

**Design and Development of
Digital Closed Questions:
A Methodology for Midsized
Projects in Higher Education**

**Active Learning, Transparent
Assessment - ALTB**

Credits

Design and Development of Digital Closed Questions: A Methodology for Midsized Projects in Higher Education.

Active Learning, Transparent Assessment - ALTB

SURFfoundation
PO Box 2290
NL-3500 GG Utrecht
T + 31 30 234 66 00
F + 31 30 233 29 60
E info@surf.nl
W www.surf.nl/en

Authors

Rob Hartog, *Wageningen MultiMedia Research Centre, Wageningen University*

Silvester Draaijer, *Centre for Educational Training, Assessment and Research (CETAR), Vrije Universiteit Amsterdam*

Mia van Boxel, *Vocational and Higher Education, Cito*

Joke Hofstee, *Information and Communication Technology, Cito*

Ignace Latour, *Information and Communication Technology, Cito*

Luuk Rietveld, *Faculty of Civil Engineering and Geosciences, Delft University of Technology*

Huub Verstralen, *Centre for Psychometric Research, Cito*

Pierre Gorissen, *Department of Educational Development, Fontys University of Applied Science*

Editor

Rob J.M. Hartog, *Wageningen MultiMedia Research Centre, Wageningen University*

The ALTB Project has been realized with support of SURF Foundation.

SURF is the collaborative organisation for higher education institutions and research institutes aimed at breakthrough innovation in ICT (www.surf.nl/en)

This publication can be downloaded from the SURFfoundation website:
www.surffoundation.nl/en/publications

© Stichting SURF
March 2008
ISBN 9789078887065

This publication is published under the Creative Commons *Attribution-NonCommercial-NoDerivatives 3.0 Netherlands License*. For more information see: www.creativecommons.nl

Preface

This book is one of the results from the project 'Active Learning, Transparent Assessment' (Actief Leren, Transparant Beoordelen – ALTB). It reflects the experience in design and development of a little more than 2000 of closed questions in fifteen small to midsized projects at four universities: Wageningen University (WU), Vrije Universiteit (VU), Delft University of Technology (TUDelft) and Fontys University of Applied Science (Fontys) and converts this experience in a methodology. Practical problems in the earlier projects led to literature searches, formation of theory and application of methodological results in subsequent projects. Thus, *design* is the core of this book. The methodology presented in this book is the result of a design process and the methodology is a methodology for the design and development of digital closed questions.

Acknowledgements

Apart from the authors, many people contributed to the ALTB project. We thank all these people for their efforts.

Many of the questions were designed by J. Vaessen (WU), A. Grefte (TUDelft), K. Teunissen (TUDelft) and C.J.L.H. Camps (VU), who worked as assistants for subject matter experts in the project. Not only did they design many questions they also realized the questions in Blackboard or QMP and carried out many other tasks that are described in Chapter 7. Other sets of questions were designed by W.M. Brandt (WU), E.P.J. Boer (WU), and J. Diederer (WU) and realized by C. Oomen (WU) and M. Weber (WU).

P. Bastings, P. Boelens, A. Bootsma, A. Goossens, J.P.A. Uijlen, J.C. Slaa and W. van der Zanden (Fontys) designed many questions in their role as Subject Matter Expert. We thank them for taking up the challenge of matching the requirements of competency based education with the possibilities of closed question types supported by N@tschool and W. van Vegchel for realizing questions in N@tschool. We thank A. van Ginneken for taking over the coordination of the ALTB activities at Fontys.

Many subject matter experts validated questions, formulated validation requirements for 'their' learning objectives, provided inspiration and designed questions as well: T. Abee (WU), G.M. Alink (WU), G. Beldman (WU), R.R. Beumer (WU), G.C. van den Bos (VU), J.C. van Dijk (TUDelft), J. Diederer (WU), L.G.M. Gorris (WU), A.J. Greven (VU), R.J.P. Musters (VU), S.G.J. Heijman (TUDelft), G. Jansen (WU), E. Kampman (WU), A.J. Murk (WU), M.W. Reij (WU), I.M.C.M. Rietjens (WU), L.C. Rietveld (TUDelft), A.G.J. Velthuis (WU) and M.H. Zwietering (WU).

J. Hofstee (Cito) contributed much to the final and very successful ALTB workshop of which the materials have been made available through the ALTB website <http://fbt.wur.nl/ALTB>

M.I. Schade (VU), J. Hofstee (Cito), T. Lampe (Cito), S. Draaijer (VU), C. Sluijter (Cito), J. van Weeren (Cito), and H. Verstralen (Cito) provided much advice on assessment and testing and valuable suggestions and pointers to literature in various stages of the project.

