

Designing onscreen simulated ICT performance tasks

Gemma Hinchliffe, Helen Harth &
Andrew Stone

The City and Guilds of London Institute,
UK

Abstract

Development and implementation procedures of innovative onscreen tests made up of simulated performance tasks require considerable assessment and technical expertise and have significant time and cost implications. In order to identify challenges for their development, delivery and scoring, we provide an account of the assessment development lifecycle of Functional Skills Information and Communication Technology tests, followed by recommendations. This paper is intended to help assessment practitioners and test sponsors make informed planning and implementation decisions.

Introduction

England has a complex assessment system; with many types of qualifications (Isaacs, 2010, p. 319). One particular area of policy initiatives – including qualifications development – has been that of ‘generic skills’; various initiatives have been put in place over the past 30 to 40 years, ranging for ‘core skills’ to ‘key skills’ (Haywood & Fernandes, 2004) or even – in the European context – to key competencies (Oates, 2003). The most recent suite of English qualifications in this arena is Functional Skills.

Technology offers many opportunities for innovation in assessment development, especially in the vocational education sector. The introduction of Functional Skills to the generic skills arena prompted a large-scale review of what could be achieved using online assessment, and also prompted, in some cases, major investment from awarding organisations looking to gain a competitive advantage in the marketplace. While there have been innovations during this process which will certainly shape

Designing ICT onscreen simulations

assessment development in the future, the difficulties encountered during the process are as interesting and perhaps more worthwhile areas of focus.

Functional Skills was piloted across England from 2006 to 2009, and consists of English, Mathematics and Information and Communication Technology qualifications at Qualifications & Credit Framework (QCF) Entry Levels 1, 2 and 3, and Levels 1 and 2. It was intended to replace Key Skills, and form a compulsory component of the Diploma, Apprenticeship frameworks¹ and Foundation Learning. The qualification was launched in September 2009.

Functional Skills qualifications are intended to test both competence and problem-solving ability, with the intention being that candidates should be assessed on their ability to complete tasks using transferable skills that are relevant to their 'everyday' life, and usable in the workplace. The assessment criteria therefore include both competence-based standards and also some wider, more nebulous criteria, which may be based in self-evaluation, planning, or problem-solving approaches. These are referred to collectively as 'skills standards'. The job of writing assessments to fit these standards is made much more difficult when embarked upon at the same time as the development of new online assessment models (see: Boyle, May & Sceeny, in press).

This paper is mainly concerned with the assessment of Functional Skills Information and Communication Technology (FS ICT). While the assessment of other Functional Skills subjects may lend itself to an amplification and development of existing item types (English Reading, for example, being assessed through carefully worded short-answer questions, with responses captured and marked through an online portal), the assessment of FS ICT at Level 1 and 2 was considered to be more suited to assessment via onscreen simulated performance tasks - capturing responses via a 'simulated desktop environment.' In this model, complex performance tasks are delivered via a simulated desktop, and automatically marked. The system is able to gather meaningful measurement evidence, previously not available in the case of externally set and locally marked assessment. This opens the door to powerful scoring, reporting and instant-feedback mechanisms, but also presents a vast array of complex problems with the design of interdependent multi-layer tasks, scoring and reporting (Clauser, Margolis, Clyman & Ross, 1997; Mislevy et al, 2006).

The development process for the FS ICT onscreen simulated performance tasks provides an excellent case study of the challenges awarding organisations face when developing new online assessment models. The high stakes nature of Functional Skills has inevitably increased commercial pressure to deliver innovative, cost-effective assessment on a wide scale.

This paper provides a transferable model of the process used to develop onscreen simulated performance tasks, using the development of FS ICT as a case study. We will consider the applicability of common test development frameworks to this case study and suggest adaptations that may benefit future developers. We will also suggest areas of further research for practitioners.

Common test development frameworks

Test development frameworks provide a starting point for analysis. There are a number of commonly used and recognised test development frameworks, proposing quality control procedures for the development, scoring and reporting of valid and

¹ Key skills qualifications use in apprenticeship frameworks has been extended until September 2012 (City & Guilds 2011; DfE 2011, p. 9).

reliable assessment. These are especially useful in complex production processes that consist of several stages and require the involvement of diverse areas of expertise (eg Downing, 2006; Allalouf, 2007; Mislevy, Steinberg & Almond, 2002; Pucel, 2005, 2008; see Appendix 1, Table 1 for a summary of each of these frameworks).

Pucel (2005) describes eight steps that aim to capture additional or unique procedures for developing computer-simulated performance tests that involve technical as well as test development procedures. Equally, Downing's comprehensive approach could in principle be adapted in the case of tests delivered onscreen (Downing, 2006).

The eleven quality control steps proposed by Allalouf (2007) relate to scoring, equating and reporting processes. Allalouf explains why mistakes happen and how to avoid, detect and deal with them. This is a particularly useful approach for assessment practitioners, especially when combined with other holistic frameworks.

A fourth and most complex approach is the evidence-centred assessment design (ECD) framework for designing, producing, and delivering a broad range of assessment types. It is argued that other step-by-step approaches such as the ones discussed above are of limited effectiveness for innovative assessments. In summary, ECD consists of an interdependent series of assessment design steps or layers, each helping to structure and guide the overarching assessment argument (Mislevy, Steinberg & Almond, 2002; Mislevy et al, 2006; Mislevy & Riconscente, 2006).

