International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 18, Number 6

September – 2017

An Evaluation Framework and Instrument for Evaluating e-Assessment Tools

UG Singh[1] and **MR (Ruth) de Villiers**[2]
[1]University of KwaZulu-Natal, Durban, [2] University of South Africa, Johannesburg, South Africa

Abstract

e-Assessment, in the form of tools and systems that deliver and administer multiple choice questions (MCQs), is used increasingly, raising the need for evaluation and validation of such systems. This research uses literature and a series of six empirical action research studies to develop an evaluation framework of categories and criteria called SEAT (Selecting and Evaluating e-Assessment Tools). SEAT was converted to an interactive electronic instrument, e-SEAT, to assist academics in making informed choices when selecting MCQ systems for adoption or evaluating existing ones.

Keywords: action research, e-assessment, evaluation criteria, evaluation framework, evaluation instrument, multiple choice questions, MCQs

Introduction

e-Assessment is the use of information technology in conducting assessment. There is a range of genres, involving the design of tasks and automated activities for assessing students' performance and recording results. Examples are multiple choice questions (MCQs); e-portfolios; onscreen marking; and development by students of electronic prototypes and artefacts (Stodberg, 2012; Thomas, Borg, & McNeill, 2014). e-Assessment is particularly valuable in assessment of large cohorts, as well as in open and distance learning (ODL), where it is crucial to successful teaching and testing. This research focuses on tools and systems that deliver and administer MCQs, addressing the need to evaluate and validate them. We describe the generation of an interactive evaluation framework that assists academics in making decisions when selecting MCQ systems for adoption or evaluating existing systems. The work was conducted in a higher education context.

MCQs include single- and multiple-response questions, true/false, true/false with explanation, matching items, extended matching items, drop-down lists, fill-in-the-blank/completion, hotspots, drag-and-drop, diagrams/video clips, simulations, ranking, re-ordering, and categorising (Singh & de Villiers, 2012;

Stodberg, 2012). Their advantages include rapid automated marking (grading); replacement of labour-intensive traditional marking; objective unbiased marking; specified durations or open-ended periods; question banks; and coverage of broad ranges of topics. MCQs provide higher reliability than constructed-response questions and are equally valid (Mendes, Curto, & Coheur, 2011; Ventouras, Triantis, Tsiakas, & Stergiopoulos, 2010). Furthermore, item analysis and item response theory allow educators to evaluate MCQs in terms of difficulty and discriminative capacity (Costagliola & Fuccella, 2009). There are cognitive benefits, as testing with good MCQs supports comprehension, knowledge verification, and achievement of course objectives (Costagliola & Fuccella, 2009). Formative assessment via MCQs is useful for revision (Farthing, Jones, & McPhee, 1998) and the feedback supports further learning (Malau-Aduli, Assenheimer, Choi-Lundberg, & Zimitat, 2013). Importantly, MCQs offer a valuable option for assessment especially in open distance learning (ODL) due to its time- and place-independence.

Drawbacks of MCQs are that they do not assess application of knowledge for problem solving (Engelbrecht & Harding, 2003) and they are also criticised as being unrelated to authentic practice. However, research shows that it is possible to test higher-order thinking through well-developed MCQs, but it requires skill, practice, and time on the part of the educator (Luckett & Sutherland, 2000; Mitkov & Ha, 2003; Singh & de Villiers, 2012).

Research Problem and Gap Identification

e-Assessment via MCQs has become an integral and increasing form of assessment (Pretorius, Mostert, & de Bruyn, 2007), particularly with large student bodies and growing faculty workloads at higher-education institutions (HEIs). Concomitantly, the need arises for frameworks and means of evaluating e-assessment systems in use or being considered for adoption (Thomas, Borg, & McNeil, 2014). Insufficient research has been conducted on requirements for e-assessment systems and their evaluation (Scalter & Howie, 2003; Valenti, Cucchiarelli, & Panti, 2002) and a gap exists:

- Wills et al. (2009) generated FREMA, an e-learning framework for assessment. FREMA is not an evaluation framework; rather, it is a reference model that provides a structured network of resources to developers of e-learning assessment.
- Thomas, Borg, and McNeill (2012) produced a process-focused life-cycle framework to link stages of e-assessment to institutional strategies for developing e-assessment.
- Factors contributing to low adoption of e-assessment at an HEI in America were analysed by McCann (2009).
- In South Africa, Pretorius et al. (2007) reported that inadequate information exists on evaluation criteria for MCQ systems.
- The above group recently compiled a list of 104 criteria in four categories against which computer-based training (CBT) systems can be evaluated to meet needs in their institution (Mostert, de Bruyn, & Pretorius, 2015). These criteria for a “perfect” CBT system are based on literature, requirements of faculty, personal experience with systems, and best-practice principles, but are not accompanied by an evaluation instrument.

The present work aims to fill the gap and address the problem by generating an innovative, comprehensive, and multi-faceted framework for evaluating electronic MCQ systems. Using an action research approach comprising six iterative studies, we developed, validated, applied, and refined a structured framework for evaluating systems and tools that deliver and assess questions of the MCQ genre. First, a framework of criteria, SEAT (Selecting and Evaluating e-Assessment Tools), was developed and evaluated. SEAT was then converted to an electronic instrument, e-SEAT, which was critiqued in further empirical studies. The final e-SEAT Instrument comprises 11 categories and 182 criteria. It generates scores and structured reports that assist faculty in selecting and evaluating MCQ tools.

We sketch the emergence of the initial SEAT Framework (Background), while the Research Methodology Section presents the research question and introduces the action research approach by which SEAT evolved to the automated e-SEAT Instrument. We then present the Development, Evaluation, Refinement, and Validation of SEAT and e-SEAT, followed by a view of the final e-SEAT Instrument. The Conclusion revisits the research question.

Background

SEAT was initially constructed by creating Component$_{\text{-LIT}}$ from literature and Component$_{\text{-EMP}}$ from empirical studies among MCQ users. The two were merged in the SEAT Framework (Figure 1), which evolved over four studies, 1a–1d, before being converted to the e-SEAT Instrument, which was validated and refined in Studies 2 and 3.

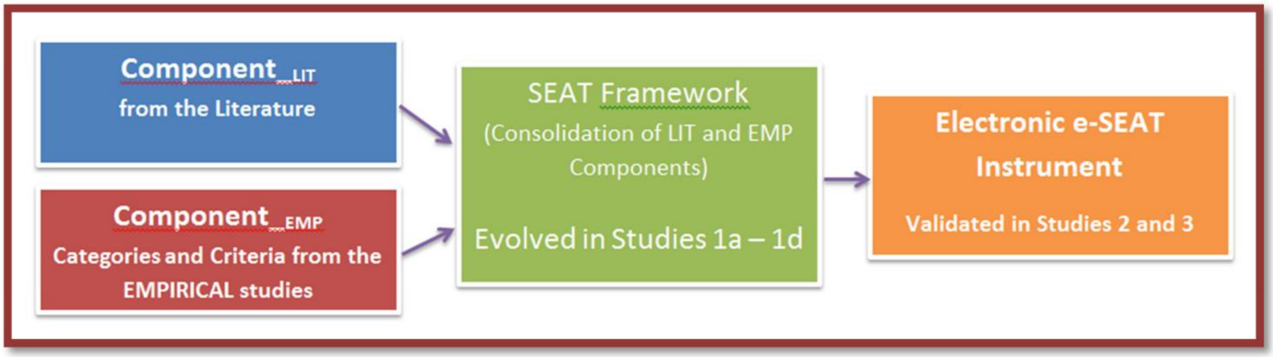

Component_LIT
from the Literature
Component_EMP
Categories and Criteria from the EMPIRICAL studies
SEAT Framework
(Consolidation of LIT and EMP Components)
Evolved in Studies 1a – 1d
Electronic e-SEAT Instrument
Validated in Studies 2 and 3

Figure 1. Evolution of SEAT and e-SEAT, adapted from Singh and de Villiers (2015).

