

# The quality management of large-scale computer based assessments

34th IAEA Annual Conference  
Cambridge, September 8-12, 2008

Peter Hermans  
Cito Arnhem

## Abstract:

This paper discusses the implications of the use of new technologies in large-scale examinations for the quality management of processing and handling examination materials. Since 2003 Cito, the Dutch National Institute for Educational Measurement, has cooperated closely with schools and the National Board of Examiners in the development and administration of experimental examinations that require the use of computers. In 2006, these examinations were distributed on a national scale for the first time.

Large-scale examinations that involve the use of new technologies require a different approach to the processing of pen-and-paper examination materials. On the basis of a systematic analysis of the processes and products of the first generations of experimental examinations, Cito is designing a system for the quality management of the development, construction, production and distribution of computer based examinations.

### *A decade of computer based assessments*

Ten years ago, computer use in Dutch school leaving examinations was limited to data processing and test- and item analyses, exams were 100% paper-based.

Today, computer-use covers the entire workflow at Cito from the first steps of item production to reporting. Computers are used to develop and construct test items; computerized item and test management allows us to store and retrieve items and tests anytime, anywhere. We can administer and distribute tests and assessments on demand. We collect and process data and to generate reports with a single mouse-click. In a decade Cito moved from golf ball typewriters to networked test construction, an item banking system, a single login test portal, testing on demand, computerized adaptive testing, digital student monitoring systems, online submission of test scores; we facilitate downloads of examination papers, marking schemes and results.

The first experiments with innovative computer based assessments within the setting of National Examinations started in the late 1990's. Our activities then were already focused on the development of assessments that would go beyond computer based delivery of paper based tests. In 1998, the National Examination Board commissioned a computer based version of the school leaving examination for basic level foreign languages, an interactive 'language village', a simulated holiday trip to village in France, England or Germany. This computer based examination should include: listening, reading, writing and speaking. The first prototype met most of the technical specifications and requirements. In the final version the speaking component had to be skipped due to limitations of the commercial software that we used and students had to register their answers on paper because the risk of errors in the answer database could not be eliminated. Despite of these drawbacks the digital language village was released as an official examination.

In 2002, the National Examination Board decided to start the "Compex" (Computers and Examinations) project to coordinate different initiatives in the development and implementation of innovative computer use in examinations. Compex examinations were taken by small groups of students at a limited number of schools. All other students sat for the paper based version. The aim of the project was to include computer use in as much of the national examinations as possible.

Compex examinations have been developed as stand alone, Internet based and local network applications, using almost every suitable software package that could be handled by the computer systems in schools. After five years, the focus of the Compex project shifted from innovative computer based assessment tasks towards the development of a standardized infrastructure for the delivery of the computer based assessments that could handle non-linear, interactive, open-ended assessment tasks leaning heavily on use of multimedia.

*Re-engineering the development of innovative computer based assessments*

More emphasis on the infrastructure the delivery of computer based assessments does not imply that the development of innovative computer based assessments came to a standstill. One important reason to continue the development of innovative computer based assessments, were the critical remarks from school principals about the innovative quality and the added value of the experimental computer assessments that had been developed in the course of the Compex project.

Doubts about the added value of multimedia use in computer based assessments are often the result of an inevitable, unconscious comparison of the look and feel of these assessments with state of the art multimedia use in other applications. Advocates of the 'new literacy' like to refer to the interactive role playing computer games (RPG's ) as an alternative for the linear computer bases assessment tasks. But state-of-the-art large scale computerized assessments, administered by schools, will always be conservative by nature because of technical constraints. The bottom 10% of the hardware specifications found in schools, more or less define the maximum hardware requirements for the delivery of a large scale computer based assessments. These constraints limit the use of the multimedia and interactivity in assessment tasks, which makes computer assessments 'boring' compared to the applications that can be found on an average personal computer at home. But the real question is, do computerized assessments have to be exciting and fun to do because computer games seem to be the only thing in life that really seem to motivate students?

It is very unlikely that the characteristics of computer games - overwhelming multimedia effects, interactivity, total immersion and a constant appeal to strategic and autonomous decision making - will be the guiding principle for the development of computer based assessments of tomorrow. In the next decade, assessments will remain standardized entities in a controlled environment, and most assessments will hardly be interactive because they will remain goal or task oriented. It is unlikely that assessments will go beyond verifying whether the task taker has mastered one of the two or three possible ways of executing a task. On the other hand, future computer based examinations might go far beyond the computer games our students are playing now. These examinations might be so intertwined with learning that one cannot tell where one ends and the other begins.