In this context, we propose that the ECD framework is the most appropriate framework upon which to base a development of this complexity. However it is worth noting at this stage that while the ECD framework provides a good basis for complex test development, it could be argued that it is more suited to the development of multiple-choice item types – hence its popularity as a model in the USA. The UK model of test development tends to rely upon a mixture of expert judgement and some data analysis, and is largely based on custom and practice. This juxtaposition is all the more interesting when discussing the development of Functional Skills assessments, which because of the substantial subjective element lend themselves to development with a dependence on Subject Matter Experts (SMEs). The purpose of this review is also in part, therefore, to reconcile this approach with a coherent theoretical process.

A unified approach

The ECD framework has four distinct development stages. Here, we will address each stage of the framework, recognising the differences and challenges in using this framework for the development of onscreen simulated ICT tasks.

Figure 1 summarises the four layers of the ECD framework in its initial state in graphical format: domain analysis, value proposition/product requirements, conceptual assessment framework/task creation and those activities relating to the implementation and delivery of the test.

Designing ICT onscreen simulations

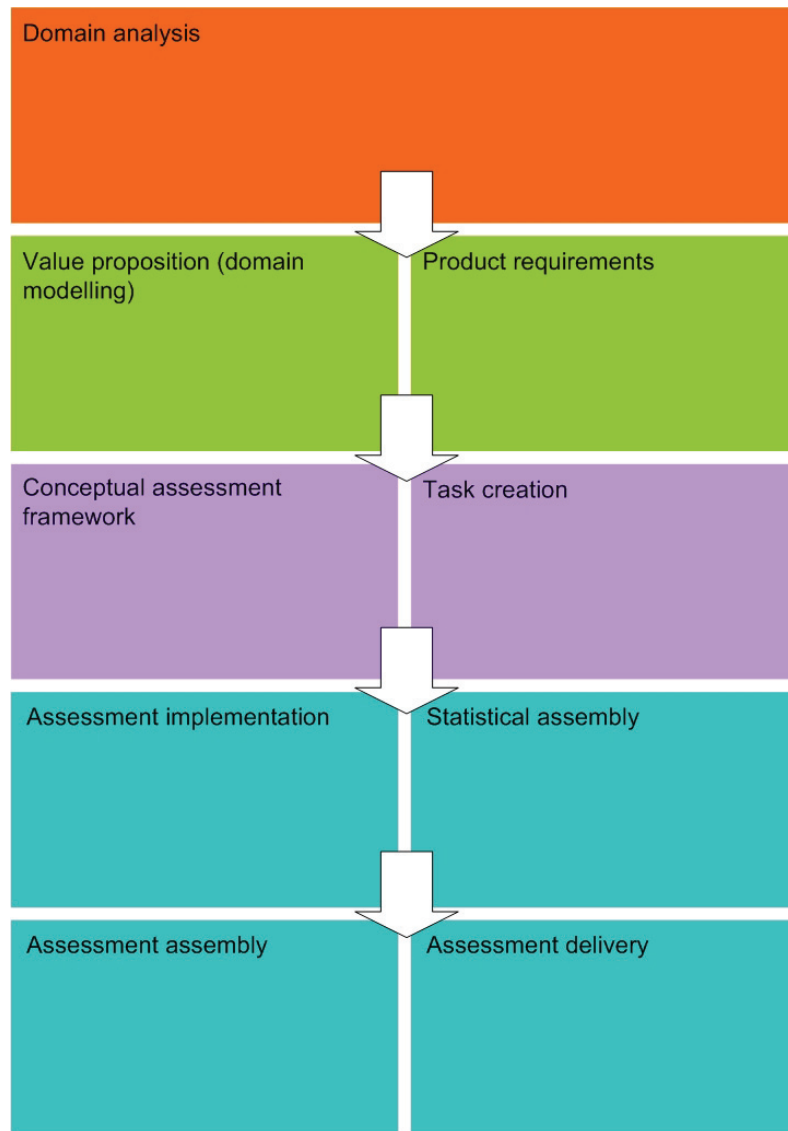


Figure 1. Graphical interpretation of the ECD framework

Domain analysis

The domain analysis stage of ECD is concerned with defining essential features of the qualification, purpose(s) of assessment, and how knowledge and skills are acquired and developed. Information about the candidate population, what they should be expected to be able to do in the domain and assessment purposes are then gathered. Developers then observe representative learners carrying out activities in the target domain to inform a conceptualisation of the kinds of behaviours expected (Mislevy & Riconscente, 2006).

As discussed above, the skill standards for Functional Skills provide a framework for their assessment and describe what learners should be able to do in 'real life' contexts. For FS ICT, candidates are required to use problem solving and evaluation skills in order to find, select, develop, present and communicate relevant or given information. This indicates the complexity and relatively higher level of Functional Skills assessment in comparison with the qualification it is intended to replace, Key

Skills, the assessments for which have over time become increasingly structured, and arguably easier to 'teach to' (Wolf, 2011).

Value proposition (domain modelling)

This second layer of the ECD model is concerned with determining key assessment design features, based on information gathered during the domain analysis layer (Mislevy & Riconscente, 2006). The prioritised aspects for the development of FS ICT at this stage were to define:

- Key assessment design features
- Design pattern
- Potential work products.

The rationale for virtual-desktop simulations

The skill standards for FS ICT indicate that candidates must be able to evidence ICT competence in unfamiliar, generic contexts. This requirement lends itself to assessment through a generic simulated platform as opposed to branded, and sometimes very different, platforms. It also favourably counteracts the problem of the advancement of these branded platforms, where the unchanging generic simulated platform provides a stable and robust test environment, as opposed to advancing branded platforms which may demand different actions from candidates completing the same task on different versions.

Also noted at this stage was that one skill standard requires candidates to use (necessarily internet-based) search engines (see Appendix 2). This poses security issues for externally-set and marked assessments not previously faced by vocational qualifications, and the secure 'simulated desktop' environment provides a neat solution to this issue, where simulated websites can be included as part of the simulated test environment.