Component$_{\text{-LIT}}$ emerged from literature studied by the primary researcher to identify pertinent criteria. Valenti et al. (2002) defined four categories of criteria for evaluating MCQ systems: Interface, Question Management, Test Management, and Implementation Issues. Pretorius et al. (2007) compiled attributes of a good e-assessment tool, using three of Valenti et al.'s categories and adding Technical, Pre-criteria (prior to usage), and Post-criteria (after usage). Component-$_{\text{LIT}}$ also contains criteria influenced by Carter et al. (2003); Lewis and Sewell (2007); and Maurice and Day (2004), resulting in a synthesis of 11 evaluation categories with 91 criteria.

Component$_{\text{-EMP}}$ emerged from interviews and questionnaires (72 and 64 participants respectively), which investigated empirically what features are required by users of MCQ systems. This research, conducted prior to Studies 1, 2, and 3 (the subject of this manuscript), generated 42 new evaluation criteria and a 12th category on Question Types (Burton, 2001; Miller, 2012; Singh & de Villiers, 2012; Wood, 2003). After integrating Component$_{\text{-LIT}}$ and Component$_{\text{-EMP}}$, there were 12 categories with 91+42=133 criteria. Some categories were merged and compound criteria were subdivided into single issues, leading to 10 categories with 147 criteria in the initial SEAT Framework: Interface Design, Question Editing, Assessment Strategy, Test/Response Analysis, Reports, Test Bank, Security, Compatibility, Ease of Use, Technical Support, and Question Types (Singh & de Villiers, 2015).

Research Design and Methodology

The research question under consideration is: *What are essential elements to include in a framework to evaluate e-assessment systems of the MCQ genre?*

The overarching research design was action research, using mixed-methods strategies (Creswell, 2014) for longitudinal investigation of the developing SEAT and e-SEAT artefacts. The studies were conducted with different groups of participants invited due to their use of MCQs and suitability for the study in hand. A number of them worked in distance education. They critiqued, evaluated, and applied the framework, facilitating evolution from the SEAT Framework to the electronic evaluation instrument, e-SEAT. Quantitative and qualitative questionnaire surveys were conducted, as well as qualitative semi-structured interviews. The questionnaires gathered uniform data from large groups, while interviews allowed in-depth exploration of interesting and unanticipated avenues with individuals. As different participants scrutinised the Framework in empirical studies, they contributed additions, deletions, and refinements to the categories and criteria of the theoretical SEAT Framework and subsequently to the practical e-SEAT Instrument.

Figure 2 shows the series of six action research studies (Singh & de Villiers, 2015).

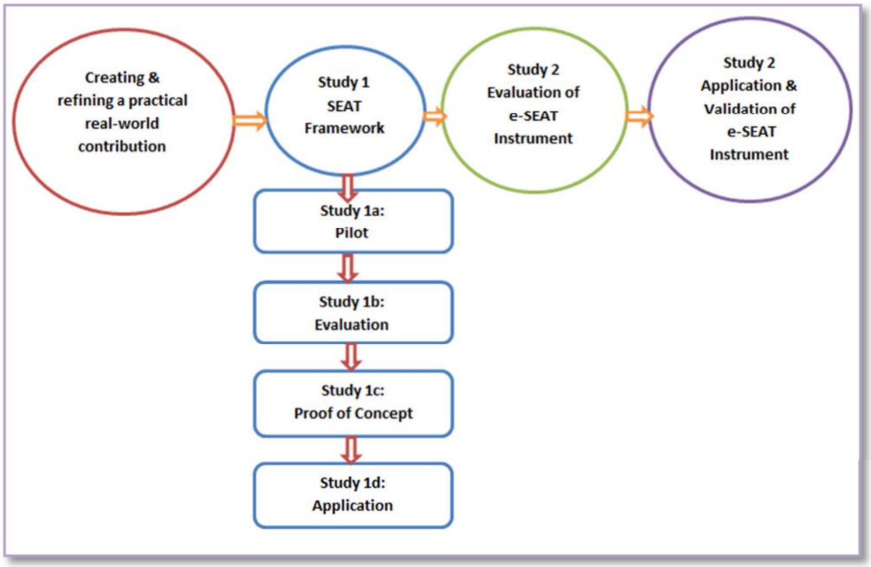

Creating & refining a practical real-world contribution
Study 1 SEAT Framework
Study 2 Evaluation of e-SEAT Instrument
Study 2 Application & Validation of e-SEAT Instrument
Study 1a: Pilot
Study 1b: Evaluation
Study 1c: Proof of Concept
Study 1d: Application

Figure 2. Action research applied to SEAT and e-SEAT instrument.

In Study 1, with four sub-studies, sets of participants inspected the SEAT Framework from varying perspectives to suggest extensions and refinements and to propose further criteria. After SEAT had been converted to the online e-SEAT Instrument, e-SEAT was investigated and refined. Study 2 evaluated the Instrument itself. Study 3 applied e-SEAT to evaluate specific MCQ systems, thus validating it by use. Participants were employees at HEIs in South Africa, particularly in computing-related disciplines: Computer Science, Information Systems, and Information Technology, but the findings are relevant to other disciplines too.

Findings of Evaluation, Application, Refinement and Validation of Seat and e-Seat

We discuss the action research series of Studies 1, 2, and 3, presenting selected findings.

Study 1 – SEAT Framework

Study 1 was iterative, comprising Studies 1a, 1b, 1c, and 1d that provided varying perspectives on the SEAT Framework with different groups of participants involved in each. The initial version of SEAT with 10 categories and 147 criteria, was a Microsoft Word document, uploaded on the online tool Survey Monkey for distribution and data analysis. In the process of conversion to Survey Monkey format, criteria were subdivided leading to 162.

Study 1a: Pilot study. In 1a, the primary researcher selected a convenience sample of two meticulous and experienced colleagues, who worked through the SEAT Framework

- rephrasing, deleting, and adding categories;
- suggesting further criteria;

- rewording criteria to improve clarity;
- identifying duplicates for removal;
- moving criteria to categories where they fitted better; and
- explaining terms.

The structural-, content-, and system-related input from the Pilot was used to create the second version of the Framework, with 10 categories and 166 criteria.

Study 1b: SEAT Evaluation study. This major study continued the action research process of refining SEAT as a prelude to developing an electronic framework. The Framework was distributed to 80 users of MCQ systems from 16 HEIs. Fifty-six (70%) returned usable responses. The survey on Survey Monkey listed the criteria, along with an evaluation item for each. Participants rated the importance of each criterion on a scale from 1 (extremely important) to 7 (not at all important). The quantitative data was statistically analysed with the non-parametric sign test to assist identification of essential criteria (mean ≤3) and non-essential criteria (3<mean<6). Six criteria in the categories of Test and Response or Ease of Use scored mean ratings significantly >3 but <6, indicating they should be removed (Table 1).

