9 million people were tested with the group-administered Army General Classification Test (AGTC) in the United States of America. Technology in those days was rather simple. Materials used to administer these tests were something to write on – paper – and something to write with – a pencil, hence the name “paper and pencil tests” or “p&p” tests.

In the 1960s the field in psychology called cognitive psychology has introduced the computer into psychology by stating that computer programs operate similarly to the human mind. According to cognitive psychologists a computer had a mind, used a language and was able to process information (Schultz & Schultz, 1992). Since then, computer technology has become a part of everyday life and has had an enormous impact on test use and test construction (Hardinge, 1997; Jansen, 1997).

## COMPUTER-BASED TESTING

In this reading we will analyse the p&p tests in order to discover what their relation is with technology and computer. Then, current and future use of computers in testing will be looked at in detail.

### PAPER & PENCIL TESTS

#### Definition

The group-administered tests described above were the first paper and pencil (p&p) tests or – as they sometimes are called – printed tests.

A p&p-test is a set of questions (or items) to be answered by the test taker with a tool to write on a document. The following materials are used in a p&p test:

- a booklet containing printed questions;
- an answer sheet on which the test taker marks the answers;
- a pencil or a pen in order to mark the answers.

P&p tests are mainly based on classical test theory. This theory assumes that each observed test score is composed of a true score and an error. So, in most cases the observed score does not equal the true score. Consequently the observed score may be higher or lower than the true score (see the section on classical test theory for more information on this subject).

Examples of p&p tests based on classical test theory are the DAT (Differential Aptitude Test) and the Bennet, a mechanical comprehension test. After a candidate completed a test, the responses given to the separate test items need to be transformed into a test score that means something. This can be done manually or by means of a computer.

We now will deal with how the computer can be helpful in scoring the answers on p&p tests to get to a meaningful test score.

#### Evolution of technology for the transformation of p&p-responses into test scores

Four important evolutions in supporting the scoring of p&p test answers can be identified. These are the implementation of

1. Punch cards
2. OMR (Optical Mark Reader)
3. OCR (Optical Character Recognition)

### 4. E-marking

Punch cards were cards in which a pattern of holes was cut to represent information, which was then read by a computer. This method is not used any more. Using the Optical Mark Reader (OMR) test takers answer the items filling in a so-called lozenge (a small area) on a form. The mark reader is a machine that recognises whether the lozenge has been filled or not. Then, this information can be transformed into test scores. OCR (Optical Character Recognition) was an improvement of the OMR system and produced more reliable test scores. OMR was created to recognise if certain blocks on the form were filled in or not, whilst OCR was developed to be able to recognise written characters and signs. Currently, so-called e-marking is on the market and makes it possible to recognize and record on-screen hand-written answers by candidates.

#### **Disadvantages of p&p-tests**

First of all, using a p&p test is time-consuming as it requires a lot of administration time during preparation, administration and processing of the scores. Second, it is slow and inaccurate when responses are transformed into test scores. Third, it is lacking flexibility from the perspective of the organization that uses the test. For instance, it is not possible to present different tests one after the other without intervention, or to change the number of tests, nor to compose different sets of tests. This means that you cannot adjust your test or its items to the candidate's competency if the test you originally presented proves to be either too difficult or too easy. Furthermore p&p tests have an old fashioned image and style. And finally, the restriction of this way of testing is that it has a fixed testing time and a fixed order of items.

These disadvantages imply that candidates have to go through a long, non-attractive test session, with many questions that could be too difficult or too easy in relation to their own competency levels. All candidates must fill out the same test items without the possibility of individualised test parts adjusted to fit the candidate. Unfortunately, there also will be learning effects when people take more than one test session. This means that people probably will obtain better scores when they take tests more often than once and this especially affects p&p tests that have fixed content.

## COMPUTER-BASED TESTING

The question then arises whether there are any alternatives to p&p-tests. Answering this question is the objective of the next section.

### COMPUTER-BASED TESTING

In the previous section we have been witnesses of how computer use can support p&p testing. Now we will check out whether the computer can be useful also in tackling the disadvantages of the p&p tests. Starting this exploration it would be good to know what is hidden behind the term computer-based testing (CBT).

