An evaluation of generated question sequences based on competency modelling

Onjira SITTHISAK[a*] & Lester GILBERT[b]
[a] *Computer and Information Technology Group, Faculty of Science, Thaksin University, Thailand*
[b] *Learning Societies Lab, School of Electronics and Computer Science, University of Southampton, United Kingdom*
*onjira@tsu.ac.th and lg3@ecs.soton.ac.uk

Abstract: In order to support lifelong learning, assessment systems have to focus on representation and updating a variety of knowledge domains, rules, assessments and learner's competency profiles. Adaptive assessment provides efficient and personalised routes to establishing the proficiencies of learners. Existing adaptive assessment systems are faced the challenge of dealing with inconsistently measuring and representing student's knowledge. We can envisage a future in which learners are able to maintain and expose their competency profile to multiple services, throughout their life, which will use the competency information in the model to personalise assessment. This paper presents an adaptive assessment system based on a competency model. The system automatically generates questions from a competency framework and sequence the questions based on the taxonomies of subject matter or of capability, making it possible to guide learners in developing questions and testing knowledge for themselves. The questions and their sequencing are constructed from a given set of learning outcomes and the subject matter recorded in an ontological database. The architecture of the system and the mechanism of sequencing the questions are discussed.

Keywords: competency, adaptive assessment, ontology, IMS QTI

Introduction

We introduce the background to the pedagogical and technological issues involved in automatic question generation. A specific approach is described for the automatic generation of questions in any domain by using a particular model of competencies. A system overview of the proposed competency framework, named COMpetence-Based learner knowledge for personalized Assessment (COMBA), is presented. We consider an implementation of COMBA with an ontological database that represents the intended learning outcome to be assessed across a number of dimensions such as levels of cognitive ability and subject matter content involved, an experiment to test its outputs, and the results. Finally, we present some discussion of generated question sequences and conclusions.

1. Background

1.1 Automatically generating questions

Questioning is useful because it challenges learners to respond and it reveals learners' abilities to reason, create, analyse, synthesise, and evaluate. A question should relate to the learning outcomes being measured. Question phrasing should be precise, clear, and easy to understand by using the simplest possible language [1-3]. Good questions should appropriately challenge learners in order to stimulate them to think more deeply about the subject matter. Finally, a good question should help the learner to identify where further study may be useful.

There are currently many systems available to generate questions automatically; these are however confined to specific domains. A number of pioneering systems such as Problets [4], ILE [5], QuizPACK [6], and Jeliot 3 [7], explored the use of automatic generation of questions using parameterised templates. The basic concept uses templates instantiated with random values to generate the questions. A question's template is able to produce a large number of different questions. Problets and Jeliot 3 generate questions about programming using computer language templates. The question generation of Problets is language independent, whereas Jeliot currently supports only Java. Problets and Jeliot are self-contained, lacking interoperability with other systems such as institutional-wide e-learning systems. ILE is a tool that automatically generates exercises for the special case of electric AC circuit problems, given global parameters such as the number of nodes and number of branches. QuizPACK works on automatic evaluation of code-execution questions. A teacher provides the core content of a question, a parameterised fragment of code to be executed, and a variable within that code. QuizPACK randomly generates the value of the question parameter, creates a presentation of the resulting question, and runs the presented code in order to generate the correct answer.

These applications of parameterised questions were developed for computer programming. A correct answer to a parameterised question can be calculated by a formula or executed by a standard language complier without the need for a teacher or author to provide it. Currently, such systems offer remarkable automatic generation of questions, but only for specific domains, and lack integration, interoperability, portability and reusability.

1.2 Adaptive assessment and its applications

Adaptive assessment system aims to assess a learner's competency by posing a minimum number of questions in order to decrease test length, which is one of the main goals in adaptive assessment [8-10]. Another main goal includes offering personalized support according to the needs and ability of each learner [11]. Work related to the proposed approach can be found in the areas of adaptive assessment system. Many adaptive assessment systems have been developed such as A Web-based English CAT prototype system [12], IDEAL [13], Personal-reader [14], COMPASS [15], SIETTE [16], and CosyQTI [17]. These systems are described below.