Table 1

Identification of Criteria for Removal

Category	Number of Criteria	Number with mean score:		
		≤3	>3 and <6	≥6
1 Interface Design	10	10		
2 Question editing	26	26		
3 Assessment strategy	9	9		
4 Test and Response analysis	44	39	5	
5 Test Bank	2	2		
6 Security	18	18		
7 Computability	8	8		
8 Ease of use	23	22	1	
9 Technical support	10	10		
10 Question types	16	16		

Participants also provided qualitative comments. In general, they found SEAT comprehensive and helpful. They confirmed that vital criteria had been identified, and recommended several others. Selected responses follow:

R2: "All statements are obvious rules."

R4: "These criteria...enable developers...to customise and enhance their tools."

R14: "There was not a single item that I would not want the option of including in an assessment tool."

R28: "The survey made us re-think our online assessment and its alignment with our lecturing."

R24 and R46 found the list too long.

Study 1c: Proof of concept study (PoC). Oates (2010) explains that not all researchers formally evaluate designed artefacts. Instead, they might conduct a PoC by generating a prototype that functions and behaves in a required way under stated conditions. In this case, the PoC involved both a functioning prototype and an expert evaluation. The researcher hand-picked a purposive sample of three experts from different HEIs, who had been closely involved with MCQ assessment. They inspected SEAT through varying lenses – Participant One (PoC1) was an e-learning manager in an ODL institution, PoC2 an academic leader responsible for strategic decisions regarding adoption of e-assessment tools, and PoC3 a senior academic, who had specialised in MCQs for more than five years. Study 1c comprised a survey and follow-up telephonic interviews regarding the participants' comments, and reasons for low ratings on some criteria that had been considered essential in Study 1b.

PoC1 suggested minor changes to wording to improve the clarity of the criteria in the framework.

PoC2 was more critical, recommending removal or rewording of several criteria. He suggested adjusting the rating scale. At that stage, criteria was rated on a Likert scale from 1, "Extremely important," to 7, "Not at all important." PoC2 advocated a more qualitative ranking, by which participants would evaluate how effectively each criterion was implemented in the system being rated. He advised a scale from "Very Effectively" to "Not at all," and a "N/A" option.

PoC3 suggested a fundamental structural improvement. He advised an 11th category, Robustness, and advocated that each SEAT category should be assigned to one of two overarching sections, "Functional" or "Non-Functional." PoC3 acknowledged SEAT's usefulness, "SEAT is invaluable to decision-makers considering the adoption of e-assessment," and stated that "It (SEAT) is a wonderful idea and might be excellent to guide an institution in decision making...benefiting most stakeholders."

After reflection on the feedback of Study 1c, SEAT's terminology was adapted considerably.

Study 1d: SEAT application study. Study 1d was the last SEAT evaluation delivered via Survey Monkey. A purposive sample of seven users with expertise in e-assessment and MCQs was selected from participants in the earlier Component-$_{\text{EMP}}$ interview study (see Background Section). They had contributed requirements that were converted to criteria, hence it was important now to get their feedback on the emerging Framework. The participants comprised five academics who had used MCQs extensively for at least five years, a reputed e-consultant, and a leading e-assessment researcher. They applied the version called SEAT Application Framework to evaluate an MCQ system they used, then provided constructive criticism. Four applied SEAT to the tool embedded in their own institution's learning management system. Participants A5, A6, and A7 completed the Survey Monkey template, but did not respond to the open-ended questions. Selected qualitative responses from A1, A2, A3, and A4 follow:

> A4 found "SEAT Framework was easy to use, easy to implement and easy to administer, but requires some thought."

Most responses mentioned using SEAT for considering potential acquisitions:

> A1: "It is very valuable for comparing e-assessment tools…saves time."
>
> A2: "In a teaching environment, the value of this instrument lies in empowering users to make a better choice between various computer-based testing packages. This is valuable…especially for the department responsible for choosing the online software."
>
> A3: "It is most useful when considering purchasing a system…I would love the criteria to be given to the owners of the system I am using, which I believe is not suitable for universities."

These remarks are encouraging, because the use of SEAT in evaluating tools for adoption, was a main intention of this research. Moreover, A2 proposed, "This can be used as a bench-marking instrument for online assessment tools." In lateral thinking, A3 believed, "This framework could be used to show non-users…the wonderful features of a system." With relation to features and scope, A4 found SEAT "one of the few comprehensive tools available. It provides an overview of the most important features." A4 felt that "the length of the instrument is essential." A1 affirmed the successful evolution of the Framework, "All the relevant questions about an e-assessment tool are already there. You can just answer the questions to evaluate the tool." There were no suggested deletions, but two new criteria were proposed for the Technical Support category.

Consolidation of Study 1. The Framework was consolidated in line with Studies 1b–1d, incorporating the category of Robustness (Study 1c) and adding/removing criteria. SEAT then comprised 11 categories and 182 criteria:

- *Functional*: Question Editing, Assessment Strategy, Test and Response Analysis, Test Bank, Question Types.
- *Non-Functional*: Interface Design, Security, Compatibility, Ease of Use, Robustness, Technical Support.

This version served as the basis for the e-SEAT Instrument. A computer programmer took the categories and criteria of the SEAT Framework and constructed an interactive electronic evaluation instrument to analyse users' input on MCQ systems/tools and to provide automated scoring for each criterion, calculations, and reports. e-SEAT generates category ratings and an overall rating. In Studies 2 and 3, selected participants evaluated, applied, and validated e-SEAT. Their feedback was used to correct problems in the automated version, and to refine terminology, but no further changes were made to the criteria.

Studies 2 and 3 – e-SEAT Instrument

These two studies are discussed together, since they overlap. Study 2, the e-SEAT Evaluation Study, was an expert review to assess the effectiveness and appropriateness of e-SEAT. A purposive sample of four expert MCQ users from different South African HEIs, was invited. None of them had participated in previous studies, hence they interacted with e-SEAT in an exploratory way and gave fresh objective impressions. After using e-SEAT, they were required to complete an evaluation questionnaire, followed by an unstructured interview.

The culmination of the action research series was Study 3, Application and Validation of the e-SEAT Instrument. Three participants, selected for their expertise and experience, critically reviewed the Instrument to validate it and to apply it in a comprehensive and complete evaluation of an MCQ system they used. They covered all the categories and criteria rigorously, validating e-SEAT by use in practice.

The common component in the two studies was the questionnaire completed by participants after they had used e-SEAT. Table 2 integrates the quantitative ratings from Studies 2 and 3, with four and three participants respectively, totaling n=7.