In terms of content, the task-based approach used in the simulated desktop model has been proven to support the measurement of skills such as problem-solving and evaluating one's own work (Mislevy et al, 1999; Steinberg & Gitomer, 1996; Williamson et al, 2004).

Moreover, the virtual-desktop simulations have the potential to decrease the logistical burden of delivery associated with the traditional model of 'paper-based' FS ICT assessment that requires printing the evidence and sending it to be marked by a human examiner, within a reasonable timescale (Stone & Dearing, 2009; Pucel & Anderson, 2006; Pucel, 2005). The Functional Skills qualifications are one of the largest scale on-demand, externally assessed testing programmes introduced in the English vocational qualifications sector; for the paper-based model of FS ICT, assessment is offered monthly, and on average approximately 10,000 candidates per month will be entered for the test. The implications of this to the business are wide-ranging, and create a huge logistical and operational burden.

While the above arguments clearly vindicate the simulated desktop approach, there are arguments against it too. Functional Skills requires candidates to evidence independent thinking, planning and evaluation; each of these poses challenges for automated marking and scoring. The criteria for FS ICT also specify that each assessment must include a minimum of 80 per cent open response assessment. This requirement, a key characteristic of Functional Skills, does not lend itself to automated marking and scoring, as 'correct' combinations of actions or responses

could potentially number into the thousands if the items are truly open-response and without 'scaffolding', as per the assessment strategy stipulated by the regulator.

These challenges were identified early in the process, and as such, the earliest assessment strategy drafts included test developers working in close collaboration with SMEs and IT developers to define the types of tasks, scenarios and experiences required for the test (as in Mislevy & Risconscente, 2006). This was key to the success of the development. However, while it enabled efficient communication between parties, it did not entirely solve the issue of how to accurately mark open-ended assessment types automatically.

Simulated task requirements

The development of complex, open-ended tasks highlights problems with attempting to address multiple domains within a task/test. Here, each SME was asked to comment on links or overlaps between the domains in the context of some proposed exemplar simulation tasks, specify linkage judgements of each skills domain to the corresponding task(s) and on how performance against one of the skills standards may be influenced by performance on another (Raymond & Neustel, 2006).

The simulations that were suggested at this stage are authentic replicas of everyday office software, including a word processor, spreadsheet, presentation graphics program, internet browser, database, email program and file manager. There is no obligation to use a specific application for a task. For example, a task requiring the production of a leaflet can be completed just as effectively with presentation graphics software as with a word processor, using information gleaned from different sources within the simulation.

Next, the requirements were captured in a functional specification; this included information on the sufficient functionality of the program in order to infer competence, task types, appropriate real-life scenarios (in order to inform background work such as simulated website builds), scoring rules, vendor neutral functionality, varying weightings for each task and interconnected tasks. The functional specification was then passed to the technology supplier's development team, and analysed internally so that the most appropriate development strategy could be selected.

Conceptual assessment framework

In this stage of the ECD model, important design decisions and processes are stipulated based on information gathered in the previous two layers.

Test specification

The next step in the process was to interpret and implement the content specification based on the qualification's skill standards in the form of a test specification. This is used to communicate the structure and content of the test, which is critical for ensuring the validity of scores and comparability across test forms. In this case, it includes the weighting of each content area, the skill standards associated with each of them, the coverage and range within each standard, and the item types and format to be used (eg tasks, constructed response or selected response items).

For the FS ICT tests, one task is made up of a varying number of items. Weights of each skill standard within a task were decided relying exclusively on subject matter expertise (as described in Raymond & Neustel, 2006). For this type of test, it is possible to design a content-based test specification that contains each of the skill standards (Raymond & Neustel, 2006; see Appendix 2). A second step to support scenario-based task development was to determine the process matrix required to

execute these tasks which contains the types of behaviours or process required in the practice setting (Appendix 3). The resulting computer-based simulations are formed of a number of tasks, each task containing a set of items, normally up to 40 or 50 in the test depending on level, each scored dichotomously (ie either 1 for a correct response or 0 otherwise).

Task development

Using the skill standards and assessment requirements, the next stage was to develop tasks that require the candidate to understand a real-life scenario. Criteria that are difficult to assess through instruction and action (for example, 'demonstrate understanding of the need to stay safe and to respect others when using ICT-based communication') may be tested using a combination of item types, including multiple choice items. The process for task development is described in Figure 2. Due to the scenario-based approach, test assembly does not apply in this instance since each test is made up of tasks, which are made up of dependent items within and between tasks.