A Web-based English CAT prototype system and IDEAL are focused on using Item response theory (IRT) to estimate the numerical value of learner's ability level, in order to determine the next item to be posed, and to decide when to finish the test, rather than to assess learners' readiness for further learning. One of the major challenges facing the use of IRT is establishing standards for usability and interpretability issues of the IRT value [18, 19]. In IRT, ability is measured by a scale point. When applied this theory to measure cognitive skills expected to be tested in each learning outcome such as Knowledge, Comprehension, Application, and Evaluation, the theory has some limitations [20].

Personal-reader is developed to personalize a learner's assessment at each moment of the learning process. There are two types of learning content: atomic learning object and linear learning object. In the case where the learner gives wrong answers, the assessment framework should detect the atomic learning objects that have to be studied again, highlights them and gives, if necessary, some additional links that could be used to better understand the current lesson. In the case where the answers are correct, the learner is allowed to continue. Then new course material is generated in the next linear learning

object. In summary, this system still has problems of representing learning knowledge and has difficulty with problem solving.
Concept MaP ASSessment tool (COMPASS) is an adaptive web-based concept map assessment tool. Based on an assessment goal that the learner selects from a set of proposed goals, COMPASS engages learners to the assessment and learning process through a set of assessment activities. The system provides different informative, tutoring and reflective feedback components, tailored to learners' individual characteristics and needs by using weight and error categories. The level of performance is represented by Gogoulou's taxonomy [15]. . In summary, this system still has the problems of collaboration with many teachers, and the use of numerous parameters associated with each question for teachers who are usually practically focused and who would have difficulty with controlling user interaction.
Spanish translation of Intelligent Evaluation System using Tests for TeleEducation (SIETTE) is a web-based tool to assist teachers and instructors in the assessment process. The system can be used in two different ways. First, teachers can use it to develop the tests that are defined by their topics, questions, parameters, and specifications. Second, learners can use it to take the tests that are automatically generated according to the specifications, and adapted to the learner's knowledge level. Question selection is based on a function that estimates the probability of a correct answer by using Item Response Theory, leading to an estimation of the learner's knowledge level. This system has the problem with estimating learner's knowledge level of each topic in each test.
The CosyQTI tool supports the authoring process and presentation of personalized and adaptive web-based assessment. The adaptation will be provided by using a form of the IF-THEN rule's trigger point which is a point for activation. This system has not been tested in full in real classroom environments. There are still some problems with estimating and representing learner's knowledge level and formal testing within real environments.

2. COMBA system

We have developed an improved competency model, named COMpetence-Based learner knowledge for personalized Assessment (COMBA), which uses ontologies. The model has been used to automate question generation in adaptive assessment systems. The system focuses on the identification and integration of appropriate subject matter content (represented by a content taxonomy) and appropriate cognitive ability (represented by a capability taxonomy) into a hierarchy of competencies. The resulting competencies structure has been shown to be able to generate questions and tests for formative and summative assessment. These questions can be expressed as IMS Question and Test Interoperability (IMS QTI) compatible XML files to enable interoperability.
The system was built on an ontological database that describes the resources (subject matter, capability, competency) and the relationships between them. An assessment for a competency often actually tests component competencies, and is supported by the linked nature of the competencies hierarchy. For example, a statistics course may test knowledge of the confidence interval [21] by testing the students' ability to calculate, explain, and define the confidence interval in a variety of situations. An assessment item can be directly formulated from a competence by using the parameters of that competence: capability, subject matter content, and other contextual elements. For example, the assessment corresponding to the learning outcome, "Students understand the concept of a confidence interval" might be something like "Calculate the confidence interval for the following situation", or "Explain the importance of the confidence interval in the following situation", or "Define standard error".

2.1 *The competency model*

COMBA is informed by the results of comparing the competency standards against the desired taxonomy of competence [22]. The improved competency model is represented in Figure 1. A competency involves a capability associated with subject matter content and optionally a contextualisation (the situation or scenario, tools, and standard of performance). A competency can be linked to one or more resources, and a student may evidence a competency in one or more ways.