Table 2

Integration of Ratings from Studies 2 and 3

The e-SEAT instrument: (n=7)	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
was useful in evaluating my e-assessment tool	2	4	1		
was intuitive to use	1	4	1	1	
provides useful information in the report	1	3	3		
e-SEAT lacks certain:					
usability features	1	2	1	3	
content features				3	4
processing features			1	2	4
I could use e-SEAT without referring to the instruction file	Yes	6	No	1	
	(It was suggested that the instructions should be online context-sensitive help, rather than a separate document)				

Disregarding outliers, the ratings were similar and positive, with six and five participants agreeing/strongly agreeing that e-SEAT was respectively useful and intuitive, while four were pleased with the report. Ratings on usability were mixed, indicating that e-SEAT's usability needs improvement. Negatively-phrased items considered lacks, whereby seven and six strongly disagreed/ disagreed that content and processing were respectively lacking.

Study 2 elicited open-ended feedback by qualitative questions and interviews on matters such as potential beneficiaries, features they liked or found irritating. Addressing e-SEAT's repertoire, ESP4 (Evaluation Study Participant 4) praised "the depth and range of questions," ESP2 reported "the comprehensive coverage is outstanding!" and ESP1 identified the "technical features were also addressed." Referring to the automation, ESP3 found "the format was easy to use," and ESP2 expressed that "e-SEAT prompted me to investigate aspects where I was unsure whether my tool had such options." Via the 182 criteria, participants encountered a range of factors related to MCQ assessment. ESP2 and ESP3 indicated they now grasped the numerous aspects. Participants discussed e-SEAT's relevance and worth to different stakeholders:

- Helpful to non-users: "academics who intend using e-assessment in future" (ESP1);
- "People choosing between tool options or wanting to evaluate their existing tool, would benefit. e-SEAT highlights positive aspects, as well as missing features" (ESP2);
- "Decision-makers would benefit greatly if they had previously worked with MCQs" (ESP4);
- "An assessor who is planning to use e-assessment... and does not fully know the features of such systems, might not be able to appropriately judge a system without such a tool" (ESP4).

ESP2, ESP3, and ESP4 requested an indication of progress, showing what was complete and what was still outstanding. Other minor problems emerged: the <Print Results> button also opened an email option, it was not possible to undo actions, and some features were not automated, but needed activation. ESP1 was concerned by the length and also requested that results be automatically emailed to users in case they inadvertently clicked <Close>. Most of these problems were fixed after Study 2.

Study 3 had less qualitative information. Participants requested a few additional processing and content features, most of which were feasible, and were implemented. Following improvements after Study 2, all three found the post-use report useful. The issue of orientation arose again, emphasising the need for a progress bar.

Participants appreciated the criteria that supported them as they evaluated systems. It also showed features "...one has not even thought of!" (VSP1 (Validation Study Participant 1)). VSP3 reflected that although "a tool may have a low score for a certain feature, that feature may not be relevant to you." Further categorisation into Essential and Optional criteria would be helpful in a future version of e-SEAT.

Responding to an open-ended question regarding beneficiaries of e-SEAT, VSP1 suggested administrators, budget managers, and decision-makers considering new purchases, while VSP2 and VSP3 mentioned academics using MCQs for testing. VSP3 posited, "you can only assess a tool once you know it well" and advocated that "a database of assessments done by users knowing a tool well" should be compiled from the results of e-SEAT, so that experts' evaluations could be consulted. This pertinent issue has been raised by other stakeholders as well. It is a sensitive matter, since the owners/designers of a poorly-rated system might object. It could only be done if the licence holder granted permission.

The Final e-SEAT Instrument

This section presents the ultimate product of the action research, namely the evaluation framework with 11 categories and 182 criteria, implemented within the interactive e-SEAT Instrument. The figures that follow, illustrate what an end-user would experience. Figure 3 illustrates the process of applying e-SEAT.

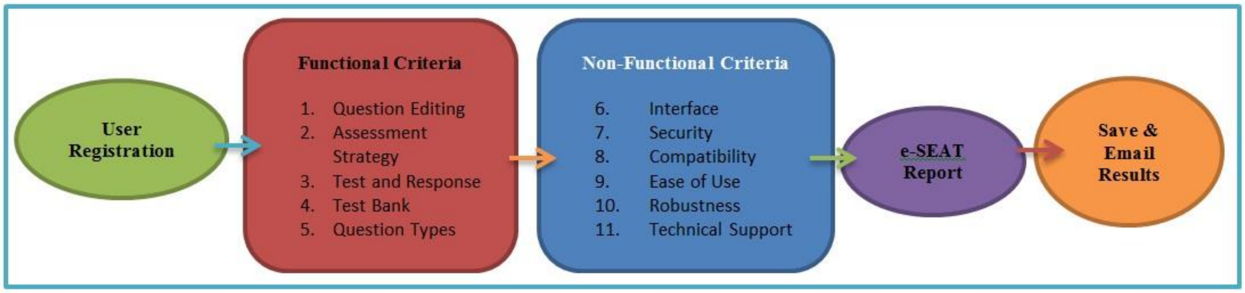
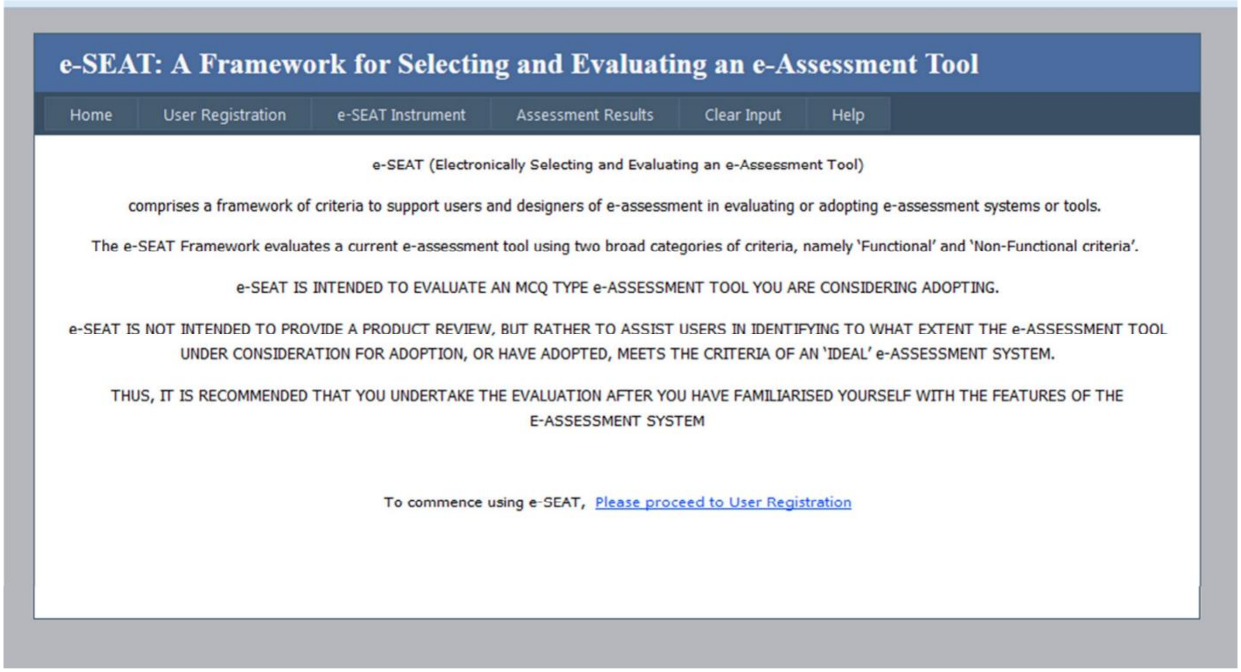


Figure 3. The e-SEAT process.

