QUALITY ASPECTS OF OPEN SOURCE TESTING TOOLS

Friedrich Scheuermann
Ângela Guimarães Pereira

Quality Aspects of Open Source Testing Tools

Friedrich Scheuermann and Ângela Guimarães Pereira
European Commission - Joint Research Centre, IPSC
Knowledge Assessment Methodologies (KAM)
TP 361, Via Enrico Fermi 1 / 21020 Ispra, Italy
friedrich.scheuermann@jrc.it, angela.pereira@jrc.it

Abstract

This paper presents work in progress concerning the definition of quality criteria for open source computer based assessment, namely platforms for the assessment of skills. The research approach undertaken so far is based on literature reviews and expert interviews which contributed to identify a number of software applications, platforms and tools being currently reviewed according to a pre-defined matrix of descriptive and normative criteria. The results of the evaluation activities will feed the setting-up of a protocol for quality assurance of e-assessment platforms in skills assessment contexts.

Background

In 2006 the European Parliament and the Council of Europe have passed recommendations on key competences for lifelong learning and the use of a common reference tool to observe and promote progress in terms of the achievement of goals formulated in *"Lisbon strategy"* in March 2000 (revised in 2006, see http://ec.europa.eu/growthandjobs/) and its follow-up declarations in selected areas (Communication in the mother tongue, communication in foreign languages, mathematical competence and basic competences in science and technology, digital competence, learning to learn, social and civic competences, sense of initiative and entrepreneurship, and cultural awareness and expression) (European Parliament and Council of Europe, 2006). Indicators for the identification of such skills are now needed, as well as instruments for carrying out large-scale assessments in Europe. In this context it is hoped that electronic testing could improve the effectiveness of the needed assessments, i.e. improve identification of skills, and their efficiency, by reducing costs of the whole operation (financial efforts, human resources etc.).

This paper describes developments within a project on e-assessment quality assessment whose overall aim is the development of quality criteria to assess e-assessment platforms and draft recommendations for such systems in contexts of skills assessment (including desirable architectures, required competencies, interoperability requirements, etc.).

In the remainder of this paper we will describe the methodology and preliminary results of a review of practice on computer based assessment, focussing on open source software applications, though including commercial options. This review's results are the basis for developing such protocol.

Research design

The research approach is framed by the need to assess skills of population groups in Europe at a large scale and to achieve accurate and comparable results for further benchmarking. Therefore, emphasis is given to tools for **diagnostic assessment and objective measurement** as the basis of research activities on *e-assessment*.

The following research questions were formulated for further orientation of the work:

1. **Potentials**: What are the potentials of testing software in relation to existing instruments for measurement? What are the implications for policy and lifelong learning?
2. **Requirements**: What types of platforms are needed in order to carry out large-scale testing in a very heterogeneous European environment which is also characterised by different infrastructures, possibilities and needs in terms of technology? What are the requirements to be respected, functionalities and features need to be taken into account for delivery?
3. **Open Source**: What is the specific added value of open source software in the context of assessment? What are the characteristics? How is it being implemented? and what are existing relevant experiences?
4. **Quality**: What are the quality dimensions to be taken into account? Which criteria can be applied for the definition of quality in open source platforms and the delivery of tests?

These questions are probed into the differential experiences of actors, such as policy-makers, test developers, test takers and test administrators, being derived from literature reviews and interviews.

Furthermore, an in-depth evaluation of a selected choice of platforms drawn from a vast range of tools identified in Internet sites and literature, using a pre-defined matrix is carried out. The evaluation is based on a mix of inspection and test methods applied to system usability as well as taking into account different phases and stages during the broader context of the assessment process.

The results of the work will be revised in several steps through a peer-reviewed process with European expert researchers and practitioners.

Instruments

The matrix of criteria for software evaluation was produced based on literature review and internet search. This is an on-going process, which also allows addressing the general context of testing and to identify relevant products and methodologies, as well as key actors in the field. Based on this research, an overall analysis of potentials and threads from a user's perspective (test taker, test developer, test administrator) was carried out and set into context of selected application areas, such as languages.

The evaluation matrix is composed of a set of categories derived from literature review and refined by the analysis of a selected number of randomly chosen applications. The matrix takes stock of initial work by Bergstrom et al. (2006) who have developed and applied a tool for assessment and online delivery. Apart from administrative data the adapted matrix contains assessment items, such as:

- Availability (URL, CD-ROM, Demo etc.)
- Licence/Costs (Open Source, Freeware, Commercial etc.)
- Delivery Method (Internet/Web-based, stand alone, secure site)
- Type (tool, platform, service etc.)
- General features (Specific assessment functionalities, administrative functionalities, communication etc.)
- Field of Application (context, such as Languages, personal skills assessment)
- Purpose (e.g. self-assessment, peer-assessment)
- Function (diagnostic, summative, formative)
- Target group(s) (Age, profession etc.)
- Outcomes (expected outcome of assessment activity, to which the tool is enabling)
- Item Types (MC, open questions etc.)
- Language(s)
- Standards (Is reference made to any applied standard?)
- Quality assurance (Is reference made to any specific quality assurance measure?)
- Interface/ Access Restrictions (e.g. open access, restricted access)
- Hardware/Software Requirements
- Stakes (high, medium, low)
- Assessment algorithms?