P. Gorissen (Fontys), I. Latour (Cito), A. Kassahun (WU) and H. v.d. Schaaf (WU) provided advice and improved our insights with respect to the QTI2 specification. In particular, we thank A. Kassahun (WU) who realized a limited QTI2 delivery system at Wageningen University for making this system available to the ALTB project as well and for all the extra efforts he took to support the project. Thanks to P. van Gremberghe (Cito) for realizing a mockup for a very simple QTI authoring and question management system.

We are indebted to A. Terlouw (WU), P.L. v.d. Togt (WU) and G.H. Folkerts (WU) who provided much help in order to realize the computer-based exams that were part of the ALTB project, and to W.M.A.M. van Dongen (WU) and D. Bouchaut (WU) for sharing their expertise on the use of QMP3 within the WU context.

We thank E. van Puffelen (Cito) who provided most of the original plans of the project, but unfortunately could not take part in its execution and to M. van Boxel who, rather ad hoc, took over his duties as co-projectleader in the first stage of the project.

We thank P. Brascamp (WU) for reading Chapter 1, R. Young (JISC CETIS UK), M. van Boxel (Cito) and J. Hofstee (Cito) for reading Chapter 2, C. Sluijter (Cito), T. Lampe (Cito), W. de Klijn (Cito), I. Latour (Cito) and H. v.d. Schaaf (WU) for reading Chapter 3, J.J. Beishuizen (VU) for reading Chapters 4 and 6, M.I. Schade (VU), M.L. Lunenberg (VU) and P.L.B. van Boxel (VU) for reading several parts of the text. Finally, we thank C. Bok (SURF) for reading the complete manuscript. We are grateful for to all these readers for their time and their valuable and constructive comments and suggestions.

Many thanks are due to C. Oomen who was project secretary and assisted the project leader throughout the project in a wide range of tasks including many administrative duties, realizing many questions according to the QT12 specification and arranging most of the layout of this book, but most of all for being always available.

The ALTB Project has been realized with support of the SURF Foundation. The SURF Foundation is the Higher Education and Research partnership organisation for network services in the Netherlands. More information about the SURF Foundation can be found on <http://www.surf.nl>

Rob J.M. Hartog
Wageningen, March 2008

Table of contents

1	Introduction.....	9
1.1	The aim of the ALTB project.....	9
1.2	A methodology for design and development.....	9
1.3	Innovative closed question types.....	10
1.4	User roles in designing digital items for higher education.....	10
1.5	Primary results of the ALTB project.....	10
1.5.1	A conceptual framework and taxonomy of item types.....	10
1.5.2	Design requirements and scoring rules.....	11
1.5.3	Design guidelines and how to use them.....	11
1.5.4	Design patterns and paradigm examples of digital closes questions.....	12
1.5.5	Scenarios and budget templates for midsized projects.....	13
1.5.6	Insight in the required functionality of future LMSs and CBA systems.....	14
1.5.7	Insight in the possibilities of question and test interoperability.....	14
1.5.8	Training materials.....	14
1.5.9	Further research.....	15
1.5.10	The 'cluster of five approach'.....	15
1.5.11	Quality.....	15
1.5.12	Design and Development of closed questions in competency based education.....	16
1.6	Methodologies are never complete.....	16
1.7	References.....	18
2	A Response-Based Taxonomy of Closed Questions.....	19
2.1	Introduction.....	19
2.2	Towards requirements for a conceptual framework.....	20
2.3	Description of the Conceptual Question Framework (CQF).....	21
2.3.1	Definitions.....	21
2.3.2	The question structure diagram.....	22
2.3.3	Taxonomy of closed questions.....	23
2.4	Application of the framework.....	24
2.4.1	Some examples.....	24
2.4.2	Positioning question types of QMP in CQF.....	32
2.4.3	Positioning QT12 interaction types in CQF.....	33
2.5	Conclusions and summary.....	34
2.6	References.....	35
3	Multiple Response Questions in Computerized Testing.....	37
3.1	Introduction.....	37
3.2	Definition and types of multiple response questions.....	38
3.2.1	Type 1: A MR-multiple true/false has a subset of independent correct options.....	38
3.2.2	Type 2: A MR-combination has one correct subset of interdependent options.....	39
3.2.3	Type 3: A MR-multi-combination has more than one correct subset of interdependent options.....	39
3.3	Design requirements and guidelines for MR-questions.....	40
3.3.1	If it complies with the aim of the question the number of options must be fixed by instruction in the stem.....	42
3.3.2	If the number of options is fixed by instruction, then the number of correct options must be half the total number of the options or less and the total number of options must be at least four.....	42
3.3.3	The maximum score of the question must be indicated in the test.....	42
3.3.4	Guidelines for writing MR-questions.....	42
3.3.5	Some examples of the use of Multiple Response Questions in higher education.....	43
3.4	The scoring of multiple response questions.....	45
3.4.1	Types of item scores for subset selection.....	46
3.4.2	Types of item scores for subset choices of fixed size by instruction.....	46
3.4.3	Types of item scores for subset choices of free size.....	47
3.4.4	Combining item scores to a test score.....	48
3.4.5	Item scoring types for rank orders.....	49