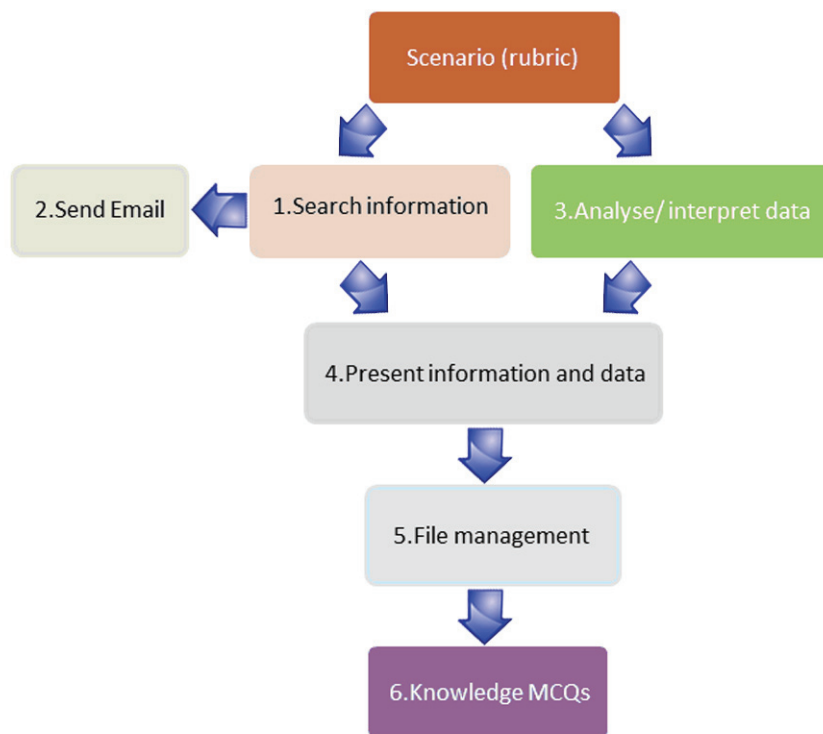


Figure 2. An example of an assessment scenario that follows the process-based specification.

Automated scoring

The credibility and usefulness of an assessment programme is directly linked to how results are quantified (scored) and reported (Aschbacher, 1991). The next issue therefore was to consider what the system records as a 'correct' response in a given task. The question at this stage was whether the 'working out' or process should be scored polytomously; or only the final product or artefact should result in a score (Downing, 2006; Thissen & Wainer, 2001; van der Linden & Hambleton, 1997).

Within the limitations of the simulation, it was feasible to record candidate actions and score them according to rule-based methods based on expert judgement, using 'if-then' logical analysis. It was judged that in order to capture the most representative cross-section of competence data, this should then be combined with scoring based

on the final product. A list of acceptable candidate actions and subsequent end-products within the simulation was created at this point, again using 'if-then' logical analysis. This was subsequently refined with real candidate answers (Mislevy et al, 2006). Once the list of acceptable correct answers was finalised, SMEs arrived at an aggregated final score. Cut-off points are then applied to classify candidates as pass or fail.

This combination of scoring types meant that the marking schemes for each task became extremely complex in order to capture the increased variation in candidate responses. This has implications for the agility of the software, and for the assessment itself.

Measurement model

As the description of the FS ICT simulation shows, assumptions of independence and dimensionality may not hold for such test items. A number of factors, such as topical knowledge, vocabulary knowledge, item locations, cognitive skills, or the presence of testlet effects may cause violations of local independence (Jang & Roussos, 2007). Item dependence may also be attributed to a number of factors, including multistage performance tasks and context dependent item sequences (Ferrara, Huynh & Baghi, 1997; Ferrara, Huynh & Michaels, 1999).

Models such as true test score theory and item response theory that assume conditional independence will overestimate the precision of the measurement, which may lead to inaccurate inferences based on the test scores (Wainer & Thissen, 1996; Wainer & Wang, 2000). In addition, local dependence of items would lead to a bias in item difficulty and discrimination parameter estimates and finally an overestimation of test score reliability (Ackerman & Spray, 1986, 1987). Based on the description of these tests, local item dependence, dimensionality and mix of item types must then be considered in the development and scoring of high stakes tests (Zenisky, Hambleton & Sireci, 2002; Sireci, Thissen & Wainer, 1991; Wainer & Thissen, 1996; Lee & Frisbie, 1999).

Data analysis

Complex assessment data may pose additional challenges for data analysis (Mislevy et al, 1999). A number of measurement models could be used to determine the reliability of scores, including classical test theory, item response theory and testlet response theory. Considering the nature of this data (assumed to be multidimensional and made up of dependent items within/across tasks), testlet response theory using a software package such as Scoright may provide additional information and allow for improved candidate feedback at section level (Wang, Bradlow & Wainer, 2005). Item statistics appropriate for polytomous items could also be used, as are passing rates and retest rates (passing rates given multiple opportunities).

Comparing simulation pass rates with those of the traditional mode of assessment may also help us to judge the accuracy of results from the simulation. In this case it may be possible that automated scoring is more precise than observations by a human assessor. Our hypothesis would be in this case that fewer people fail the onscreen rather than the computer-completed test due to increased objectivity in scoring and improved feedback mechanisms.

Other considerations are about whether the tasks corresponding to each skill standard provide stable scores for each of the content areas. Furthermore, the possibility of taking this test as soon as the candidates are considered ready for it should be considered in item and test analysis as well as in maintaining standards over time (He, 2010).

Implementation and delivery

This final stage of the ECD framework could be seen as extraneous to the test development itself, but in the development of a qualification such as Functional Skills, with its high volumes and the pressure to cut results delivery timescales, it is crucial that it is managed by people with an understanding of the complexity of the development already completed in the run up to this stage. A major requirement for such a development project is that the system is able to cope with the demand for test administration and delivery of results.

Resulting simulation

At the beginning of the test, the virtual desktop opens up to present the software applications which the candidate has at their disposal. The right hand side of the screen contains the rubric that guides the candidate through the requirements of each task (see Figure 3). Before taking the test, candidates view a 'navigation' screen, which allows them to familiarise themselves with the simulation environment before commencing the test. Once tasks are scored, provisional summary scores, section scores and classification decisions are immediately reported to candidates, which are then confirmed following data analysis and post-administration standard setting.

