IMPROVING THE VALIDITY OF ASSESSMENTS IN COMPUTER BASED ASSESSMENT

Desirée Joosten-ten Brinke, Open University of The Netherlands, The Netherlands
Co-author: Maurice L. de Volder, Open University of The Netherlands, The Netherlands

Summary

Computer based assessment has the opportunity to improve the efficiency of assessment processes by delivery, administration, storage and scoring of assessments. However, above these practical benefits, computer based assessment provides the possibility to improve the assessment validity. This paper presents the implementation strategy of computer-based assessment by the Open University of the Netherlands and the interventions made to improve the assessments’ quality. First, the development towards computer-based assessment at the Open University is described. Secondly, the implementation strategy towards a full computer-based assessment organisation is given and thirdly, an elaboration is given on the improvement of assessments’ validity.

Introduction

Computer based assessment has the opportunity to improve the efficiency of assessment processes by delivery, administration, storage and scoring of assessments. However, above these practical benefits, computer based assessment provides the possibility to improve the assessment validity. This paper presents the implementation strategy of computer-based assessment by the Open University of the Netherlands and the interventions made to improve the assessments’ quality. First, the development towards computer-based assessment at the Open University is described. Secondly, the implementation strategy towards a full computer-based assessment organisation is given and thirdly, an elaboration is given on the improvement of assessments’ validity.

From paper and pencil to computer based assessment

The testing tradition at the Open University started in 1984 when the Open University was established. Each course with the volume of 100 hours of study was finished with an exam. All these exams were taken by paper and pencil. This testing system became very expensive and the first ideas about an automated test service system arose. The testing system should be flexible, so every student is free to choose an own preferred time to make the finishing test. This idea fits the philosophy of the Open University of freedom of time, place and pace. Only place was then limited by the free choice of one of the 18 study centres.

The development of an individual automated testing system, named SYS, was divided in several stages (De Roode, 2009). In the first stage, the computer selected ad random a set of item numbers. These item numbers correspond with items, only multiple-choice items, in
an item book. The item book consisted of all the available items. The next stage was the automation of the item books. The selection of item numbers and items was now automated, but the delivery to the students was still on paper. With the exam, the computer produced an optical readable form on which the students gave the answers to the question. This form was read by a scanner and translated in a (tentative) score.

SYS started with three courses. The development of SYS met the expectations of the examination committee of the Open University. Moreover, although some of the students were at the start rather reserved to the computer-based exams, the acceptability raised quickly. Based on efficiency reasons, the Open University decided in about 1995 that 50% of the courses should be finished by multiple-choice exams in SYS. In 2000, sys was available on all the study centres in the Netherlands as well as the study centres in Flanders.

At that time, all the exams were multiple-choice exams. However, there was a need for the same flexibility of SYS for other exams with essay questions. The use of essay questions in SYS is mainly an organisational solution. Students could not get directly after their exam the tentative score, but examiners had to evaluate the answers on the essay questions. The study centres had to send the students’ answers to the examiners and after the examiners scored the exams, they were sent to the administration for further processing.

At this stage, the Open University had a computer-based system available for about half of the summative assessments, mostly consisting of multiple-choice questions. The computer-based system covered the item construction, test construction, and reporting phases of the assessment process. The delivery phase was still on paper.

**The implementation strategy**

In 2007, the board of the Open University decided that the development of computer-based assessment belongs to one of the main cornerstones of the university. A project team was set up with people of different departments of the Open University. The project’ assignment was to make a well-considered choice for item banking software.

The project team started their activities by setting up a list of requirements for item banking at the Open University. The requirements list had to do with the different aspects of assessment (Kousen, 2008):

1. item construction and management
2. assessment profiles and assessment construction
3. import- and export facilities
4. assessment planning, scheduling and delivery
5. scoring, analysing and reporting
6. workflow support
7. system requirements

Each aspect was further elaborated. For example, the aspect of item construction and management was elaborated in:

a. use of different item formats and possible integration of multimedia;
b. use of cases in combination with different item formats;
c. use of templates
d. structuring item banks
e. support of workflow
f. management of item versions and security of data.

Based on an extended comparison between different software packages, the Open University choose for Question Mark Perception (QMP). Although, this package did not cover all the requirements, it was the best fit.

The implementation strategy then comprised first two small pilots, an evaluation of the pilots, and a stress test of the system. Because the evaluation and the stress test gave positive results, the preparations of the large-scale implementation started. These preparations will
end in a transitional stage, in which all assessment is done by QMP or with the support of QMP.