3.5	Conclusions.....	52
3.6	References.....	53
4	Guidelines for the Design of Digital Closed Questions for Assessment and Learning in Higher Education.....	55
4.1	Introduction.....	55
4.1.1	Limitations in current literature on design guidelines.....	56
4.2	The guidelines: dimensions of inspiration.....	56
4.2.1	A: Professional context.....	57
4.2.2	B: Interactions.....	57
4.2.3	C: Design patterns.....	58
4.2.4	D: Textbooks.....	59
4.2.5	E: Learning Objectives.....	59
4.2.6	F: Students.....	60
4.2.7	G: Sources.....	60
4.2.8	H: Motivation.....	61
4.2.9	I: Validity.....	61
4.2.10	J: Equivalence.....	61
4.3	Case studies to investigate the appropriateness of the developed guidelines.....	62
4.3.1	Criteria for assessing the value of the guidelines.....	62
4.3.2	Observations.....	63
4.3.3	Evaluation of the set of guidelines.....	65
4.4	Conclusions.....	66
4.4.1	A set of guidelines is an inspirational source for question design but must be embedded in a broader approach.....	66
4.4.2	Design patterns have the potential to be a powerful aid.....	66
4.4.3	A question design methodology must be geared towards educational technologists.....	66
4.5	References.....	66
	APPENDIX 1 - OVERVIEW OF CASE STUDIES.....	68
	APPENDIX 2 - OVERVIEW OF CASE AND THE USE OR NON-USE OF GUIDELINES.....	69
5	Design of Digital Closed Questions: Procedures for Deciding when to Use which Guidelines.....	81
5.1	Introduction.....	81
5.2	A procedure for Design & Development of closed questions for the CBA role.....	82
5.3	How to use guidelines for Design & Development of closed questions for the ALM role.....	83
5.3.1	No clusters of five and no complete coverage.....	83
5.3.2	The two major situations.....	83
5.3.3	Situation 1: ready available presentations, lecture notes and/or textbooks.....	84
5.3.4	Situation 2: the set of closed questions is intended to be THE learning material.....	85
5.4	Concluding remarks and further research.....	85
5.5	References.....	86
6	Design Patterns for Digital Item Types in Higher Education.....	87
6.1	Introduction.....	87
6.1.1	Focus on design patterns.....	87
6.1.2	New opportunities for designing items for computer-based assessment and learning management systems.....	88
6.1.3	User roles in designing digital items for higher education.....	88
6.1.4	ALTB project.....	88
6.1.5	Information sources on the Design and Development of digital items.....	88
6.1.6	Design patterns.....	89
6.1.7	Design patterns for item design.....	90
6.2	A template for describing design patterns for digital items.....	90
6.2.1	Introduction.....	90
6.3	Selected design patterns for digital items.....	93
6.4	Conclusions.....	114
6.5	References.....	115