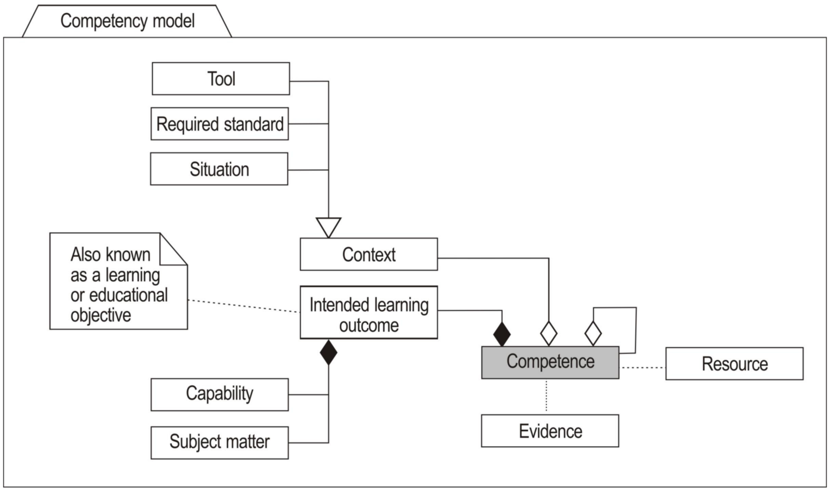

Figure 1 Competency model

Capability is behaviour that can be observed, based on a domain taxonomy of learning such as Bloom's [23], Gagné's Nine Areas of Skill [24], or Merrill's Cognitive Domain [25]. Subject matter content is the subject domain of what the student can do by the end of course. The competency evidence substantiates the existence, sufficiency, or level of the competency, and might include test results, reports, evaluation, certificates, or licenses. External knowledge resources and tools support and promote the problem solving, activity performance or situation handling of the competency. The situation identifies the particular circumstances and conditions of the competency, for example, its time limit.

The proposed competency model involves three important principles: an orientation towards, and focus upon, activity-based teaching and learning, the identification and integration of appropriate subject matter content within a broader teaching and learning context, represented by a hierarchy of linked competencies, and the identification of the assessment that would demonstrate successful teaching and learning has been accomplished.

2.2 *Architecture of COMBA system*

The COMBA implementation consists of a number of modules, illustrated in Figure 2. The Competence navigator is responsible for retrieving the requested competence, based on the domain request from the student, and passing the competence to the Subject Matter Content and Capability navigator modules. In using the model for the automatic generation of questions, the relevant subject matter and capability data, together with the authoring question template files, are assembled to generate questions derived from the matrix of competencies crossed with cognitive abilities. Given a question which is now ready for further use, it is formatted using the QTI specification.

The QTI specification facilitates the sharing of questions and tests, enabling investment in the development of common tools such as Web-based authoring and delivery applications. For an adaptive test, this specification supports the use of pre-conditions and branching, allowing the embedding of sequencing and adaptive logic into a test. Adaptivity is limited to

the questions referred to within the test. As a result, if the student answered, it may not be possible to branch in directions not provided in the test. In addition, the inability to import external data may limit adaptivity. In order to develop a test, the generated questions are linked together for storing in a test bank. For the delivery of the test, the system deploys an assessment delivery service (QTI tools[1)]) to allow a student to view a question, to answer it, to receive feedback, and to view the assessment results.
In the COMBA system, the ontology was based on OWL-Lite [26] which was sufficiently expressive to describe the subject matter hierarchy and provides for higher performance reasoning. The ontologies adhere to the criteria of ontology design: clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment [27].