Figures 4 and 5 depict screenshots of e-SEAT's Introductory Screen and a typical rating screen respectively.



e-SEAT (Electronically Selecting and Evaluating an e-Assessment Tool)

comprises a framework of criteria to support users and designers of e-assessment in evaluating or adopting e-assessment systems or tools.

The e-SEAT Framework evaluates a current e-assessment tool using two broad categories of criteria, namely 'Functional' and 'Non-Functional criteria'.

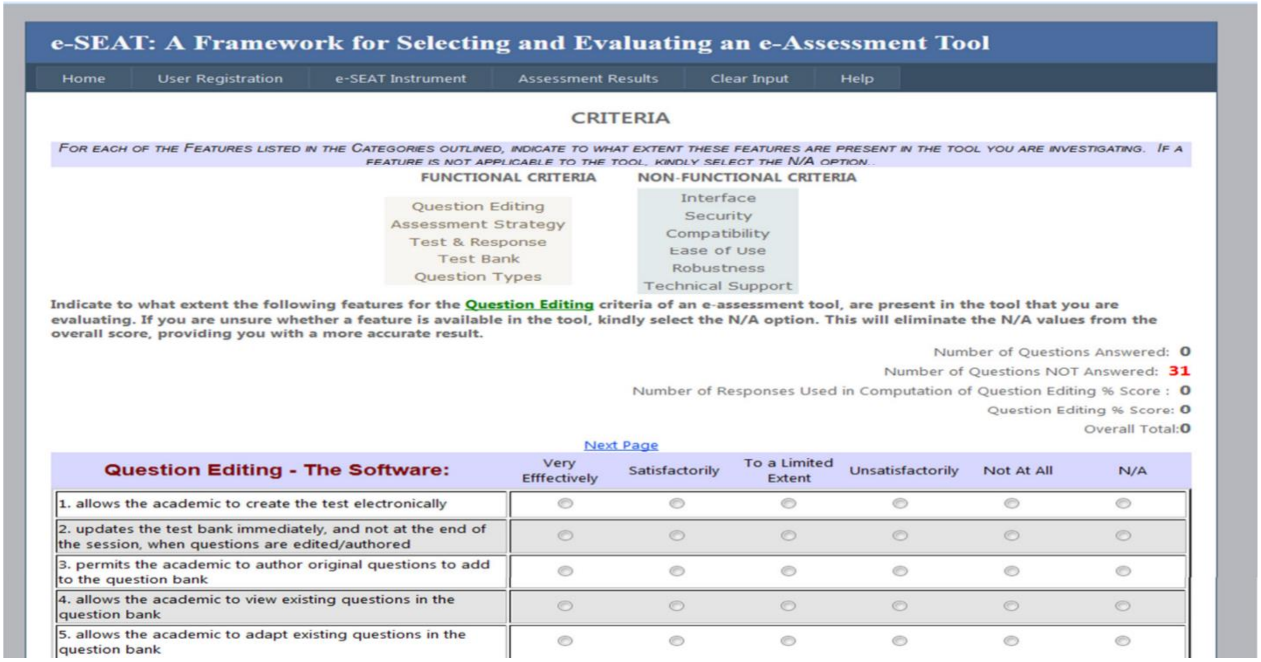
e-SEAT IS INTENDED TO EVALUATE AN MCQ TYPE e-ASSESSMENT TOOL YOU ARE CONSIDERING ADOPTING.

e-SEAT IS NOT INTENDED TO PROVIDE A PRODUCT REVIEW, BUT RATHER TO ASSIST USERS IN IDENTIFYING TO WHAT EXTENT THE e-ASSESSMENT TOOL UNDER CONSIDERATION FOR ADOPTION, OR HAVE ADOPTED, MEETS THE CRITERIA OF AN 'IDEAL' e-ASSESSMENT SYSTEM.

THUS, IT IS RECOMMENDED THAT YOU UNDERTAKE THE EVALUATION AFTER YOU HAVE FAMILIARISED YOURSELF WITH THE FEATURES OF THE E-ASSESSMENT SYSTEM

To commence using e-SEAT, Please proceed to User Registration

Figure 4. e-SEAT Introductory screen.



e-SEAT: A Framework for Selecting and Evaluating an e-Assessment Tool

CRITERIA

For each of the Features listed in the Categories outlined, indicate to what extent these features are present in the tool you are investigating. If a feature is not applicable to the tool, kindly select the N/A option..

FUNCTIONAL CRITERIA	NON-FUNCTIONAL CRITERIA
Question Editing	Interface
Assessment Strategy	Security
Test & Response	Compatibility
Test Bank	Ease of Use
Question Types	Robustness
	Technical Support

Indicate to what extent the following features for the Question Editing criteria of an e-assessment tool, are present in the tool that you are evaluating. If you are unsure whether a feature is available in the tool, kindly select the N/A option. This will eliminate the N/A values from the overall score, providing you with a more accurate result.

Number of Questions Answered: 0
Number of Questions NOT Answered: 31
Number of Responses Used in Computation of Question Editing % Score : 0
Question Editing % Score: 0
Overall Total:0

Next Page

Question Editing - The Software:	Very Efffectively	Satisfactorily	To a Limited Extent	Unsatisfactorily	Not At All	N/A
1. allows the academic to create the test electronically	○	○	○	○	○	○
2. updates the test bank immediately, and not at the end of the session, when questions are edited/authored	○	○	○	○	○	○
3. permits the academic to author original questions to add to the question bank	○	○	○	○	○	○
4. allows the academic to view existing questions in the question bank	○	○	○	○	○	○
5. allows the academic to adapt existing questions in the question bank	○	○	○	○	○	○

Figure 5. Part of the interactive screen for Question Editing Criteria.

Table 3 lists the Functional Criteria and Table 4 the Non-Functional Criteria of the e-SEAT Instrument.

Table 3

e-SEAT's Functional Criteria

Question Editing - The Software:
1. allows academics to create the test electronically
2. updates the test bank immediately, and not at the end of the session, when questions are edited/authored
3. permits academics to author original questions to add to the question bank
4. allows academics to view existing questions in the question bank
5. allows academics to adapt existing questions in the question bank
6. supports importing of questions in non-proprietary, interoperable format to the question bank
7. supports exporting of questions in non-proprietary, interoperable format from the question bank
8. permits a range of parameters/options to be specified in questions (e.g. four or five options per question)
9. supports feedback creation for each question
10. allows the incorporation of question metadata (e.g. categories, keywords, learning objectives, and levels of difficulty)
11. facilitates offline question creation within the tools