7	Practical Aspects of Task Allocation in Design and Development of Digital Closed Questions in Higher Education.....	117
7.1	Introduction.....	117
7.2	Method.....	118
7.3	Description of the Design and development contexts.....	119
7.3.1	Question authoring and delivery environments.....	119
7.3.2	Different functions and different competencies.....	119
7.4	Practical TASK ANALYSIS.....	122
7.4.1	Defining the Project.....	122
7.4.2	Setting Up the Project.....	122
7.4.3	Collecting and defining learning objectives.....	123
7.4.4	Design and intermediate representation of questions.....	124
7.4.5	Validating questions.....	125
7.4.6	Revising questions.....	125
7.4.7	Image processing.....	126
7.4.8	Realization in CBA system.....	126
7.4.9	Additional CBA-related tasks.....	127
7.4.10	Additional management and communication within the team.....	127
7.5	Ten scenarios.....	128
7.6	Conclusions and discussion.....	128
7.7	References.....	132
8	Computer Support for Design and Development of Innovative Closed Questions.....	133
8.1	Introduction.....	133
8.2	IMS Question and Test Interoperability Specification.....	134
8.2.1	The IMS Global Learning Consortium, Inc.....	134
8.2.2	Short history of QTI.....	134
8.2.3	What is new in the QTI 2.1 specification.....	134
8.2.4	Implementation related changes.....	137
8.3	Innovative items and QTI.....	137
8.3.1	Item format.....	138
8.3.2	Response action.....	140
8.3.3	Media inclusion.....	140
8.3.4	Level of interactivity.....	140
8.3.5	Scoring method.....	141
8.4	Item development in existing systems.....	141
8.4.1	Workflow control.....	142
8.4.2	Data Entry Templates.....	142
8.4.3	Import / Export.....	142
8.5	An Integrated Item Design and Development Environment.....	143
8.5.1	Service oriented architecture.....	143
8.5.2	Generic functionality.....	143
8.5.3	Design and Development of items.....	143
8.5.4	Combining items.....	144
8.6	Conclusions.....	144
8.7	References.....	145

List of tables

Table 1:	Overview of intended audiences for each of the chapters in the book.....	17
Table 2:	Definitions of the basic concepts in the CQF	22
Table 3:	Taxonomy of Closed Questions (limited)	24
Table 4:	Question example 1	26
Table 5:	Question example 2	27
Table 6:	Question example 3	28
Table 7:	Question example 4	30
Table 8:	Question example 5	31
Table 9:	Positioning QMP question types in the CQF	32
Table 10:	Mapping of QTI2 interaction types onto the CQF	33
Table 11:	Linking the MR question types to the concepts in the ALTB Conceptual Question Framework.....	38
Table 12:	General item writing requirements and 'guidelines' (Haladyna, 2004p 99 table 5.1) ...	41
Table 13:	Scoring type dependencies derived in this section	48
Table 14:	Two Dimensional Framework by Anderson & Krathwohl (2001).....	91
Table 15:	Overview of 15 small/midsized D&D projects.....	120
Table 16:	Roles, competencies and costs for question design and development.....	121
Table 17:	Benchmark test for entering set of standardized questions with different authoring environments	127
Table 18:	Five scenarios and corresponding budgets for the development of 300 questions for the CBA role.....	130
Table 19:	Five scenarios and corresponding budgets for the development of 300 questions for the ALM role.....	131

List of figures

Figure 1:	Question Structure Diagram as an UML class diagram.....	23
Figure 2 (a & b):	Question example 1 (left) and CQF concepts based on the structure of the response (right)xx	25
Figure 3:	Question example 2	27
Figure 4:	Question example 3	28
Figure 5:	Question example 4	29
Figure 6:	Question example 5 (adapted from (Busstra, 2007)	31
Figure 7:	Fisher information of rank order data and the selection partial credit score of an MR-item with 5 options of which 2 correct, and 2 to be selected for the score	50
Figure 8:	A similar graph for an MR item with six options of which 3 correct and 3 to be selected for the partial credit score	51
Figure 9:	Six options of which 3 correct and 2 to be chosen for the partial credit score.....	51

1 Introduction

Rob Hartog
Wageningen University

Silvester Draaijer
Vrije Universiteit Amsterdam

Motto:

"We believe that the model of basic research by a group of scientists, with results that inform practice by a group of educators, is misconceived.

The search for knowledge and understanding and the development of educational resources must be concurrent concerns and interactive activities.

The alternative vision, which we prefer, has inquiry coupled with the development of resources, so that development is guided by and informs the growth of scientific principles and concepts, and scientific inquiry addresses questions that are important in practice" (Gardner et al., 1990)

1.1 The aim of the ALTB project

This book presents the main results of the SURF project "Active Learning, Transparent Assessment" (Actief Leren, Transparant Beoordelen – ALTB) (Hartog, 2004). The primary aim of the ALTB project was to provide a methodology for design and development of digital closed questions in higher education. The research question of the ALTB project was essentially: 'How and under what conditions is it possible to support the design and development of digital closed questions in higher education? The answer should support the rationale for the methodology.