**De pilots**

In November 2008 and January 2009, five courses were finished with a computer-based exam in QMP. Four of the exams were multiple-choice exams and one exam consisted of essay questions. The delivery of the assessments went without technical problems and results of the evaluation of the computer-based assessments were positive. Table 1 shows that the lowest evaluation score for the assessments was a 5.9. This score was influenced by the fact that the assessment contained some content errors. The highest score ($M = 8.1$, $SD = 1.3$) was given for the assessment with the essay questions. The students appreciated the assessment form and the examiner appreciated the online correction procedure.

**Table 1. Mean score and standard deviation for the use of Questionmark Perception**

<table>
<thead>
<tr>
<th>Course</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>C02222</td>
<td>13</td>
<td>7.5</td>
<td>1.3</td>
</tr>
<tr>
<td>S30211</td>
<td>13</td>
<td>5.9</td>
<td>1.4</td>
</tr>
<tr>
<td>O17311</td>
<td>22</td>
<td>8.1</td>
<td>1.3</td>
</tr>
<tr>
<td>C08321</td>
<td>10</td>
<td>6.0</td>
<td>2.2</td>
</tr>
<tr>
<td>R08181</td>
<td>78</td>
<td>7.3</td>
<td>1.5</td>
</tr>
<tr>
<td>Total</td>
<td>136</td>
<td>7.2</td>
<td>1.6</td>
</tr>
</tbody>
</table>

**The preparation of the large-scale implementation**

While the evaluation of the pilots was positive, the preparation of the large-scale implementation started. This preparation was done in several work packages. Work package 1 covers the vision and policy to computer-based assessment at the Open University and the testing and editorial issues. Work package 2 covers the elaboration of the processes. Work package 3 involves the issues of specific rules and the legal framework. Work package 4 covers the manuals of all procedures. Work package 5 covers all technical aspects, such as infrastructure and interfaces. The information from all these work packages comes together in a plan for the transition phase between the current organisation and the new organisation. The large-scale implementation will start on January 1, 2010.

**Quality improvement of the assessments**

The main reason for the implementation of computer-based testing lays in efficiency arguments. However, the software for computer-based testing, in this case Questionmark Perception, has the opportunity to improve also the quality of the assessments. Because in the past, choices for the assessment form were prompted by the possibilities of the systems and policy of that moment (50% of the courses should have a multiple-choice exam). With new assessment types in the computer-based environment it is possible to use other assessment forms that better fit the learning objectives of the courses. In addition, the use of multimedia elements gives the opportunity to connect better to competence-based education.

The paper-and-pencil delivery of exams restricted the content specialist in their test format choices. The main choice of an examiner was between multiple choice and open-ended questions. The use of pictures, videos, animations was not possible. Scalise and Gifford (2006) give the basis for new forms of assessments. They mention these new forms of assessment as innovative item formats. Figure 1 gives an overview of their innovative item formats. The overview is horizontally organized from most constrained selected response items to items that require a performance under real or simulated conditions.
The use of items out of this taxonomy can improve the test validity. The validity of an assessment is about the correctness of inferences, decisions, or descriptions made about individuals, groups, or institutions from test results. Validity can be described in three types of validity: content-related validity, construct-related validity and criterion-related validity (Messick, 1989). Especially the content validity can better be covered by these item formats than only by multiple-choice items. Content validity means that the sample of questions in a test represents the important content, skills, or behaviours of the domain of interest. Content validity is usually obtained by having knowledgeable people look at the test items and make judgements about the appropriateness of each item and overall coverage of the domain.

The choice for one of the item formats influences the guessing factor of the items. In the standard multiple-choice items with four distracters, the guessing factor is 25%. In some of the used multiple-choice question this guessing factor became higher if students knew part of the answer. In the example in Figure 2 you can see that a guessing factor of 25% in a standard 4-choice question (type 1C in the taxonomy of Scalise and Gifford) in one situation becomes a 12.5% guessing factor if the question is presented in another way (like type 4A).

**Figure**
Comparison of the same question in two different item formats.
In question 1 the student has to choose from four distracters, in question 2 the student has to make two time a selection from the distracters in the rolboxes.

Both, the connection with the learning objectives and the influence of the queeasing factor are important aspects of the implementation of computer- based assessment at the Open University of the Netherlands.

References