In a way, these doubts about to the innovative quality and added value of computer based assessments also reflect the lack of quality standards for the development of these kinds of tests. Innovations in computer based testing are driven by software engineers rather than by test developers and only very few innovations in testing software are the result of test development. Most quality control is focused on technical issues, especially optimizing technical conditions for a reliable delivery of the assessment, overlooking or ignoring possible consequences for the content (the validity) of the assessment. In most occasions:

- computer based assessments are designed and developed by people with very limited experience in IT and multimedia;
- most assessments are produced by IT and multimedia experts with little or no experience in test development;
- the process of developing computer based assessments is managed in the same way as the development of paper based tests.

*New actors, new roles*

Compared to paper based tests, developing computer based assessments brings more and different fields of expertise into the process of test development than just item writing skills and psychometrics. A team developing computer based assessments might include multimedia designers and programmers, animators, interaction and interface designers, graphic designers, audio technicians and cameramen. In our experience these specialists need to be actively involved in the process of test development from start to finish.

It is vital for a successful production of computer based assessments these different fields of expertise are included in teams of developers rather than transfer the content of the assessment from one department to another. In contrast to paper based test, development of computer based assessment imply a dynamic and interactive production process and not the linear workflow that is typical for paper based tests.

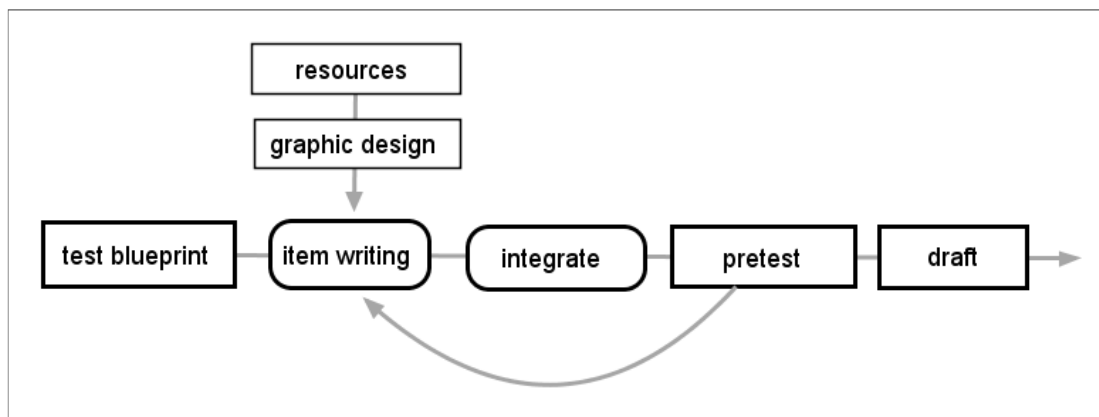


Figure 1: basic model of paper based test development

Figure 1 shows a basic model of the development of a paper based test, ending with the draft version of the test to be submitted to a government body for approval as a national examination.

In the case of computer based assessments a test blueprint or table of specifications is too limited to serve as a design brief for the multidimensional processes that guide the development. Such test blueprints have to be translated into a narrative, a scenario and finally a detailed script, describing not only the required and expected thinking processes and actions of the test taker, but also the possible sequences of events and actions that are expected.

And the analogy with movie making does not stop here, managing the development of computer based assessments, combines major characteristics of the roles that producers and directors have in the case of a movie.

Computer based assessments add a second new dimension to test development. In order to guide the work of graphic and interaction designers, it is often necessary to visualize the content of a computer based assessment in a series of screen designs similar to making of a storybook in movie making.

Figure 2 shows a basic model of the development process of a computer based assessment.