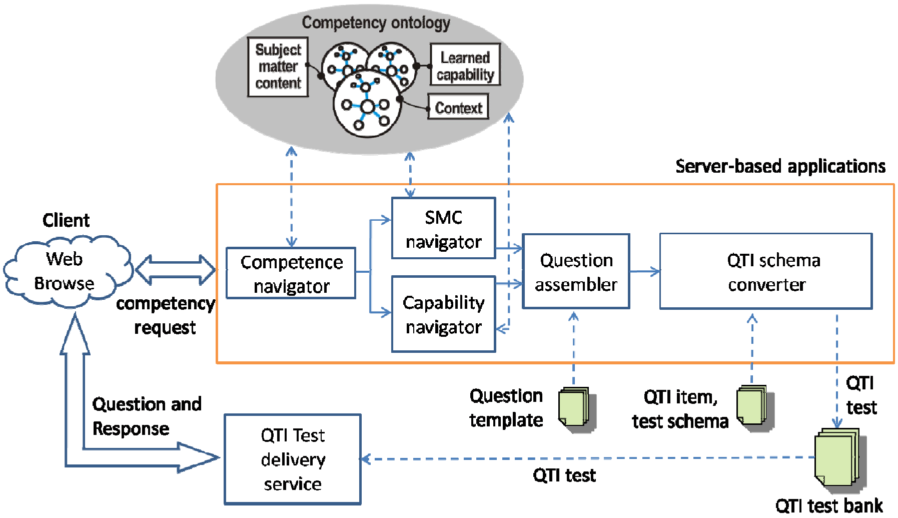

Figure 2 Architecture of the COMBA system

3. Using COMBA for generating adaptive question sequences

```
<?xml version="1.0" encoding="UTF-8" ?>
- <assessmentTest xmlns="http://www.imsglobal.org/xsd/imsqti_v2p1"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.imsglobal.org/xsd/imsqti_v2p1
   http://www.imsglobal.org/xsd/imsqti_v2p1.xsd" identifier="TEST"
   title="Function Point Analysis Test">
 - <testPart identifier="part1" navigationMode="linear" submissionMode="individual">
     <itemSessionControl showFeedback="true" />
   - <assessmentSection identifier="sectionquestion1" title="Section question:"
       visible="true">                                                    A
       <assessmentItemRef identifier="question1" href="question1.xml" />
     </assessmentSection>
   - <assessmentSection identifier="sectionRquestion1" title="SectionR question1"
       visible="true">
     - <branchRule target="EXIT_TEST">
       - <equal toleranceMode="exact">                                    B
           <variable identifier="question1.SCORE" />
           <baseValue baseType="float">1.0</baseValue>
         </equal>
       </branchRule>
     </assessmentSection>
   - <assessmentSection identifier="sectionquestion2" title="Section question:"
       visible="true">                                                    C
       <assessmentItemRef identifier="question2" href="question2.xml" />
     </assessmentSection>
   </testPart>
 + <outcomeProcessing>
 + <testFeedback access="atEnd" showHide="hide"
     outcomeIdentifier="outcomeIdentifier" identifier="outcomeValue" title="Test
     Feedback">
</assessmentTest>
```

Figure 3 Example of QTI branching rules in XML format

In this section, we present the use of the model in automating question sequence generation. A competency hierarchy supports a variety of adaptive rules to adjust questions to the

[1] http://playr.qtitools.org/playr/

students' capability and to the nature of their knowledge. Many methods of traversing the competency hierarchy may be applied, involving different starting points and algorithms. These methods may lead to interesting issues which should be considered in adapting to the students' particular talents, strengths, weakness, and own learning preferences. Within a test constructed according to the IMS QTI specification, the sequencing and adaptive logic are expressed in branching rules. For example, an adaptive sequence may provide a question at a slightly higher level if a student succeeds or a question at a lower level otherwise. Figure 3 presents an example QTI question file for adaptive assessment using QTI constructs which may be incorporated into a test. Portions labelled A and C show the student items called "question1" and "question2" respectively. The portion labelled B illustrates a branching rule. If the student succeeds on question1, the test jumps forward to the end of the test (shown as branchRule target= 'EXIT_TEST') or goes to "question2" in the section labelled C otherwise.

3.1 Experimental validation of generated question sequences

An experiment was designed to validate a sequence of questions, generated using the COMBA model. The particular sequence experienced by a student was dependent upon the student's answers, and so was adaptive. If the student succeeded on a question, where possible the next question was a question at the same capability level and at a higher subject matter level than the previous question. If the student failed the question, the system presented where possible an easier question. This was a question at the same capability level and at the lower subject matter level than the previous question. Questions started from the highest subject matter level and the highest ability level, and the sequence stopped when the student answered a question correctly.

The experiment focused on the opinions of students on the efficiency and effectiveness of the adaptive sequence. The questions explored student ratings of the sequencing, on the criteria of fairly assessing their knowledge (TestAssessKw), helping them to understand how a given learning outcome separated into "learning outcome components" (DecomposeLO), helping them to separate a given learning outcome into "topics" (DecomposeTopic), adapting to their level of knowledge (AdaptQuestion), being useful for self-assessment (UsefulForSelfAssessment), identifying their lack of knowledge (IdentLO), and providing appropriately difficult questions (ShowDifficultQ).