12. grants academics previews of assessments created offline
13. grants previews of assessments created online
14. incorporates an automatic grammar check facility
15. incorporates a spell checker
16. flags questions which learners have not answered in an assessment, so that they can be deleted or amended by the academic
17. allows academics to add comments to a question created by other academics, before adding to/rejecting from the question bank, where multiple editors are working on one test bank
18. allows academics to approve or reject all questions created, before adding to/rejecting from the question bank
19. directs comments regarding questions submitted to the question bank directly to the author of the question
20. allows academics to create a marking scheme for an assessment
21. allows academics to combine questions from different test banks into a single test
22. allows academics to pilot tests prior to the assessment going live
23. supports printing of tests for moderation purposes
24. records average time taken by learners for each question
25. facilitates allocation of marks to questions to support manual marking
26. provides support for incorporating graphics in questions
27. provides tools to do automatic analysis of learner responses
28. supports printing of tests to support taking the test offline
29. facilitates allocation of marks to questions to support overriding the mark automatically assigned
30. flags questions as easy, average or difficult (metadata) to support better randomisation
31. displays the IP address of the individual learner taking the test
Assessment Strategy - The Software:
1. supports random generation of questions from the test bank in multiple versions of the same assessment
2. incorporates branching of questions, depending on learners' responses (e.g. if a learner selects option (a) questions 5 to 10 are displayed, else questions 11 to 15)
3. displays feedback as/if required
4. displays results as/if required
5. specifies how many attempts a learner is permitted to make on a question
6. permits learners to sit a test as many times as they like, in the case of self-assessments
7. permits a learner to take the test at different times for different sections, in the case of self-assessments (e.g. complete Section A today, Section B tomorrow and eventually complete assessment when he/she has time)
8. permits learner to take a self-assessment offline
9. supports test templates that facilitate many types of testing including formative, peer-generated, practice, diagnostic, pre/post and mastery-level testing
10. automatically prompts learners to redo an assessment (with different questions covering

the same topics) if they get below a specified percentage
Test and Response - The Test:
1. allows groups to be set up
2. allows learners to be added to a group
3. permits questions to be viewed by metadata fields (e.g. categories, keywords, learning objectives, and levels of difficulty)
4. allows learners access to previous assessment results
5. allows learners access to previous assessment responses
6. allows learners access to markers' comments on prior assessments (in cases where a human assessor reviewed the completed test)
7. allows results to be accessed after a specific date, as required
8. allows learners to compare their results with other learners' results
9. allows learners to compare marks with group averages
10. presents results immediately to learners, when appropriate
11. provides learners with the option/facility to print assessment responses
12. distributes academics' comments to learners via the system
13. distributes academics' comments to learners via email
14. emails academics automatically if the marking deadline is not met
15. presents mean (average) score statistical analysis per assessment
16. presents discrimination index statistical analysis per assessment
17. presents facility index statistical analysis per assessment
18. presents highest score statistical analysis per assessment
19. presents lowest score statistical analysis per assessment
20. presents frequency distribution statistical analysis per assessment
21. incorporates an automated 'cheating spotter' facility
22. supports the ordering of the results tables in various ways (e.g. by marks, student numbers, names, etc.)
23. displays marks as percentages
24. presents, to the academic, all attempts at a question
25. permits the academic to view individual responses to questions
26. allows the learner to view the whole test, as he/she had completed it
27. displays a comparison of mark data of different groups
28. displays a comparison of the performance in different subtopics/sections
29. permits mark data to be viewed without having access to names of learners
30. flags questions which were poorly answered
31. flags questions which were well answered
32. the statistical analysis per assessment presents the difficulty index statistic
33. the statistical analysis per assessment presents the percentage answered correct
34. the statistical analysis per assessment presents the percentage of top learners who got the question correct
35. the statistical analysis per assessment supports correlation of assessment data across different class groups
The Test Bank:

1. draws random questions from a question bank, as required
2. only contains questions which have been moderated for the required standard and cognitive levels
3. assigns global unique identifiers to all questions created or revised in the question bank
4. has the potential to include questions that test learners’ "Higher Order Thinking Skills" (HOTS)
Question Types – The System supports:
1. Multiple choice: Single response
2. Multiple choice: Multiple response
3. True/false
4. True/false with explanation
5. Fill-in-the-Blanks/Completion
6. Simulation
7. Matching Items
8. Extended Matching Items (EMIs)
9. Selection/Drop-down-lists
10. Ranking
11. Diagrams/Graphics
12. Video/Audio Clips
13. Drag-and-Drop
14. Reordering/Rearrangement/Sequencing
15. Categorising
16. Hotspots
17. Hotspot (Drag and Drop)
18. Text Input (short answer – marked manually)

Table 4

e-SEAT’s Non-Functional Criteria

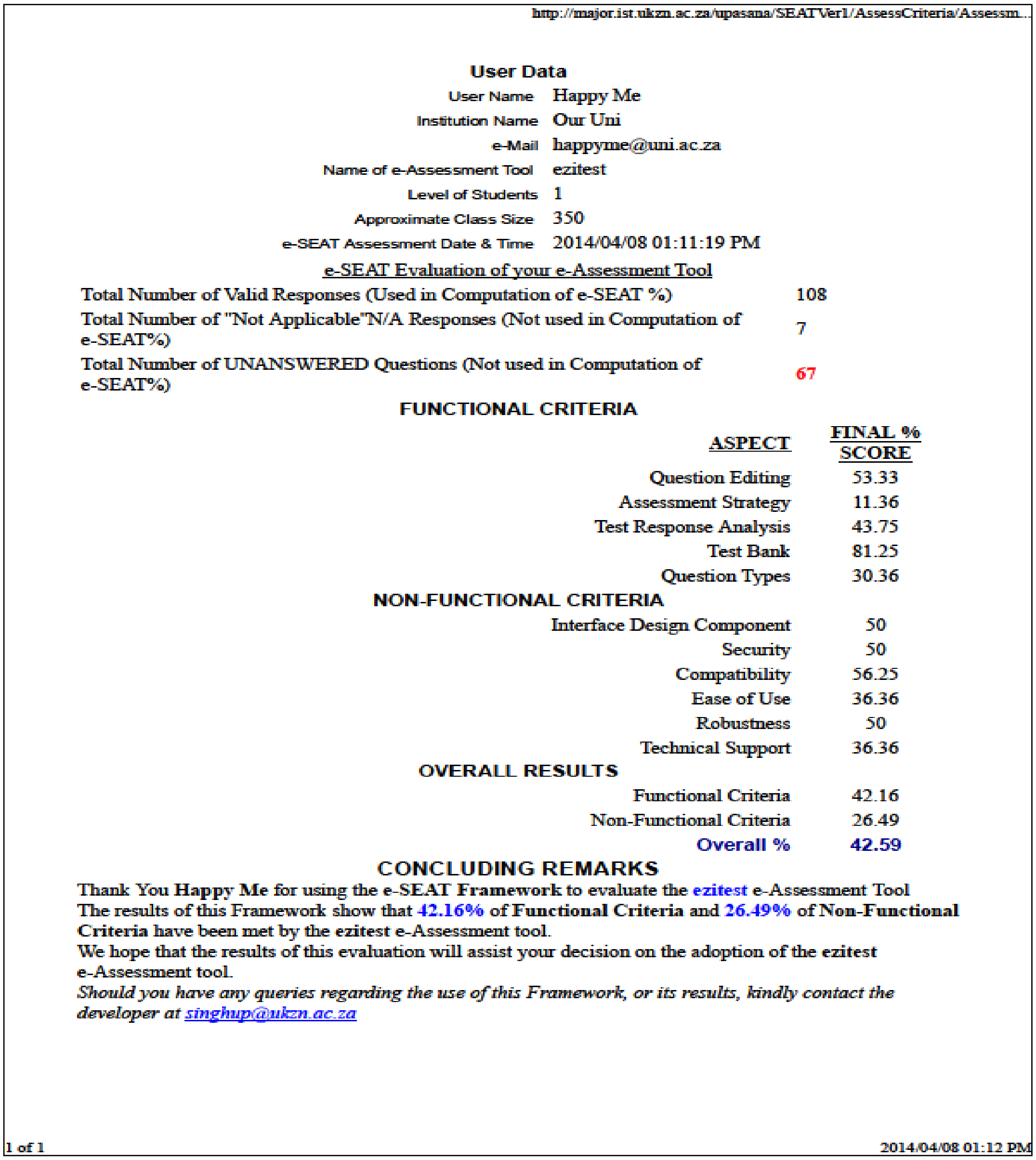
The Interface:
1. is intuitive to use
2. caters for users with special needs, by including features such as non-visual alternatives, font size variety, colour options
3. facilitates ways of varying the presentation of tests
4. allows learners to view all tests available to them
5. permits learners to view logistical arrangements in advance, such as times and venues of assessments
6. permits viewing of multiple windows as required for assessments
7. allows academics to email reminders to students of assessments due
8. provides an option to clearly display marks for each question
9. provides an option to clearly display marks for each section
10. displays a clock to keep track of time allocated/remaining for formative assessment