#### Definition

A variety of terms and definitions are used to describe testing with computers. Sooner or later the reader who is studying the relation between computers and testing will be confronted with the following terminology and abbreviations:

- CBT (Computer-Based Testing);
- CAT (Computer-Assisted Testing; Computer-Aided Testing; Computer-Administered Testing; Computer-Adaptive Testing);
- CBA (Computer-Based Assessment);
- Computerised Assessment;
- Computerised Testing;
- CAA (Computer-Assisted Assessment).

Let us now turn to the meanings of these terms.

An analysis of definitions and descriptions of these terms reveals that some descriptions are rather vague, for instance: “Administering tests electronically using a computer” or “using an electronic testing system”. These kind of descriptions do not make a clear distinction between p&p tests and computer based testing (CBT), because computers may also be used as a support to facilitate certain parts of p&p test procedures, as we saw in the previous section on p&p tests. Furthermore it can easily be seen that it is confusing to use “CAT” as an abbreviation, because this abbreviation can point to different meanings. When “CAT” is used the author or speaker certainly must explain in what sense he uses this abbreviation.

## COMPUTER-BASED TESTING

The main difference between computer based testing (CBT) and p&p testing seems to be the use of on–screen devices by CBT. It therefore seems justified to give two descriptions of CBT: a narrow one and a broad one.

When we speak about CBT as the assessment of a person’s capacities through software and hardware using on–screen devices, we use the narrow definition of CBT. The term CBT should perhaps best be used for this narrow definition only. When we talk about administering tests with support of hardware and software, or with support of a computer, we use the definition in a broad sense. This second definition could be used for instance for a p&p test which is scored by a computer. Here we suggest to better use the term “CAA” (Computer Assisted Assessment).

Because the widely used term “CAT” is mostly linked to so-called adaptive testing (during adaptive testing the difficulty of the items is adapted to the competencies of the test taker) it is advisable to preserve this abbreviation solely for Computer Adaptive Testing.

### **Stages in the evolution of CBT**

CBT finds its origins in the early 1960s (Jansen, 1997). From then on computer elements were regularly used as administrative support for testing, mainly to convert test answers into test scores as discussed in the p&p test section before, and to introduce names and scores of the candidates. The Belgian Department of Defence introduced automated processes at the Recruitment and Selection Centre during the 1970s. Punch cards were used for the correction of the tests. During the eighties, gradually several computerised test systems came into use. Meanwhile, OMR (optical mark reader) systems were replacing punch cards. In the second half of the nineties the CBT–system was further elaborated. Recently in the Belgian Department of Defence, from 2001 onward, no more p&p tests were used and a fully automated system was introduced.

With respect to general CTB, McBride (1998) observed *three stages* in the evolution of CBT and called them the three generations of problems.

During the *first generation* cost was the central problem. At first, the computer equipment, development, software and maintenance were extremely expensive. This is the case with each newly implemented technology. McBride mentions the case of the U.S. Department of Defence’s ASVAB system (Armed Services Vocational Aptitude Battery). This battery is a series of tests developed by the Department of

## COMPUTER-BASED TESTING

Defence in the 1960s for enlistment purposes (Hardinge, 1997). In 1988, costs for a CBT-version were estimated between 25 million and 50 million US dollars, which was more expensive than the p&p version of the ASVAB. Eight years later, the cost was estimated less than three million US dollars, which made it beneficial to use the CBT-version. In 2002 the situation with respect to cost has changed even more. For instance, primary and secondary schools in South Bend, Indiana, U.S.A., now administer their tests on the computer. The new test program in South Bend costs only 3000 US Dollar more a year than the old p&p-testing program. Nowadays, computer equipment is much cheaper than before and this of course decreases costs.

McBride called the *second-generation problems* “Problems in *converting* Printed Tests to Computer Administration”. This second generation was about problems that arise when p&p tests are transformed into CBT-tests. A first problem is that transformations could cause differences in test results between the p&p version and the CBT-version (Neuman & Baydoun, 1998). This is especially the case when test takers are confronted with a time limit, the so-called speed or speeded tests. One of the reasons for differences in results may be that the act of marking the answer sheet is absent in the CBT-version; the absence of this act could increase the speed of answering. A second problem within this generation is of a different kind and has to do with differences in computer types. What happens when the types of computers that were used in the CBT-system are no longer available? For example, more recent computers are much faster than the ones of just a few years ago. Also at present, computers may have a different keyboard, which could cause a difference in response reaction time.