A first categorisation of products aimed at selecting platforms according to their relevance for the project. Categorisation and relevance of software is based on the degree of compliance of the platforms for the following features:

- Diagnostic testing

- Objective measurement
- Platform, covering all phases and steps of assessment
- Proctored, internet-based assessment features
- Multilingual or potential to deliver in multilingual versions
- Availability/accessibility for evaluation
- At a later stage: open source license

Finally, contextualised experiments will be carried out with a limited number of software products in order to identify and verify quality indicators, which in turn will contribute to a first version of a quality criteria checklist for e-assessment platforms. The work will be peer reviewed by experts' workshops, leading to a tuned version of such platforms.

Platform evaluation

The starting point of the analysis is the expected benefit from testing measurements in general and from supportive electronic environments. Testing activities can be fully based on ICT platforms or just enhanced by ICT in addition to other forms of the assessment process (e.g. some types of "blended assessment" mixing different ways of delivery). From our revision of the existing literature it seems as though that there are almost as many criteria as there are contexts, scenarios and stages for testing. Such criteria relate to the adequateness of assessment methodologies (from a psychological/psychometrical, pedagogical perspective), technical features and specifications as well as to socio-economic reflections. However, few experiences are documented to provide a sound overall picture of the complete scope and process of effective and efficient computer-based test delivery.

A first classification of products and services aimed at separating those items into those of relevance for this project. They were classified according to the above mentioned types and then selected on the basis of availability, features provided and licences. Separation of software into open-source and commercial (including shareware, freeware etc.) types was not considered to be appropriate at this stage since we would like to keep an overview of the state-of-the-art and innovative solutions, which outlines promising potentials for future applications in skills assessment, in particular.

There exist a large number of electronic tools on the market supporting assessment activities. Such tools are offered either as

- specific functionality of (educational) platforms that enable the management of (usually multiple-choice) items together with the administration and server- or web-based delivery of tests (e.g. Moodle, http://www.moodle.org),
- survey development tools (e.g. Hot Potatoes, http://hotpot.uvic.ca),

- tools dedicated to data collection and analysis of results/measurement (e.g. OpenSurveyPilot, http://www.opensurveypilot.org/)
- management tools, e.g. for documentation, reporting (including grading tools, classroom/pupil assessment administration) (e.g. Gradebook 2.0, http://www.winsite.com/bin/Info?2500000035898)
- assessment platforms, covering the complete process of assessment activities (e.g. TAO, http://www.tao.lu), or
- assessment services (e.g. Pan Testing) covering a wide range of (tailor-made or standard) activities proposed depending on specific needs. Such services are usually offered by commercial enterprises (ASP).

So far, based on literature review and internet search, more than 460 products and services were identified which then have been explored and classified according to the categories defined earlier. As a consequence, based on the features listed earlier, a list of assessment platforms was derived, out of which 3-5 will be tested in a next step of the project.

Many tools and applications are being developed by commercial enterprises with specific services on well-focussed areas. However, availability of platforms for test delivery is limited. An example for such a platform is TAO (*Test Assisté par Ordinateur*) system (See: Plichart, Jadoul et al. 2004 and http://www.tao.lu)TAO is a modular platform for internet-based computer aided testing. The platform allows the management of knowledge pertaining to subjects (individuals whose competencies and knowledge may be assessed), groups of subjects, tests and items (elements of tests requiring an answer from the user). TAO is said to be a flexible and distributed system since it uses meta-data for resource description formalised through Semantic Web standard language RDF/S. In the words of the TAO authors any sort of testing in several domains, including accreditation and even surveying could usefully deploy this open source (OSS).platform. This system is still under development, although a full prototype already exists. The TAO system has not undergone major testing. Also, according to the authors it has much more potential than existing assessment platforms, being a dedicated assessment platform, the elements and properties of which, provide the link with psychometric theory (item parameters and characteristics, testing algorithms etc.) being explicitly built into TAO, but still open for relevant tailoring. The platform is in principle interoperable with other electronic applications.

Its main assets, regard the open shell concept that allows easily specific functionality to be added as a plug-in; currently it includes a variety of assessment models, as well as possibilities for having construction of items other than just multiple choice, in addition to a user friendly interface from the point of view of the test taker. However, the platform is not yet developed on industrial standards due to lack of funding.

One of the reasons to go Open Source in these types of platforms is to try to boost through a community of users further developments. This project will try

to verify this statement at a later stage. During our software review, a great deal of that what is presented under this *branding* is not corresponding to that what is commonly understood as “Open Source” in terms of the availability of open source code (see for instance the OSI, http://www.opensource.org/). In many cases this software is declared as “work in progress” to be published at a later stage or, as in most cases, out of date and not anymore accessible.

Final remarks

Results of the analysis of selected platforms will be presented during the conference event. Furthermore, a preliminary version of quality indicators and criteria will also be presented.

References

Bergstrom et al. (2006). Defining Online Assessment for the Adult Learning Market. In: Online Assessment and Measurement. M. Hricko and S.L. Howell. Hershey, London, Information Science Publishing: 46-47.

European Parliament and Council of Europe (2006). Recommendation of the European Parliament and of the Council on key competences for lifelong learning**:** 10-18

Plichart, P. et al (2004). TAO, a collaborative distributed computer-based assessment framework built on Semantic Web standards. In: International Conference on Advances in Intelligent Systems – Theory and Applications; AISTA2004. Luxembourg