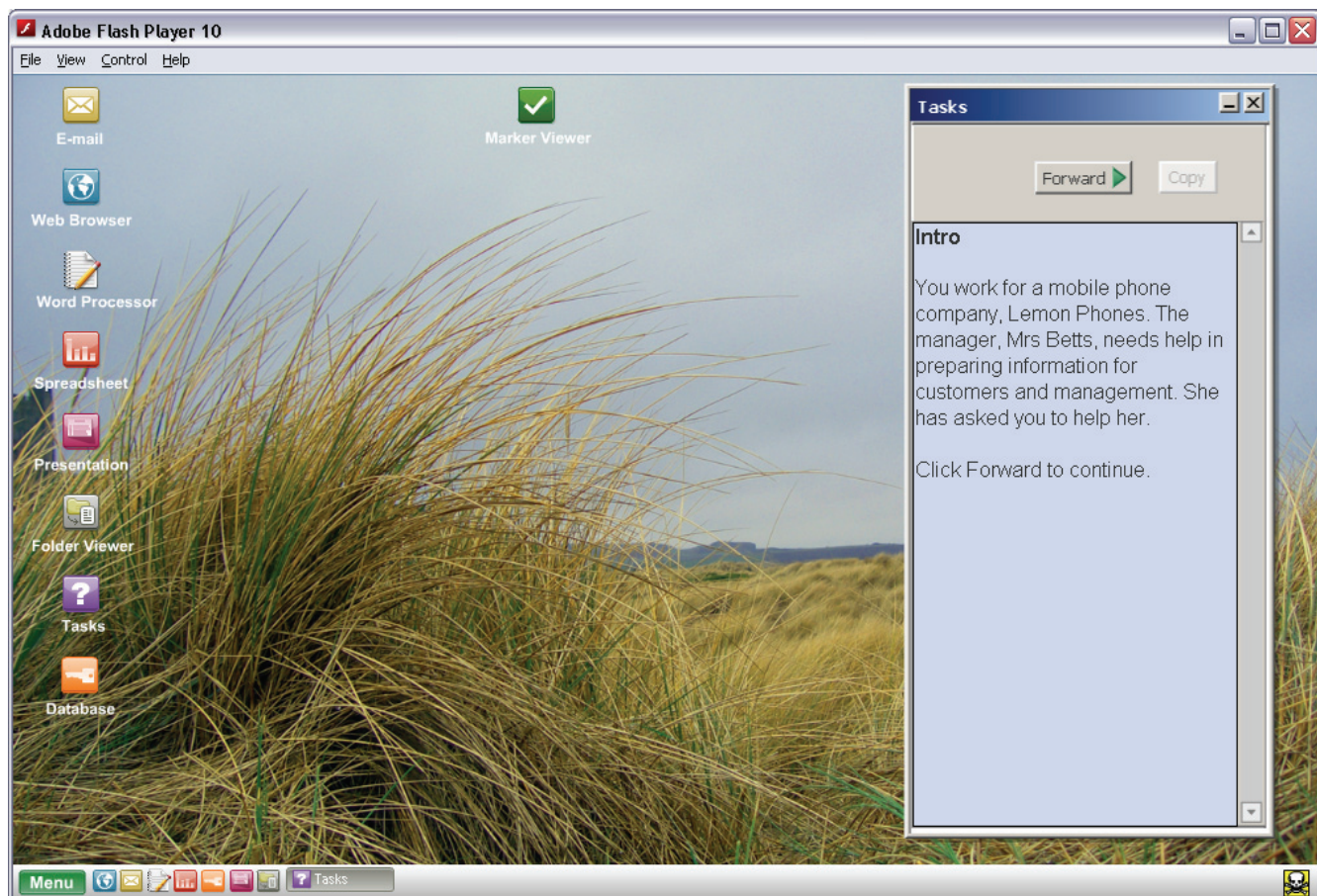


Figure 3. An example of a FS ICT simulation. Task instructions sit in a floating pane, seen here on the right hand side of the screen.

Summary

Figure 4 (see following page) summarises the test development process followed during the development of the FS ICT simulations, reclassified within the ECD structure. This represents the amendments and extra elements of the development as discussed thus far.