In the *third generation*, problems were centred on developing innovative tests to measure abilities. The most interesting thing about CBT is the possibility to develop new measures of ability that could not be realised through p&p. These new kinds of testing will be discussed now.

### ADVANTAGES OF CBT

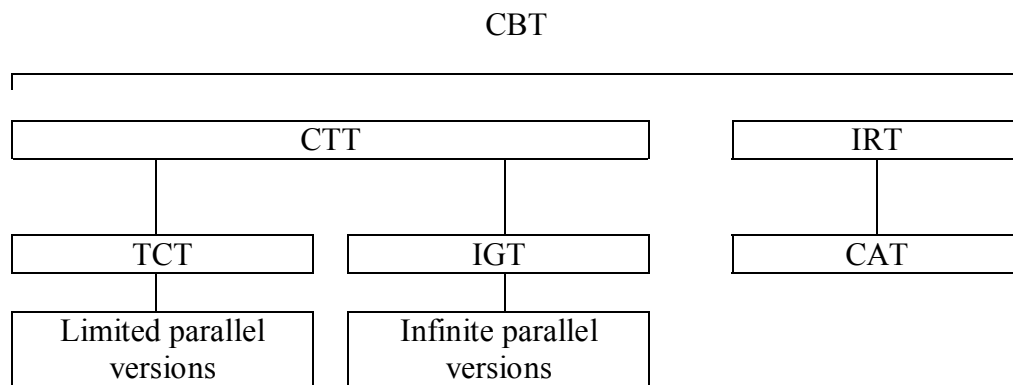
#### **Psychotechnical innovations**

The features of CBT can go beyond the limitations of the p&p tests. P&P tests are only based on classical test theory (CTT). CBT tests can either use CTT as a basis as well, but it can also use a new test-theoretical approach (Crocker & Algina, 1986).

## COMPUTER-BASED TESTING

This new approach is what McBride is referring to when he speaks about third generation problems.

In Figure 1 three CBT-procedures are shown. TCT's (Traditional Computerised Testing) main worries are with transferring p&p tests into computer versions. IGT (Item Generative Testing) delivers an engine on a computer that generates items. TCT and IGT are based on classical test theory. In contrast, CAT (Computer Adaptive Testing) is based on a new theory, IRT or Item Response Theory. This theory estimates the ability level of a candidate on the basis of his or her responses to previously administered items. Sometimes IRT is referred to as Modern Test Theory as opposed to Classical Test Theory.



**Figure 1. Computer-Based Testing concepts**

An overview of Item Generative Testing (IGT) is elaborated in Irvine and Kyllonen (2002). The reader may get introductions to CAT in Wainer, Dorans, Eignor, Flaughner, Green, Mislevy, Steinberg and Thissen (2000), and in Weiss (1983). Furthermore, Steege and Fritscher (1991) have treated some of the advantages of CAT. CTT and IRT are compared with each other in the work of Crocker and Algina (1986).

*Traditional Computerised Testing (TCT).* With the rise of new computer technology it becomes possible to construct different test versions measuring the same competencies with different items. In other words, now it is possible to develop equivalent forms of the same test. These equivalent forms are called parallel test versions or parallel versions. The items are stored in the memory of a computer. The act of creating such a database is often referred to as “item banking”. However, only a limited number of parallel versions per test are available using this technique of test construction: in most cases there are only two parallel versions. Two examples of

## COMPUTER-BASED TESTING

companies that deliver this kind of test batteries are CPM (Consultants in Personnel Management) and SHL (Saville and Holdsworth Ltd). They present tests in different languages and are able to deliver or develop parallel versions of tests.

*Item Generative Testing.* This form of testing makes it feasible to produce an infinite number of parallel versions. There is no need to keep items in a data bank, because items are created automatically by a computer algorithm (a set of rules to solve a problem) at the time of testing. In other words, items are generated automatically just before testing starts, hence the name Item Generative Testing. Hardinge (1997) mentioned that this new model serves as a basis for a test battery for selection purposes in the United Kingdom Army. The BARB (British Army Recruit Battery) was introduced in 1992. Instructions are given on-screen and each candidate answers a unique set of items, because of the unique feature of endless parallel versions. Nevertheless, all sets of items are of equal difficulty. Touch-screens are used: by touching the computer screen a candidate marks the answers. Test results are corrected in real time. Another advantage is that IGT principles provide the basis for developing adaptive test forms.