The SURF ALTB project was carried out in 2005 and 2006 at four universities. A number of tasks in the ALTB project were carried out by a testing and assessment company. The ALTB project incorporated fifteen small and mid-sized projects divided on the design and development of digital items. The aim of these various subprojects was to develop sets of questions for summative use, and for use in quizzes intended for formative applications. A systematic approach to the design and development of digital items was used under a range of conditions, in situations involving various forms of collaboration and types of task division. The intention was to identify the potential of digital items and to determine how they can best be used, to collate people's experiences in design and development teams, and to formulate the lessons learned. These experiences were essential input for the development of a methodology for digital item design.

1.2 A methodology for design and development

In this book, a methodology for design and development (D & D) of closed questions is defined by:

- a conceptual framework and taxonomy ;
- design requirements, i.e. requirements that must be satisfied by the resulting questions;
- design guidelines, i.e. guidelines that give designers direction and help them to arrive at results that satisfy the requirements;
- procedures that define how to use these guidelines;
- design patterns and paradigm examples;
- scenarios that match tasks and resources in different contexts;
- links between the near past, the present and the near future.

At a more detailed level, such a methodology will also provide many do's and don'ts.

1.3 Innovative closed question types

Currently available Computer-based Assessment systems (CBA systems) offer a great variety of digital item types (Bull et al., 2001; Mills et al., 2002; Parshall et al., 2002) such as multiple answer, drop-down lists, numeric, hot-spot, drag-and-drop. These systems also enable a variety of item types to be deployed within a single assessment. The availability of CBA systems and the Internet make it easier than ever before for Subject Matter Experts (SMEs – professors, academics, lecturers, tutors, instructors) to use such innovative item types. In addition, other digital options can be used such as the inclusion of images. Several authors have referred to these item types as innovative. SMEs in many higher education courses are already using digital item types that are made available via CBA systems and Learning Management Systems (LMSs). One recurring problem, however, is how to make optimal use of these new possibilities.

1.4 User roles in designing digital items for higher education

Within the field of higher education, digital test items are usually developed within the context of a course taught by SMEs and their assistants. In general, SMEs and their assistants have limited time for designing and developing such items, as well as limited skills and experience. In practice, Educational Technologists (ETs) experience a growing demand for advice. Furthermore, ETs receive more requests to participate in small to midsized projects on design and develop pools of digital test items. These items are generally used in summative assessments, and in quizzes aimed at stimulating active learning. ETs need a methodology for the design and development of digital items if they are to provide the best possible advice to those involved in projects of this kind.

1.5 Primary results of the ALTB project

1.5.1 A conceptual framework and taxonomy of item types

In answer to requests from the case studies in the ALTB project, a response based closed question framework and a taxonomy of question types was developed. The framework is more fundamental than existing frameworks, precisely because it is only based on the response structure. An important problem with frameworks from literature is that question types with essentially the same response structure are considered as fundamentally different. Other problems with existing frameworks were that the same name often referred to different question types or that the same question type had different names.

Chapter 2 describes the framework developed in ALTB and the corresponding taxonomy. The framework is illustrated with examples of closed questions and comparisons with concepts and types, which are usually defined at an operational level in CBA systems. Finally, the closed question types are matched with the interactions described in the Question and Test Interoperability 2 specification (QTI2). In reading the other chapters of this book the reader should keep in mind that the concepts and type definitions are primarily the *result* of interaction between theory development and needs that were revealed during the projects. The concepts and type definitions were *not input* for the case studies in ALTB.

The framework is directly important for researchers in assessment, system developers and educational technologists. It is not likely that SMEs or their assistants can directly benefit from the framework. However ETs will use the conceptual insights provided by the framework in answering questions from SMEs or their assistants.

	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
H2. Concepts & taxonomy	+/-	+/-	++	++	++	-

In order to enable the non-specialist readers to read any chapter as a stand alone chapter, the newly developed terminology is mainly restricted to Chapter 2 and more traditional labels for question types are still used in the sequel.

1.5.2 Design requirements and scoring rules

For traditional multiple-choice questions literature provides many design requirements. Among digital closed questions three types of questions required much attention in the case studies, in particular in relation to the issue of scoring. These types are in the traditional terminology sometimes called Multiple Response questions ('MR questions'), but they also appear under various other labels. For these question types, the ALTB project provides in Chapter 3 extensions of existing theory, additional design requirements and scoring rules. Chapter 3 relates practical issues with the three types and with essential design requirements and additional guidelines. This chapter aims primarily at ETs and system developers. It is up to system developers to implement these scoring rules. The main function of Chapter 3 in current design and development projects, is that it helps ETs to avoid confusing discussions on the use and scoring of 'MR-questions' and warns ETs to check available systems for support of the ALTB scoring rules.