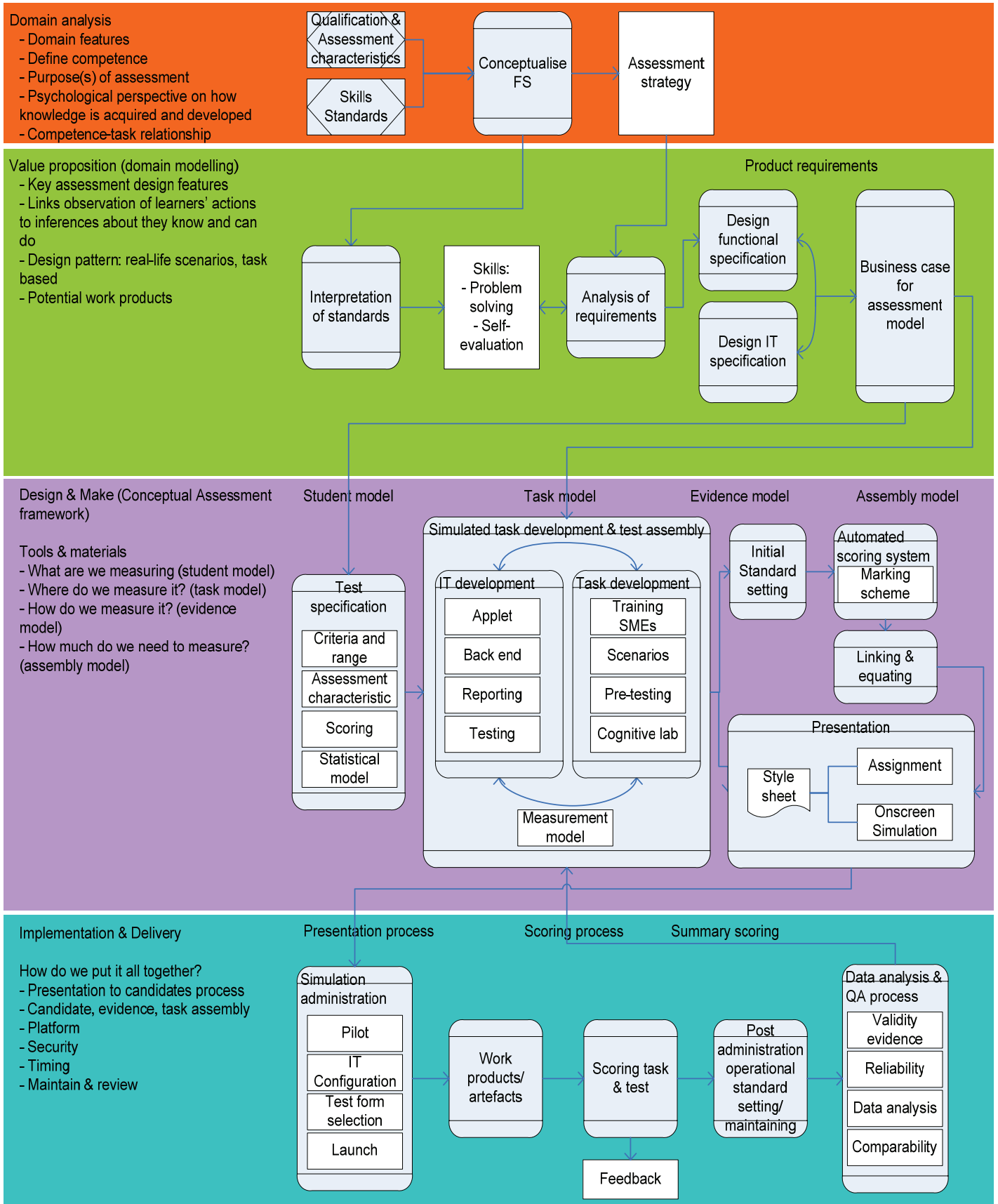


Figure 4. An amended version of the ECD framework to reflect the FS ICT simulation development process.

Recommendations

Developing, delivering and scoring onscreen tasks of the complexity described here is a challenge for awarding organisations and other test providers. A first challenge is one of developing sufficient expertise in innovative assessment methods and psychometrics for informed decisions during the initial phases of the development. SMEs are challenged to understand new ways of working as well as additional requirements of the technology design and implementation. Such a system can provide opportunities for development as well as initial constraints that should be captured at the design modelling layer/phase. This highlights yet another issue with the availability of technical expertise able to meet the assessment demands within time and budget. ECD highlights the need for common terminology in the development process and the role each team has in building up the validity argument – what can be feasibly inferred based on the test results. What follows are three recommendations for organisations developing similar assessment instruments.

1. Scoring

The scoring model used should be considered during the test specification development phase. Item dependency should be considered when selecting an appropriate measurement model. A number of methods should be used with the data set for scoring and reliability studies, including classical test theory, item response theory and testlet response theory. For multi-dimensional tasks, multiple scores should be derived, so that individuals' differential scoring on various dimensions can be understood by test developers and (perhaps) reported to test users.

2. Practice 'space'

The availability and standard of practice material is crucial when offering test material that significantly differs from a candidate's normal way of working. In this context, we suggest that practice material is made readily available, to demonstrate the capability of the software, but with limited assessment content, therefore operating as practice 'space' as opposed to 'material'. This enables learners to become familiar with the test environment, without become overly familiarised with the content.

3. The tension between technology and assessment content

In the initial stages of assessment strategising, communication between the SMEs and the technology developers is crucial. From the very beginning of the development project we found close partnership working between SMEs and technologists to be essential. The technology developers must be flexible, communicative and have sympathy for the SMEs' curricular aims; the SMEs must understand the technology's limitations and be imaginative enough to apply their in-depth subject knowledge and assessment development expertise to the available functionality (Boyle & Hutchinson, 2008). The balance between writing valid tasks that may not work with the technology, and writing less valid tasks to fit the limitations of the software, must be met. The two are interrelated to such an extent that should one party advance ahead of the other in the development, retracing steps and making amendments becomes almost impossible. Therefore, in order to retain aspirations for fully valid assessment in a practical and demanding setting, this partnership working is essential.

Future research

This project has shown that it is possible to apply an established framework like ECD to a complex development such as that outlined above. Further research is required for the development of psychometric measurements that are able to describe this complex data set and confirm the validity of the inferences originating from a virtual environment.

Alternative models to classical test theory may be beneficial in the analysis of items, tasks, test form assembly and evaluation of the FS ICT simulations, such as IRT-based scoring and test development methods (Wainer, Bradlow & Wang, 2007). Simulated and live data studies comparing IRT methods provide us with highly improved scoring mechanisms and help us to understand the impact of local item independence and dimensionality on measures obtained using classical test theory (Roussos, Stout & Marden, 1998). Item characteristics such as type, difficulty, order or content specification should be investigated further since they may affect the measurement properties of these ICT tests and result in an overestimation of test score reliability (Ackerman & Spray, 1986, 1987).

Another area of further research includes the degree of specificity that is necessary for such complex tasks and the number of performances that should be sampled in order to obtain valid and reliable results (Aschbacher, 1991). Furthermore, the simulation fidelity or realism should be investigated, especially given the requirement for vendor-neutral functionality.

All of these issues highlight the importance of future validation work, taking into account characteristics of onscreen simulated performance assessments (Clauser, Kane & Swanson, 2002). Ultimately innovations in assessment and technology can only achieve their aims when there is evidence for increased construct representation and improved measurement efficiency (Sireci & Zenisky, 2006).