11. allows academics to SMS reminders to students of assessments due
12. provides a toggle button to allow students the option to answer individual questions, or the whole assessment
13. presents help facilities for users
14. provides an option to allow/disallow printing
Security Criteria - The Tool:
1. ensures that tests are accessible only to learners who have explicit authorisation, granted by access administrators
2. encrypts all data communicated via the network
3. ensures that mark data held on the server can be accessed by authorised persons only
4. logs the IP address where each learner sat
5. logs which questions were marked by which lecturer
6. logs when the academic marked the question
7. prevents answers to questions already completed from being altered (in cases where second opportunities are not permitted)
8. requires permission of the academic before any question can be modified or deleted from a test
9. prevents learners from amending a test once taken
10. prevents learners from deleting a test once taken
11. automatically allocates a global unique identifier to tests
12. provides the ability to view entire tests for verification without the ability to change them
13. restricts tests to particular IP addresses and domains
14. allows academics to enter details of learners who cheat to alert other colleagues of 'problematic' students
15. permits academics to modify results after communication with a learner regarding the reason for the change
16. permits test results to be changed or corrected when a memorandum error is discovered
17. logs modifications to original marks
18. records motivations for modifications to original marks
19. provides password access to tests
20. allows academics to restrict assessments to a specific IP address
21. prevents learners from opening any other windows not required for the assessment (similar to Respondus lockdown facility)
Compatibility - The Tool:
1. is accessible from a standard, platform- independent web browser, without additional plugins
2. is downgradable for learners with previous versions of browsers
3. is customisable to provide a uniform interface with the rest of the institution’s intranet or virtual learning environment
4. links seamlessly with other institutional systems, so that learners can use their existing username and passwords
5. permits results to be exported to spreadsheets or statistical analysis software
6. uses a common logon facility, that integrates with other institutional systems

7. links seamlessly with other institutional systems so academics can export marks directly
8. specifies which browser must be used for an assessment in the setup details
Ease of Use - The System:
1. requires little time to capture data related to learner profiles and assessments
2. requires a short time period to set up an assessment online
3. requires little/no training on how to use the tool
4. provides simple and fast login procedures
5. includes an intelligent help system – dependent on the academic role and current activity
6. incorporates speech synthesis for 'special needs' learners
7. is intuitive to use – academics should not require any special programming language skills to adopt the tool
8. makes it easy to include multimedia elements in test items
9. allows academics access to details of times of an assessment
10. permits all learners in a group to be removed from the system simultaneously
11. allows access to details of learners sitting a test at a particular time
12. permits learners to return to the point at which they had exited an incomplete self-assessment
13. makes it easy, where necessary, to enter foreign characters and symbols
14. automatically distributes electronic certificates of test submission to learners
15. allows learners access to details of room numbers and venues of an assessment
16. allows learners access to details of times of an assessment
17. simplifies the task of adding learner access
18. simplifies the task of removing learner access
19. simplifies the task of editing learner access
20. allows learners to be enrolled on the system by an administrator
21. allows learners to be removed from the system
22. permits academics to enter learner details (name and student number) in the test directly
23. allows academics to limit a test by giving learners a unique number to access the test
Robustness - The Tool:
1. does not hang while a student takes a test
2. is stable, even when a large number of learners access the system or take a test simultaneously
3. does not crash frequently
4. is able to recover the test from the point at which the learner stopped, in the event of an unforeseen system error or crash
5. processes responses given by learners in an acceptable time period
Technical Support - The System:
1. incorporates a resilient network
2. if not web-based, includes software that is easy to install, requiring little effort and time
3. runs on multiple platforms
4. includes installation software that is easily available
5. allows new functionality to be incorporated without reinstalling the system
6. supports large numbers of concurrent learners are logged in simultaneously

7. supports multi-format data storage – Oracle/Access or ODBC (Open Data Base Connectivity) format
8. facilitates the use of existing database systems
9. grants academics access to details of all test purchases relevant to that academic, where tests are purchased from the supplier of the assessment software
10. automatically prompts learners to redo an assessment (with different questions covering the same topics) if they get below a specified percentage
11. automatically prompts learners to redo an assessment (with questions they previously answered correctly removed from the new assessment) if they get below a specified percentage

After a user has evaluated an MCQ tool, e-SEAT generates a report. It appears onscreen and is also e-mailed to the user and the researcher-designer. Figure 6 shows a report regarding a fictitious system called *Ezitest*.



http://major.ist.ukzn.ac.za/upasana/SEATVer1/AssessCriteria/Assessm...

User Data

User Name	Happy Me
Institution Name	Our Uni
e-Mail	happyme@uni.ac.za
Name of e-Assessment Tool	ezitest
Level of Students	1
Approximate Class Size	350
e-SEAT Assessment Date & Time	2014/04/08 01:11:19 PM

e-SEAT Evaluation of your e-Assessment Tool

Total Number of Valid Responses (Used in Computation of e-SEAT %)	108
Total Number of "Not Applicable"N/A Responses (Not used in Computation of e-SEAT%)	7
Total Number of UNANSWERED Questions (Not used in Computation of e-SEAT%)	67

ASPECT	FINAL % SCORE
FUNCTIONAL CRITERIA	
Question Editing	53.33
Assessment Strategy	11.36
Test Response Analysis	43.75
Test Bank	81.25
Question Types	30.36
NON-FUNCTIONAL CRITERIA	
Interface Design Component	50
Security	50
Compatibility	56.25
Ease of Use	36.36
Robustness	50
Technical Support	36.36
OVERALL RESULTS	
Functional Criteria	42.16
Non-Functional Criteria	26.49
Overall %	**42.59**

CONCLUDING REMARKS

Thank You **Happy Me** for using the **e-SEAT Framework** to evaluate the ezitest e-Assessment Tool
The results of this Framework show that **42.16%** of **Functional Criteria** and **26.49%** of **Non-Functional Criteria** have been met by the **ezitest** e-Assessment tool.
We hope that the results of this evaluation will assist your decision on the adoption of the **ezitest** e-Assessment tool.
Should you have any queries regarding the use of this Framework, or its results, kindly contact the developer at singhup@ukzn.ac.za

1 of 1 2014/04/08 01:12 PM

Figure 6. e-SEAT Sample Report Screen.