Competencies were collected from the INFO2007 Systems Analysis and Design course at the University of Southampton. The topic of the course instantiated in the model involved function point analysis and associated issues including: adjusted function points, unadjusted function points, complexity adjustment, the formula for complexity adjustment, degrees of influence, the formula for unadjusted function points, and calculating function points from an ER Diagram. The participants were voluntary 2nd year undergraduate students. Instruction sheets were distributed to all attending students at the end of a lecture, and asked the students to rate the generated questions against the criteria on a 4-point forced-choice Likert scale ('Strongly disagree', 'Disagree', 'Agree', 'Strongly agree', coded as 1, 2, 3, and 4 respectively) that best described their opinion.

3.2 Results and discussion

The study gathered data from 19 students. A one-sample t test was used to test differences between the observed sample means and an expected sample mean of 2.5, being mid-way between agreeing and disagreeing on the measurement scale. As can be seen in Table 1, the mean rating was significantly higher than 2.5 for 9 of the 12 measured variables.

The students did not think that the test particularly assessed their knowledge on average. It is not clear why they thought this; one hypothesis is that the 'stopping rule' (at the first correct answer) did not give them confidence that their knowledge had indeed been thoroughly tested.
Interestingly, the students agreed that the adaptive sequence helped them to understand how a given learning outcome separated into "learning outcome components", but they did not agree that it helped them to separate a given learning outcome into "topics". Whilst a learning outcome component involves capability and subject matter, a topic involves only subject matter. This suggests that the generated questions helped the students to understand the decomposition of capability, but were not particularly helpful in understanding the decomposition of topics.

Measured Variables	Test Value = 2.5			
	t	**df**	**Sig. (2-tailed)**	**Mean Difference**
TestAssessKw	–0.224	18	0.826	–0.026
AdaptQuestion	5.786	18	0.000	0.711
UsefulforSelfAssessment	2.471	18	0.024	0.500
IdentLO	3.269	18	0.004	0.500
DecomposeLO	3.139	18	0.006	0.447
DecomposeTopic	0.907	18	0.376	0.184
ShowDifficultQ	8.367	18	0.000	0.605

Table 1 t Test

The results of the remaining t-tests were straightforward: the students agreed that their question sequence was well adapted, was useful for self-assessment, helped identify their lack of knowledge, and provided appropriately difficult questions. Broadly speaking, this experiment and the earlier one (reported in [28]) show that the questions and the adaptive test sequences were acceptable to students, and hence that the COMBA model is capable of generating good assessments.

4. Conclusion

While this study successfully demonstrates a data model and a method of automatically generating acceptable and useful questions and sequences, representing competencies and the subject matter is the critical challenge. In addition, more effective algorithms are needed for generating questions and sequences. Any generating mechanism must ensure a high standard of English grammar in the resulting questions. The revised generating mechanisms in this experiment reduced some inappropriate format of questions by using revised SPARQL queries to expand the returned results. This indicates that not only the format of the template itself is important for generating questions and sequences using parameterised templates, but also the algorithm of querying is critical.
The key contribution is supporting a variety of ways of developing adaptive sequences. Future work could focus on methods for generating adaptive question sequences and considered their pedagogical value. For example, it is possible that students might have differing abilities in quite similar content areas. In this case, learners may not achieve an appropriate level of their capability and content. New adaptive question sequences could employ different traversal algorithms. If the learner failed a question, the system could present the next question at a lower capability level and at the same subject matter level; or at the same capability level and at the nearest subject matter level to the previous question.

The pedagogical value of a particular method would need further investigation for successful learning and teaching, but having such varieties of methods could provide fruitful areas of exploration.

References

[1.] Horton, W.K., *E-Learning by Design*. (2006), San Francisco, USA: Pfeiffer An Imprint of Wiley.

[2.] Silyn-Roberts, H., *Writing for Science*. (1996), Auckland, New Zealand: Addison Wesley Longman New Zealand Limited.

[3.] McComas, W.F. and L. Abraham, *Asking more effective questions*. (2005), University of Southern California.

[4.] Dancik, G. and A. Kumar. *A Tutor for Counter-Controlled Loop Concepts and Its Evaluation*. in *Frontiers in Education Conference*. (2003). USA.

[5.] Cristea, P. and R. Tuduce, *Automatic Generation of Exercises for Self-testing in Adaptive E-learning Systems: Exercises on AC Circuits*, in *3rd Workshop on Adaptive and Adaptable Educational Hypermedia at the AIED'05 conference, (A3EH)*. (2005).