*Computer Adaptive Testing (CAT).* Computerised Adaptive Testing is a very clever way of testing. The computer program searches for an appropriate test item based on the candidate's response to the previous question. On the basis of an algorithm, items are selected from a bank of items. In fact, each time the candidate has responded, the computer program adapts to the candidate's answer. At the start of the testing session, an item of medium-difficulty is shown. Thereafter, each good answer is followed by the selection of a harder item and each incorrect answer is followed by an easier item (Wainer, et al., 2000; Weiss, 1983). This sounds familiar to our ears. Indeed, this procedure reminds us of the technique Alfred Binet used to assess young children with his intelligence test.

### **Technological progress**

Alongside psychotechnical innovations, CBT joins various technological novelties together, solving many disadvantages of p&p tests. Scores are accurately calculated in real time and can be communicated to the candidate immediately after the test session. Reports and scores are generated automatically in order to allow interviewers or teachers to interpret test results as soon as possible. Test time can be reduced by flexible directing candidates through a test according to the candidate's capacities.

## COMPUTER-BASED TESTING

Flexibility is further enhanced by integrating tests in a network, which permits to regulate test administration. The name for such a network is Test Manager or Test Administration System. Automated calculations and storage of data and test scores of candidates make research a lot easier. Ideally, the CBT system must allow the use of different kinds of tests and assessment instruments, such as multiple choice tests, questionnaires, tests using images or graphics, and open-ended answers.

Using such psychotechnical and technical advantages should result in a system that is more up-to-date, attractive and client-based than comparable p&p test systems.

Studies indeed show that candidates' attitudes towards taking selection tests via computers generally are positive (Steege & Fritscher, 1991).

### TEST STAGES AND AUTOMATION

Figure 2 gives an overview of the psychotechnical and technological progresses during the last century. The first column gives a test example of a certain time period, which was discussed earlier in this reading. In the second column the test concept or test stage is described and in the third column the appropriate degree of automation is shown. For instance, the Binet – Simon test is described as an individual adaptive non- computerised (or manually administered) test.

Test	Test concept or stage	Degree of automation
Binet–Simon / Terman –Merill	Individual adaptive	Manually
Army $\alpha$ , Army $\beta$	Group – p&p	Manually
BDSB70	Group – p&p	CAA
BDSB80	Colt stage	Computerised
BDSB90	Conversion stage	
ASVAB, IGT	Innovative stage	

**Figure 2. Test stages in relation to degree of automation**

There are two types of group-administered p&p tests: on the one hand the manually filled out form with its manual conversion into test scores, on the other hand the CAA (computer assisted assessment) form, including computerised support for administrative purposes and conversions of answers into test scores. In Figure 2, the

## COMPUTER-BASED TESTING

three CBT generations used by McBride are grouped as on–screen devices or CBT. Both CAA and CBT can be seen as computerised test procedures. Because some tests exist in different forms – or have been developed in different time periods – it is possible for the same test to be classified under different test concepts. An example of CAA and CBT systems is the selection battery used by the Belgian Defence Forces, namely the Belgian Defence Selection Battery (*BDSB*). In its origin it was developed as a p&p test battery, *BDSB70* was the first step in the CAA and started in the 1970s. Later on, the Belgian Defence Staff was very reluctant to develop a CBT battery because of the costs. They changed their opinion when automation of other personal management activities was introduced during the 1980s: *BDSB80* was the second step towards CBT. Then, in the late 1990s, the decision was taken to duplicate the p&p test on the computer resulting in fifty percent p&p test delivery and fifty percent CBT test delivery. Finally, in 2000, the choice was made to create a fully computerised test battery.

In Figure 2 the small upward arrows indicate that companies who try to introduce a new sophisticated technology may be confronted with failures and then can be forced to fall back on technology of a less sophisticated nature. An example of this again is the *BDSB*. During the cost period CBT technology was too expensive, resulting in the continuation of the p&p system with a CAA nature. Meanwhile, CAA changed from punch cards to OCR and at a certain point it became logical to introduce CBT. The cost period therefore can be interpreted as a transition period between CAA and CBT. In sum, it is possible that companies are situated in different test stages with regard to different parts of their test batteries.