References

- Ackerman T.A. & Spray J. A. (1986). A general model for item dependency. Paper presented at the 1986 AERA annual meeting, San Francisco.
- Ackerman T.A. & Spray J. A. (1987). The effect of item response dependency on trait or ability dimensionality. Iowa City, IA: ACT.
- Allalouf, A. (2007). Quality Control Procedures in the Scoring, Equating, and Reporting of Test Scores. *Educational Measurement: Issues and Practice*, 26(1): 36–46.
- Aschbacher, P. A. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education*, 4(4), 275-288.
- Boyle, Andrew and Hutchison, Dougal (2008) 'Sophisticated tasks in e-assessment: what are they and what are their benefits?' *Assessment & Evaluation in Higher Education*, 34(3), 1-14.
- Boyle, A., May, T. & Sceeny, P. (in press) A history of three e-assessment programmes in England.
- City & Guilds (2011) *Functional Skills update*. Available online at: <http://www.cityandguilds.com/63273.html>
- Clauser, B. E., Kane, M. T., Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4), 413-432.
- Clauser, B.E., Margolis, M.J., Clyman, S.G. & Ross, L.P. (1997). Development of Automated Scoring Algorithms for Complex Performance Assessments: A Comparison of Two Approaches. *Journal of Educational Measurement*, 34(2), 141-161.
- Department for Education (DfE) (2011) *Wolf review of vocational education: government response*. Available online at: <http://media.education.gov.uk/assets/files/pdf/w/wolf%20review%20of%20vocational%20education%20%20%20government%20response.pdf>
- Downing, S.M. (2006). Twelve steps for effective test development. In: S.M. Downing & T.M. Haladyna (Eds), *Handbook of Test Development*. LEA Pub, NJ.
- Ferrara, S., Huynh, H. and Michaels, H. (1999), Contextual Explanations of Local Dependence in Item Clusters in a Large Scale Hands-On Science Performance Assessment. *Journal of Educational Measurement*, 36(2), 119-140.
- Ferrara, S., Huynh, H., Baghi, H. (1997). Contextual Characteristics of Locally Dependent Open-Ended Item Clusters in a Large-Scale Performance Assessment. *Applied Measurement in Education*, 10(2), 123-144.
- He, Q. (2010). Maintaining standards in on-demand testing using item response theory. *Ofqual*/10/4724.
- Heyward, Geoff, and Rosa M. Fernandes. 2004. From core skills to key skills: fast forward or back to the future? *Oxford Review of Education* 30(1): 117-145.
- Isaacs, Tina. 2010. Educational assessment in England. *Assessment in Education: Principles, Policy & Practice* 17(3): 315-334.
- Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance based nonparametric approach. *Journal of Educational Measurement*, 44, 1-21.

- Lee, G. & Frisbie, D.A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, 12, 237-255.
- Mislevy, R.J. & Riconscente M.M (2006). Evidence-centred assessment design. In: S.M. Downing & T.M. Haladyna (Eds), *Handbook of Test Development*. LEA Pub, NJ.
- Mislevy, R.J., Steinberg, L.S., Almond, R.G. & Lukas J.F. (2006). Concepts, terminology, and basic models of Evidence-Centred Design. In: D.M. Williamson, R.J. Mislevy & I.I. Bejar (Eds), *Automated Scoring of Complex Tasks in Computer Based Testing*. LEA Pub, NJ.
- Mislevy, R.J., Steinberg, L.S., Breyer F.J., Almond, R., Johnson, L. (1999) A cognitive task analysis with implications for designing a simulation-based assessment system. *Computers and Human Behaviour*, 15, 335-374.
- Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2002). Design and analysis in task-based language assessment. *Language Assessment*, 19, 477-496.
- Oates, T. (2003) Key skills/key competencies: avoiding the pitfalls of current initiatives. In Rychen, D.S., Salganik, L.H. & McLaughlin, M.E. (Eds.) (2003) *Definition and selection of key competencies: contributions to the second DeSeCo symposium*, Geneva, Switzerland, 11-13 February, 2002. (Neuchâtel: Swiss Federal Statistical Office). Available online at: <http://www.oecd.org/dataoecd/48/20/41529505.pdf>
- Pucel, D. J. & Anderson, L.D. (2006). Computer simulation performance testing. Proceedings of the Ninth IASTED International Conference on Computers in Advanced Technology in Education 2006 (CATE), Lima, Peru.
- Pucel, D.J. (2005). *Developing and Evaluating Performance-Based Instruction* (third edition). New Brighton, MN: Performance Training Systems, Inc.
- Pucel, D.J. (2008). *Developing and Implementing Computer Simulated Performance Tests*. 34th International Association for Educational Assessment (IAEA) Annual Conference 2008, Cambridge University, UK.
- Raymond, M. & Neustel, S. (2006). Determining the content of credentialing examinations. In: S.M. Downing & T.M. Haladyna (Eds), *Handbook of Test Development*. LEA Pub, NJ.
- Roussos, L. A., Sout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*. 35(1), 1-30.
- Sireci S.G. & Zenisky A.L.. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In: S.M. Downing & T.M. Haladyna (Eds), *Handbook of Test Development*. LEA Pub, NJ.
- Sireci, S. G., Thissen, D. & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, 28(3), 237-247.
- Steinberg, L.S. & Gitomer, D.G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science*, 24, 223-258.
- Stone, A. & Dearing, M. (2009) Simulations: a case study of City & Guilds' newest assessment, *Education & Training*, 51(5/6), 422-433.
- Thissen, D. & Wainer, H. (2001). Overview of Test Scoring. In: D. Thissen & H. Wainer (Eds), *Test Scoring*, pp. 1-19. Hillsdale, NJ: Lawrence Erlbaum Associates.