[6.] Brusilovsky, P. and S. Sosnovsky, *Individualized exercises for self-assessment of programming knowledge: An evaluation of QuizPACK.* Educational Resources in Computing (JERIC), (2005). **5**(3): p. 6.

[7.] Myller, N., *Automatic Generation of Prediction Questions during Program Visualization.* Electronic Notes in Theoretical Computer Science (ENTCS), (2007). **178**: p. 43-49.

[8.] Welch, R.E. and T.W. Frick, *Computerized adaptive testing in instructional settings.* Educational Technology Research and Development, (1993). **41**(3): p. 47-62.

[9.] Askins, D., *Adaptive Assessment*, in *Toolkit - The Nuts and Bolts e-Newsletter*. (2004).

[10.] Way, W.D., *Practical Questions in Introducing Computerized Adaptive Testing for K-12 Assessments*, in *Research Report 05-03*. (2005).

[11.] Brusilovsky, P., *Methods and Techniques of Adaptive Hypermedia.* User Modeling and User-Adapted Interaction, (1996). **6**(2-3): p. 87-129.

[12.] Yu-Huo, T., W. Yu-Lung, and C. Hsin-Yi, *A Practical Computer Adaptive Testing Model for Small-Scale Scenarios.* Educational Technology & Society, (2008). **11**(3): p. 259-274.

[13.] Wang, F.H. *Application of componential IRT model for diagnostic test in a standard conformant eLearning system*. in *the Sixth International Conference on Advanced Learning Technologies*. (2006). Kerkrade, The Netherlands.

[14.] Cheniti-Belcadhi, L., N. Henze, and R. Braham. *Towards a Service Based Architecture for Assessment*. in *Proceedings of 13th Annual Workshop of the SIG Adaptivity and User Modeling in Interactive Systems (ABIS 2005)*. (2005). Saarbrücken, Germany: German Informatics Society (GI).

[15.] Gogoulou, A., et al., *Adapting the Communication-Scaffolding Tools in a Web-based Collaborative Learning Environment.* Proceedings of the ED-MEDIA 2004, (2004). **2004**(1): p. 1153-1161.

[16.] Conejo, R., et al., *SIETTE: A Web–Based Tool for Adaptive Testing.* International Journal of Artificial Intelligence in Education, (2004). **1**(33).

[17.] Lalos, P., S. Retalis, and Y. Psaromiligkos, *Creating personalised quizzes both to the learner and to the access device characteristics: the Case of CosyQTI.* (2005).

[18.] Hays, R.D. and J. Lipscomb, *Next steps for use of item response theory in the assessment of health outcomes.* Quality of Life Research, (2007). **16**: p. 195-199.

[19.] Gibbons, R.D., J.C. Immekus, and R.D. Bock, *The Added Value of Multidimensional IRT Models*. (2007).

[20.] Junker, B.W. and K. Sijtsma, *Nonparametric Item Response Theory in Action.* Applied Psychological Measurement, (2001). **25**(3): p. 211 - 220.

[21.] Field, A., *Discovering Statistics Using SPSS*. (2005), London: SAGE Publications Ltd.

[22.] Sitthisak, O., et al. *Adapting health care competencies to a formal competency model*. in *the ICALT*. (2007). Niigata, Japan: IEEE Computer Society Press.

[23.] Bloom, B.S. and D.R. Krathwohl, *Taxonomy of educational objectives: The classification of educational goals by a committee of college and university examiners*, ed. H.I.C. Domain. (1956), New York: Longman.

[24.] Gagne, R., *Robert Gagne's Nine Steps of Instruction*. (1970).

[25.] Merrill, M.D., *Instructional Transaction Theory (ITT): Instructional Design Based on Knowledge Objects*. (1999).

[26.] W3C, *OWL Web Ontology Language (OWL)*. (2004).

[27.] Kalfoglou, Y., *Exploring Ontologies*. Handbook of Software Engineering and Knowledge Engineering. Vol. 1: Fundamentals. (2001), Washington, D.C.: World Scientific Publishing.

[28.] Sitthisak, O., L. Gilbert, and H.C. Davis, *An evaluation of pedagogically informed parameterised questions for self assessment.* Learning, Media and Technology, (2008). **33**(3): p. 235-248.