### DEVELOPING A CBT-SYSTEM

#### Ways to develop a CBT-system

When developing a CBT-system the ultimate goal must be an integrated network that allows the user to adapt the kind and number of assessment instruments depending on his objectives and the context of the assessment: it allows to flexibly handle or manage test use. For this reason a CBT-system is often called a “Test Manager”. A Test Manager is not a person, but a system that gives the user the opportunity to choose tests out of a database holding a diversity of tests, and to compose different test batteries. The following four ways to develop a CBT-system will be discussed:

## COMPUTER-BASED TESTING

- buying a system
- developing a system independently
- an in-company system
- hiring a system.

For a user three ways exist to *buy* a network: the user may buy a commercial-of-the-shelf product (COTS product, that is a product that is purchased like it originally was constructed by the delivering company); a tailor-made network; or a semi-tailor-made network.

The Test Manager bought as a COTS product must be accepted and used as it was designed by the delivering company. Adjustments do not belong to the contract. The user sometimes depends totally on the company that delivers, which might result in a less flexible but less expensive system. One of the companies delivering a COTS test management system – the “Career Harmony Assessment Management System” or CHAM – is Consultants in Personnel Management.

In the case of *developing a tailor-made network independently*, the system is fully constructed on the basis of the needs expressed by the user. Adjustments are part of the contract. Tailor-made work is very expensive and takes a lot of planning and time. When using the semi-tailor-made approach, the existing *in-company network* is adapted, or a COTS product is adapted, in co-operation between the user and the delivering company. One possible solution is to co-operate with universities. In this way a flexible network might be put down, meeting the user’s criteria as closely as possible.

Besides buying a network, the user may wish to either develop a brand-new CBT network in-company or to hire the services of a company that organises the whole CBT assessment procedure. Of course, developing a CBT-system without any help from an external consultancy company requires a lot of technical and psychotechnical expertise and resources inside the company. If there is not enough expertise in the own organization, an alternative may be to *hire* a CBT assessment procedure. There are many companies specialised in organising CBT assessments and the user might have candidates tested in-company or at location. The next cases illustrate which services are available.

## COMPUTER-BASED TESTING

*ETS (Educational Testing Services) annually delivers more than 1,3 million CBT tests worldwide. CAT\*ASI (Computer Adaptive Technologies, Inc. \*Assessment systems, Inc.) manages a nation-wide network of test centres and accommodates CBT sessions for high stake tests (high stakes tests are used for decisions of great importance, such as admission sessions for universities or for a job). They offer services such as testing technology (graphics, different question types), the development of new items, management of test sessions, deployment and delivery of the materials, and processing of data and reports.*

### **The development of a CBT-system**

In most of the cases, users wanting to develop a CBT-system will have to perform a market analysis and write Requests for Proposals (RFP) in order to receive tenders from companies. The basic underlying principle is the same as for traditional test development. The test developer should bear in mind that the computer is only a means to support administration and construction, and not an end in itself (for an introduction in testing psychology, see McIntire & Miller (2000); for designing a tests, Russel and Peterson (1997) forms a good starting point).

Of course, CBT requires specific needs. For instance, it must be clear whether the test objectives need special hardware, such as joysticks for psychomotor assessment in the case of pilot or driver selection. In Parshall, Spray, Kalohn and Davey (2002) advice can be found concerning the practice of implementing CBT. Special matters of concern are furthermore security and maintenance (International Test Commission, 2005). In order to keep items and data safe, the CBT system must offer the possibility of restricted access via passwords. In that way the user can decide who has access at what moment and at what level. Different levels could include proctors (test administrators), developers, psychologists and managers. Second, the database containing personnel data of the candidates must be secured. Third, the possibility must exist to create backups of the data. And finally, it is important that the system is secured against current or electricity interruptions.

The main points with regards to maintenance of a CBT system can be split in two ways. It can be assured firstly by the delivering company who can offer a traditional help desk, which means that problems are solved via e-mail or via local interventions. A more sophisticated form of help desk can be established via a remote control or distance help desk, on the basis of a direct link between two computers over the

Internet. The name for this remote control help desk is “Virtual Private Network” (VPN).

### **IMPLEMENTATION OF CBT**

Depending on the setting, the question arises whether it is practical and cost-effective to apply CBT. First, CBT is especially attractive and effective in the field of large-scale selection, large-scale educational programs and where there is a need for classification and allocation of large numbers of candidates (Bennet, 2001; Steege & Fritscher, 1991; Zakrewski & Bull, 1998). On the other hand, CBT can also be very useful in situations with only a few candidates where complex skills need to be measured. CBT can measure such abilities much more effectively, they are not measured as well by p&p tests.