Conclusion

We revisit the research question: What are essential elements to include in a framework to evaluate e-assessment systems of the MCQ genre?

This research makes a valuable theoretical contribution, filling a gap with the comprehensive SEAT Framework. The contributions from experts provided pertinent judgements and further content to the evolving artifacts. The extensive compilation, comprising 11 evaluation categories and 182 criteria, provides a conceptual understanding of the requirements and features of tools that administer questions of the MCQ genre. It emerged that different MCQ systems cumulatively provide multiple functionalities, beyond the familiar ones. Hence, the generic Framework includes a broad set of criteria relating to features and facilities that are not required in all systems, but that can be reduced and customised to specified requirements or contexts. Furthermore, the criteria can serve as sets of design guidelines for producing new assessment systems, thus benefitting designers and developers. In some cases it would be excessive for users to evaluate their systems with 182 criteria, hence e-SEAT could be made customizable in future so that users could choose which categories are essential to them.

The work is innovative in its practical contribution, namely, the e-SEAT Instrument, which is the interactive artifact on which the theoretical SEAT Framework resides and is delivered to users. The action-research approach served well in supporting step-by-step development, as the series of studies facilitated e-SEAT's evolution and improvement. Participants acknowledged its utility for evaluating e-assessment systems, particularly when such were under consideration for potential acquisition. Importantly, participants identified inadequacies that were corrected. With its click functionality and automated ratings, e-SEAT expedites the process of evaluating MCQ systems thoroughly, prompting users to consider factors that, independently, they might never have investigated.

The outcomes of this work are useful to various stakeholders: educational institutions, due to the accessibility of information on the quality of their existing MCQ tools or tools they are considering adopting; academics/faculty who wish to implement e-assessment in the courses they teach; students who appreciate rapid-results MCQ technologies to supplement traditional assessment; and in particular to ODL institutions where in the absence of class-based teaching, some degree of e-assessment is essential.

In future research, work could be undertaken to improve e-SEAT's usability and the report it generates, since responses on these aspects were tentative. Finally, measures are under way to convert e-SEAT to a fully operational system and obtain gatekeeper consent to make it officially available to other institutions.

References

Burton, R.F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment and Evaluation in Higher Education, 26*(1), 41-50.

Carter, J., Ala-Mutka, K., Fuller, U., Dick, M., English, J., Fone, W., & Sheard, J. (2003). How shall we assess this? *ACM SIGCSE Bulletin, 35*(4), 107-123.

Costagliola, G., & Fuccella, V. (2009). Online testing, current issues and future trends. *Journal of e-Learning and Knowledge Society (Je-LKS), 5*(3), 79-90.

Creswell, J.W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed.). Thousand Oaks, CA: SAGE Publications.

Engelbrecht, J., & Harding, A. (2003). E-assessment in mathematics: Multiple assessment formats. *New Zealand Journal of Mathematics, 32,* 57–66.

Farthing, D.W., Jones, D.M., & McPhee, D. (1998.) Permutational multiple choice questions: An objective and efficient alternative to essay-type examination questions. *ACM SIGCSE Bulletin, 30*(3), 81-85.

Lewis, D. J., & Sewell, R. D. (2007). Providing formative feedback from a summative computer-aided assessment. *American Journal of Pharmaceutical Education, 71*(2), 33.

Luckett, K., & Sutherland, L. (2000). *Assessment practices that improve teaching and learning: Improving Teaching and Learning in Higher Education,* 98-130. Witwatersrand University Press.

Malau-Aduli, B. S., Assenheimer, D., Choi-Lundberg, D., & Zimitat, C. (2013). Using computer- based technology to improve feedback to staff and students on MCQ assessments. *Innovations in Education and Teaching International,* 1-13.

Maurice, S.A., & Day, R.L. (2004). Online testing technology: Important lessons learned. *International Journal of Engaging Education, 20*(2), 152-160.

McCann, A.L. (2009). Factors affecting the adoption of an e-assessment system. *Assessment and Evaluation in Higher Education, 35*(7), 799-818.

Mendes, A.C., Curto, S., & Coheur, L. (2011). Bootstrapping multiple-choice tests with the-mentor. *Computational Linguistics and Intelligent Text Processing*, 451-462. Heidelberg: Springer.

Miller, P.A. (2012). Use of computer-based assessment strategies. *Unpublished paper at E-Learning Update Conference*. Johannesburg. Available at https://app.box.com/s/bcb47906350a31d1ab8c

Mitkov, R., & Ha, L.A. (2003). Computer-aided generation of multiple-choice tests. *Proceedings of the HLT-NAACL 03 workshop on building educational applications using natural language processing, 2,* 17-22. Stroudsbury PA: Association for Computational Linguistics.

Mostert, E., de Bruyn, E., & Pretorius. G. (2015). What should the perfect online assessment system look like? *Proceedings of International Association for Management of Technology (IAMOT) 2015 Conference*. Available at http://www.iamot2015.com/2015proceedings/documents/P241.pdf

Oates, B.J. (2010). *Researching information systems and computing* (5th ed.) London, UK: Sage Publications.

Pretorius, G.J., Mostert, E., & de Bruyn, E. (2007). Local innovation: Development of a computer-based testing system. *9th Annual Conference on World Wide Web Applications*. Johannesburg. South Africa.

Scalter, N., & Howie, K. (2003). User requirements of the ultimate online assessment engine. *Computers & Education, 40*(3), 285-306.

Singh, U.G., & De Villiers, M.R. (2012). Investigating the use of different kinds of multiple-choice questions in electronic assessment. *Progressio: South African Journal for Open and Distance Learning Practice, 34*(3): 125-143.

Singh, U.G., & de Villiers, M.R. (2015). E-SEAT: An electronic framework for evaluating e-assessment systems. *Proceedings of E-Learn 2015 – World Conference on E-Learning*. Kona, USA: AACE.

Stodberg, U. (2012). A research review of e-assessment. *Assessment and Evaluation in Higher Education, 37*(5), 591-604.

Thomas, C., Borg, M., & McNeil, J. (2014). E-assessment: Institutional development strategies and the assessment life cycle. *British Journal of Educational Technology, 46(3).*

Valenti, S., Cucchiarelli, A., & Panti, M. (2002). Computer-based assessment systems evaluation via the ISO90126 quality model. *Journal of Information Technology Education, 1*(3), 157–175.

Ventouras, E., Triantis, D., Tsiakas, P., & Stergiopoulos, C. (2010). Comparison of examination methods based on multiple choice questions and constructed-response questions using personal computers. *Computers & Education, 54*(2), 455-461.

Wills, G.B., Bailey, C.P., Davis, H.C., Gilbert, L., Howard, Y., Jeyes, S., ... & Young, R. (2009). An e-learning framework for assessment (FREMA). *Assessment and Evaluation in Higher Education, 34*(3), 273-292.

Wood, E.J. (2003). What are extended matching sets questions? *Bioscience Education eJournal, 1*(1), *1-2*.