- van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer.
- Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15, 22-29.
- Wainer, H. & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37, 203-220.
- Wainer, H., Bradlow, E.T. & Wang X. (2007). *Testlet Response Theory and its Applications*. New York: Cambridge University Press.
- Wang X., Bradlow E.T. & Wainer H. (2005). User's Guide for SCORIGHT (Version 3.0): A Computer Program for Scoring Tests Built of Testlets Including a Module for Covariate Analysis. Research Report, ETS.
- Williamson, D.M., Bauer, M., Steinberg, L.S., Mitlevy, R.J. & Behrens, J.T. (2004). Design rationale for a complex performance assessment. *The International Journal of Testing*, 4, 303-332.
- Wolf, A. (2011) *Review of Vocational Education - The Wolf Report*. London: Department for Education.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Effects of local item dependencies on the validity of item, test, and ability statistics. *Journal of Educational Measurement*, 39 (4), 1-16.

Appendix 1

Table 1. Three frameworks for test development

Steps in the test development cycle			
Issues and challenges in the development of computer simulated tests (Pucel; 2008)	Twelve Steps for Effective Test Development (Downing, 2006)	Eleven quality control steps for scoring, equating and reporting (Allalouf, 2007)	ECD framework components (Mislevy, Steinberg & Almond, 2002)
<ol style="list-style-type: none"> 1. Identifying the skills to be tested 2. Determining the reasonableness of testing through computer simulation 3. Developing the performance test instrument 4. Establishing computer simulation design parameters 5. Developing the multi-media presentation 6. Preparing test-takers 7. Validation 8. Data analysis and review (validity and reliability evidence) 	<ol style="list-style-type: none"> 1. Overall plan 2. Content definition 3. Test specification 4. Item development 5. Test design and assembly 6. Test production 7. Test administration 8. Scoring test responses 9. Passing scores 10. Reporting test results 11. Item banking 12. Test technical report 	<ol style="list-style-type: none"> 1. Knowing your examinees in advance 2. Obtaining the examinees' answers 3. Storing examinee data in a database 4. Scoring (temporary raw scores) 5. Item analysis 6. Computing final raw scores 7. Equating new test forms and items 8. Computing standardised scores 9. Test security checks 10. Reporting test scores 11. Documentation of the scoring process 	<ol style="list-style-type: none"> 1. Domain analysis – define the field of study/practice 2. Domain model – define relevant competence 3. Conceptual Assessment framework (CAF) <ul style="list-style-type: none"> • Student model – competence, characteristics of test • Evidence models – behaviours that indicate levels of proficiency, statistical model • Task models – tasks/activities, rubrics, scoring rules 4. Implementation & delivery system model – test administration, delivery of work product, scoring, summary scoring and feedback <ul style="list-style-type: none"> • Assembly models • Presentation models • Evidence identification • Evidence accumulation

Appendix 2 (following pages)

Exemplar test specification template for FS ICT Level 1.

Level 1 Functional Skills ICT – Paper-based

Name of paper



Total marks available: 40

Marks allocated to fixed response: **X (XX%)** (Maximum fixed = 20%) Marks allocated to open response: **XX (XX%)** (Minimum open = 80%)

		Skill Standard	Coverage and Range	Marks	Covered (task no)	
Using ICT	Weighting: 20 - 30%	1	Identify the ICT requirements of a straightforward task	Use ICT to plan and organise work		
		2	Interact with and use ICT systems to meet requirements of a straightforward task in a familiar context	A	Select and use software applications to meet needs and solve straightforward problems	
				B	Select and use interface features effectively to meet needs	
		3	Manage information storage	C	Adjust system settings as appropriate to individual needs	
	4	Follow and demonstrate understanding of the need for safety and security practices	Work with files, folders and other media to access, organise, store, label and retrieve information			
	Finding & Selecting	Weighting: 10 - 20%	5	Use search techniques to locate and select relevant information	Search engines and queries	
6			Select information from a variety of ICT sources for a straightforward task	A	Recognise currency, relevance and bias when selecting and using information	
		B		Recognise copyright when selecting and using information		
Weighting: 10 - 20%		XX%				XX
						XX

Developing presenting & communicating		Weighting: 50 - 70%		Apply editing, formatting and layout techniques to meet needs, including:		
7	Enter, develop and refine information using appropriate software to meet requirements of straightforward tasks	A1	Text			
		A2	Tables			
		A3	Graphics			
		A4	Records			
		A5	Numbers			
		A6	Charts and graphs			
		A7	Other digital content			
	8	Use appropriate software to meet requirements of straightforward data-handling task	A	Process numerical data		
			B	Display numerical data in a graphical format		
			C1	Use field names to organise information		
			C2	Use data types to organise information		
9	Use communications software to meet requirements of a straightforward task	D1	Enter, search, sort and edit records			
		A1	Receive and read electronic message with attachments			
		A2	Send electronic message with attachments			
		B1	Demonstrate understanding of the need to stay safe when using ICT-based communication			
10	Combine information within a publication for a familiar audience and purpose	B2	Understand the need to respect others when using ICT-based communication.			
		A	For printing and viewing on screen			
11	Evaluate own use of ICT tools	B	Check for accuracy and meaning			
		At each stage of a task and at the task's completion				
					XX	

Appendix 3

Activity	Skill Standard ID	N Tasks	N items	%
1. Use application to source information		1		
Open application	1			
Read instructions/information, plan action	2,6,9			
Select appropriate information	1,6			
Create folder	3			
Save information	3			
Communicate information	9			
2. Use application to solve problems		3		
Select appropriate application	2			
Search information	5			
Select or produce information	6,7,8,10			
Communicate information	9			
Evaluate/improve artefact, product submitted	11			
Save output	3			
3. Answer questions	4,6,9	2	4	
Total			40	100%

Table 2. An example of a process-based test specification for FS ICT Level 1.