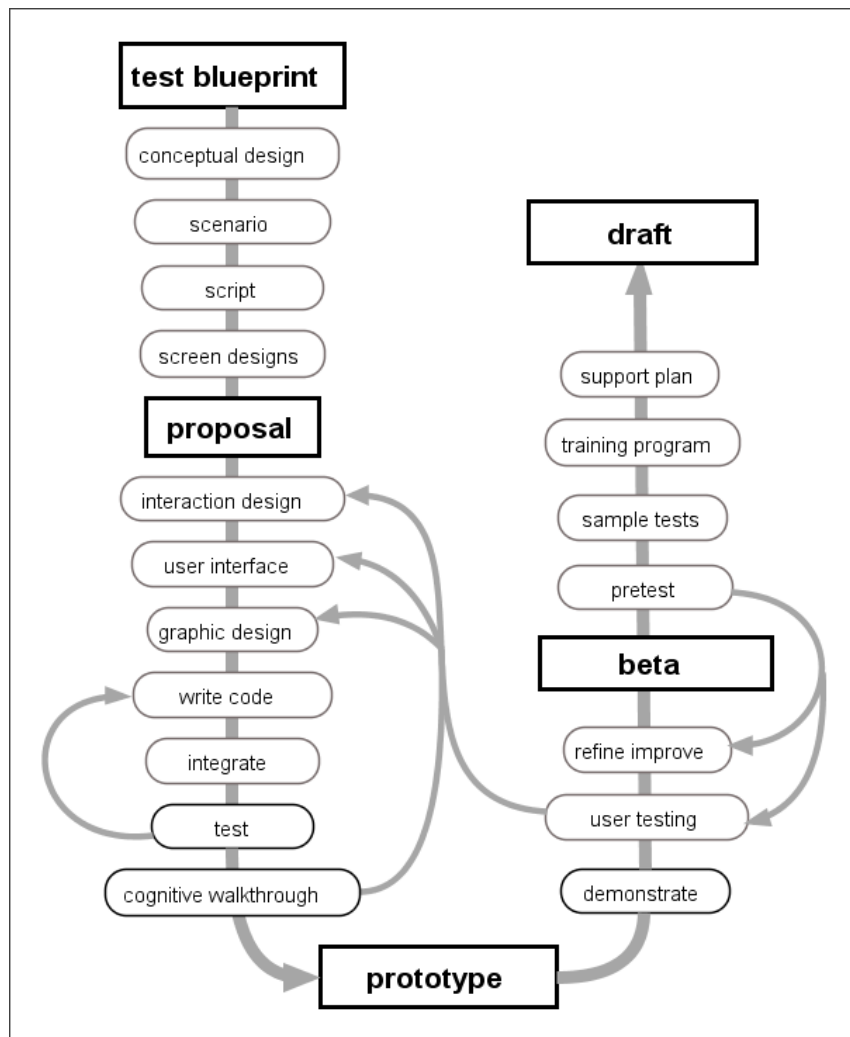


Figure 2 basic model of the development of computer based assessments

The first part of the process is of a creative nature. Test developers, item writers and designers try to imagine narratives that describe a task (a ‘quest’) that includes as much of the knowledge and skills included in the table of specifications or test blueprint as possible.

Narratives like this need to be translated into scenarios and at the end in very detailed scripts describing sequences of events, situations and interactions ('dialogues').

A proposal, describing the assessment in detail, including assignments, tasks, questions and answers, is then translated into a computer application by designers, animators and programmers. This proposal might include first prototypes of parts of the assessment task or previously developed prototypical applications used in other computer based assessments.

After a series of technical and other tests, the application is then demonstrated as a prototype to the commissioning governmental body. After approval by the commissioning body, a next series of tests, including testing the application with potential users of the test, leads to a beta for testing in the school environment, mostly by using sample tests in the context of a school examination. A support plan and a training program for system engineers and examination administrators are an essential part of a large scale computer based assessment and should therefore be part of the draft of the final examination that is submitted for final approval.

Paper based tests seem to follow a more or less linear development process. Development of innovative computer based assessments implies a process that on the one hand needs to be as rigid as a software development and on the other hand should leave enough room for creative contributions from all the parties involved. Managing this process means finding a delicate balance between control and creation. It also means establishing a common working culture for professionals with different ideas about creativity, standards and control.

Crucial in this process is the development of design standards, the establishment of common design rules and procedures for quality control.

Before the draft of a computer based assessment is submitted to for approval as a national examination, the Board of Examiners has at least reviewed three earlier versions of the assessment task: the proposal, the prototype and the beta version. Planning 'go - no go' decisions at these points in the development process is essential for the success of the project. Developing computer based assessments is expensive and most of the time funds are limited, and therefore the commitment from the governing body from beginning to the end is vital. The decision about the prototype is the most important moment in the development process and there are three kind of testing routines related to this phase in the development of a computer based assessment task.

1. A technically reliable delivery of computer based assessments requires a thorough testing routine on a series of different configurations and under all of the conditions imaginable during a national examination, long before the actual examination takes place. Even the slightest change in the application, makes it necessary to go through the testing routine again. Not only changes in the program code require a new series of tests, changes in the content of the test can also lead to unexpected error messages or even computers crashes or loss of data.

2. It is vital for a successful delivery that a second series of technical tests is done at every school. For this purpose schools need sample tests, versions of the assessment that equal the actual examination technically but with a different content. Running these kinds of tests well in advance of the examination period also guarantees a flawless installation of the computer based assessments included in the actual examinations.