In the next section examples of CBT in three domains will be discussed. These three domains are selection, educational and school programs and training and appraisal.

#### **Selection**

In the domain of the military selection the ASVAB and BARB projects have already been mentioned. Other CBT batteries are Micropat (a test battery for helicopter crew) and pilot selection batteries. In the domain of State Department selection or Government selection CAT\*ASI is one of the companies which organises CBT sessions. Bell, IBM and AT&T are large private organisations which have developed their own CBT selection tools (McBride 1988).

#### **Educational and School Programs**

In this context CBT and CAT are frequently used for admission issues in relation to a school or educational program. A well-known admission test is the GMAT (Graduate Management Admission Test). Other CBT tools, like TOEFL (Test of English as a Foreign Language), form a part of evaluation and assessment procedures to determine the mastery level of the English language. In The Netherlands the Central Institute for Test Development (CITO) is the State Department producer of nation-wide CBT-tests, and is specialised in CAT techniques. A final example is that of Lievens and Coetsier (2002) who have reported on Situational Judgment Tests in student selection.

### **Training and appraisal**

Examples of CBT with regard to the assessment of complex skills or competencies – as for instance needed for aircrew, air traffic controllers or pilots – are management tests and simulation tests. The Canadian Air Force implements such a simulator, the Canadian Automated Pilot Selection System (CAPSS). CBT may also be appealing when it is difficult to train or evaluate skills in real situations. This may occur when a trainer needs to evaluate whether trainees master observation techniques to be used during Assessment Centre (AC) procedures. AC procedures are a series of individual and group tests during which candidates are observed and evaluated. Imagine that the trainee needs to practice observation techniques during a real selection procedure. This would be very inefficient. In order to avoid these practical problems the trainees could take a CBT AC exercise. Such an exercise was developed for use in the Dutch–Flemish Open University study program for social sciences. Trainees interact with the computer program using different media techniques. Trainees get instructions, observe four candidates on the computer screen and evaluate the four observed candidates. The scores given by the trainees then are compared to expert scores, which appear on–screen. In addition, trainees perform an interview with a candidate via the computer screen. Finally, trainees decide on a final score, which is compared to a score given by an expert. On-line help and comments are produced by the computer. In this way personnel costs for trainee guidance is minimised. Another advantage is that trainees generally seem satisfied with this procedure.

Every advancement has its downsides, and CBT makes no exception to this saying. Therefore, in the next section some CBT-problems will be highlighted.

### **PROBLEMS RELATED TO CBT**

In general CBT is used more and more. Nevertheless, many potential users are reluctant to implement CBT systems. Six problem areas form possible grounds for this reluctance.

First of all, there is a general rule concerning the acceptance of new technologies: at the start of the new technology, acceptance is very low, because people are used to the old technology (Jansen, 1997). Some people still think in terms of p&p tests. In addition, it should be noted that computer business is currently recovering from the recent crisis in the technological world.

## COMPUTER-BASED TESTING

Second, a lot of organisations still are struggling with problems related to the p&p time period, such as how to integrate the administrative part of the test taking into the computer (Jansen, 1997). These problems are situated in the CAA generation.

Thirdly, another set of problems relates to the first generation of CBT problems. Purchasing or developing a CBT system could appear to be very expensive, certainly if a computer system is not already installed. Furthermore, as Jansen (1997) specifies, a budget for CBT is no priority. Instead, organisations prefer to invest in the computerisation of their administration and offices.

Fourth, equivalence of CBT test versions and p&p test versions should be checked (Neuman & Baydoun, 1998), especially in the context of speeded tests. This kind of problems can be referred to as CBT problems of the second generation.

In the fifth place, the creation and facilitation of the innovative test generation of CBT have their own specific difficulties. The danger of complete dependence on technology is realistic. There is a need for logistic support, e.g. enough computers, electricity, rooms – if not, CBT simply cannot take place. In contrast, p&p tests are more mobile, because they can be taken to every location one can imagine. Also, there are some specific problems related to CAT. First, even the most difficult items can be memorised by candidates. Consequently, the item bank must be large enough so that the smarter people will not be able to memorise items and recognize them later on. For the construction of a CAT data bank in general, many items and candidates are needed. Furthermore, the construction of open-ended items is a difficult matter in CAT. As for the IGT, critiques mention that these tests are a little biased towards measuring candidates' working memories and speeded aspects of their performance (Hardinge, 1997).

Finally – to close this section focusing on CBT related problems – the development and implementation of CBT requires a team of experts composed of test developers with experience in the field of CBT technology, computer scientists for analysis and programming, and statistical experts.

### **FUTURE Developments**

According to Russel and Peterson (1997) technology will probably continue to rapidly change the world for the next decades. As more and more companies become accustomed to computers and CBT, new ways of CBT will spread more and more. The following changes are very likely to impact upon the future.

## COMPUTER-BASED TESTING

A change seems plausible from keyboard and mouse, to touch–screen and pen–based computers. At the moment research is focusing on how to convert written material onto the computer–screen. Touch–screens have already been implemented, for instance during the BARB–project.

A second tendency might be the increase of tests using visual stimuli in CD-ROM technology. These kinds of tests have already been conceived, mainly as Video–Based Situational Tests.

Third, another future development could be the implementation of new test features. In the past, tests were limited to written and oral assessment procedures. Currently, features like (stereo) sounds, animation and full–motion video, and interactive test simulations, seem to be promising and typical for an expanding multimedia technology (Drasgow & Olson–Buchanan, 1999; McBride, 1998; Russel & Peterson, 1997).

In the fourth place, CBT may create possibilities to discover and analyse new psychological constructs, as was done by Kyllonen and his team setting up the LAMP (Learning Abilities Measurement Program) (Steege & Fritscher, 1991). CBT also makes it possible to improve the measurement of traditional psychological constructs. In the fifth place there is the growing importance of the Internet (Web-Based Assessment, on-line–assessment, e–assessment, or e–applications) in all domains of testing. Different companies have already put complete Internet–based test systems on the market. The Swedish company Enlight has developed “Enlight TestStation” that enables to produce e–assessments. Another example is that of the Dutch company Van der Maesen. They also developed a Multimedia Situational Judgement Test for social competencies. One of the latest developments in this domain is so-called e–marking, which opens the opportunity, in contrast to multiple choice questions, for the analysis of natural language processing techniques (Sturman & Kispal, 2003).

Sixth and lastly, one of the latest innovations is the palm–top, a small computer that fits in your hand or palm. This small instrument may be usable for administering psychotechnical tests. For the moment, developments five and six are not applied on a large–scale, mostly because of problems with the security of the data.

## CONCLUSIONS

Psychological testing has evolved from an individual oriented testing via a group–oriented p&p administration to a diversified approach based on computer technology,

## COMPUTER-BASED TESTING

which allows testing of people individually or in groups on simple behaviours or more complex behaviours. These changes took place during the period 1905 – 2005. The changing agent was technology, mainly the computer, moving through three stages each characterized by a central problem namely 1) cost, 2) conversion of p&p tests to CBT, and 3) the creation of innovative tests.

The computer has become a part of everyday life, but CBT still is not fully accepted. There remain problems. Organisations struggle with problems ranging from p&p and CAA problems, to CBT problems of the innovative generation. Certainly, CBT continues to be costly (Bennet, 2001).

CBT-systems may be developed in different ways and are especially efficient in meeting large-scale assessment needs. Under certain conditions it may be useful to develop CBT-procedures for other than large-scale settings. The ways to develop a CBT-system and the cost-effectiveness of the effort to construct a CBT system in relation to the objectives must always be considered. As a general rule, CBT-systems ought to be realised according to the principles of traditional test development, apart from additional new issues, such as security and maintenance. It is also important not to let technology determine the content of assessment.

Although CBT has some disadvantages, the advantages outnumber these disadvantages. In the last decades some exciting innovative CBT-systems, such as CAT and IGT, have been utilized in a lot of contexts and hold promising prospects for future research and practice.

It is expected that computer and audio-visual technology will continue to change the attitudes towards assessments in learning and work contexts (Bennett, 2001; Hardinge, 1997; Jansen, 1997). Adaptive learning and flexibility in ways of thinking and behaving will become more important. One of the answers to this is CBT. CBT has the potential of becoming the tool with which it will be possible to assess the impact of these changes on the abilities of future candidates.