3. Before the assessment is released as a prototype, a series of tests, so called ‘cognitive walkthrough’, is carried out by experts with special focus on interaction design, interfacing and ergonomics. In a ‘cognitive walkthrough’, the sequence of actions refers to the steps that an interface will require a user to perform in order to accomplish some task. The experts step through that action sequence to check it for potential usability problems.

#### *Quality control and test bias*

Although most computer based assessments are well designed, that in itself is no guarantee that they are well designed assessments. Cognitive walkthroughs are a necessary step in the quality management of computer based assessments, because it is the only way to reveal sources of test- and item bias caused by the use of the computer for the assessment. The origins of this kind of test bias are no different then some of the annoyances we encounter in using software in general:

- inconsistent use of terminology (violation of rule no. 1 of interface design: ‘same name, same thing, different name, different thing’)
- using ‘geek’ terminology (incomprehensible error messages)
- instructions that are easily overlooked
- (long) instructions that disappear suddenly
- inconsistent use of symbols and metaphors
- parts missing on the screen or parts that are not displayed properly
- no ‘undo’ or going back (or a ‘cancel’ that does not cancel)
- unclear effect of requested actions
- confusing settings of buttons and checkboxes
- intolerant data fields
- asking for unneeded input (including asking 2x)
- application gives no feedback about what it is doing

Most users will accept these kinds of design flaws (because that is what they are) when they encounter them during their daily computer use, but they are unacceptable in the context of an assessment task, especially in the case of high stakes testing.

Item- and test bias caused by computer use can be minimized by introducing a quality control program to check the application for slip-ups in the interface and interaction design, but the

quality of computer based assessments can only be assured if the development team works with the same set of quality standards based on a common frame of reference.

All developers must know the audience of test takers and understand the goals of this audience. They must realize that they are supposed to create an application that can be used by all test takers, including colour blind, hearing impaired, physically handicapped or those who are not native speakers. They must develop an application that is consistent with itself and with other applications that test takers are familiar with. Consistency, both in appearance and in behaviour, enables users to use their existing knowledge and computer skills and helps create a sense of comfort and confidence. Perhaps most difficult for most developers is that the application should put the user in control by giving the user accurate and precise feedback at any appropriate time, by keeping the appearance simple and pretty and by allowing test takers to undo their actions at any time

Computer based assessments are not part of our daily computer use, but more and more students will be confronted by digital forms of assessment and testing in the near future. Educational institutions are easily convinced of the many advantages of computer based testing. It saves time, the system handles most of the planning and test administration and even takes care of the time consuming scoring of student answers. The advantages for test takers are less obvious. There are still many students who are reluctant to sit for a computer based examination. The most obvious reason why computer based testing is not everyone's favourite is probably the technology related anger that all computer users share. Ranging from uncertainty about what the computer is doing to stalling software and crashing computers: although the test might be reliable, technology will never be 100% proof.

Ten years of developing computer based assessments have taught us that:

- Close cooperation with management, teachers and IT staff in the schools is a prerequisite for managing the quality of computer based assessments.
- Managing the development of computer based assessments is a mix of classical (product oriented) project management and process based (vision oriented) management. Both need careful monitoring during the entire development process.
- Members of a multidisciplinary development team have different training, ways of working and culture and the roles and boundaries between the different disciplines change and shift constantly. This will not only affect how team members work together as a team but it will also cause confusion about responsibilities and decision making.
- Every member of a development team needs clarity on their own role and on what other team members do. Mechanisms of power (competing for control, status,

marginalising certain professional identities) and the way decisions are made need to be addressed regularly.

- Multidisciplinary teams rarely share a common language or jargon, which can be alienating. Team members must learn to value each other's contributions, look at how the group communicates and be aware of making judgements and holding prejudices. Emotions and egos guide creative processes but at the same time should not get in the way of the team effort.
- The basis for a successful development of computer based assessment is a precise script describing the actions the test taker does or is supposed to do (mental operators as well as physical actions) to complete the assessment task.
- Multidisciplinary development teams are like families: they either work together smoothly or they are totally dysfunctional.

Even test takers learn to live with the technical shortcomings of computer systems.

But fact remains that many students prefer paper based tests for the same reasons that most taxpayers prefer paperwork over filing their annual income tax forms online: computers are prescriptive black boxes that do not give users the confidence that they are in charge and can follow their own strategy.

Computer based assessments will only be successful when the developers learn to see the test taker not as just another user of their computer application but as a valued customer entitled to a fair, valid and reliable assessment.