	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
3. Requirements & scoring	+	+	++	++	++	-

1.5.3 Design guidelines and how to use them

In the ALTB project, extensive experience with sixty design guidelines was recorded in the fifteen case studies. The main conclusions were the following:

1. In each project only a small subset of the ALTB list was applicable
2. This subset was different for different projects
3. Presenting the complete ALTB list to SME's and their assistants is not beneficial and sometimes even counterproductive
4. ETs should select a small subset from the ALTB list and present this subset in the design and development team and focus on this subset and corresponding design patterns.

The ALTB list of design guidelines was derived from literature. This was not as trivial as it may sound. The literature contains long lists of explicit design 'guidelines' for multiple-choice items (T/F, alternate choice, four options) to be used in assessments. See, for example, Haladyna and Downing(2002). The ALTB project, however, made clear that SMEs regard most of these 'guidelines' to be unhelpful. This is due to the fact that such 'guidelines' often actually are requirements in stead of pointers for inspiration. The ALTB case studies made clear that ETs should avoid focusing their advice and participation on the promotion of such requirements in disguise.

Chapter 4 presents and analyzes experience with guidelines in fifteen case studies. Chapter 4 primarily aims at ETs and at the community of assessment researchers.

	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
4. Guidelines	+/-	+/-	++	++	-	-

In order to support ETs in selecting a small subset of guidelines to be used in a specific project for the design and development of pools of closed questions, two selection procedures have been developed in the ALTB project. Chapter 5 presents one procedure for selecting and applying guidelines for design and development of closed questions that will be used in assessments and one procedure for selecting and applying guidelines for design and development of closed questions that will function as activating learning material.

SME's and assistants who want to know more about design guidelines might want to read Chapter 4.

SME's and assistants who just want a small subset of guidelines have two options: they can involve an ET or they can use the procedures in Chapter 5 to select an adequate subset.

	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
5. How to select adequate guidelines	+	+	++	+/-	-	-

1.5.4 Design patterns and paradigm examples of digital closes questions

The ALTB project revealed that ETs and SMEs were seldom able to use example items from literature as paradigm examples or as a source of inspiration. One major problem was that SMEs encountered great difficulty in abstracting the examples. That imposes a barrier to subsequent transformation of those examples for applicability for their own courses.

The ALTB project produced a set of design patterns and corresponding paradigm examples in order to support the design of digital closed questions. In order to present these patterns also a pattern representation format was developed in the ALTB project.

The term "Design Pattern", which was introduced by (Alexander, 1979, p. 206) in the seventies of the last century is a concept used in architectural design. It was adopted for use in software engineering (Gamma et al., 1994) about 15 years later. Design patterns are generic combinations of components or solutions to recurring problems in designs. It is not realistic to suppose that designers design from scratch. On the contrary, an experienced designer is supposed to have many design patterns in mind. "It is only because a person has a pattern language in his mind, that he can be creative when he builds" (Alexander, 1979, p. 206) . Competent designers can instantly match a problem to the appropriate design pattern to arrive at satisfactory solutions to given problems and contexts. Design patterns are therefore an integral component of design methodology.

To date, it is likely that most ETs have internalized only a few design patterns for digital closed quesitons, or that they have very limited numbers of these resources to hand. Yet ETs have the most to gain from the design pattern approach. It would enable them to provide better support for the SMEs, by supplying appropriate design patterns at just the right moment in item-development projects. The design pattern approach allows for a faster, more economical, yet more varied deployment of digital items.

In the ALTB project, more and more design patterns for innovative questions were being developed or derived. Experience in the last case studies is in keeping with the rule that design patterns are important for good design and that design of digital closed questions is not an exception to this rule. In particular, every ET who is involved in design and development of closed questions should have internalized a large set of design patterns.

Presenting examples as a means to help designers in the initial design stages only helps if the subject matter of the example is very close to the subject matter to be covered in the course for which closed questions are being designed. While guidelines often proved to be too abstract and generic, and examples were too concrete and specific, design patterns proved to be more adequate for supporting question designers. The importance of the concept of design patterns as an instrument for a methodology derives from the limitations of individual examples, and the limitations of factors such as the usefulness of guidelines and the value of frameworks.

Chapter 6 presents a set of design patterns and corresponding paradigm examples. The primary intended audiences for this chapter are educational technologists, subject matter experts (professors, lecturers) and their assistant question designers.

	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
6. Design Patterns	++	++	++	+	++	-

1.5.5 Scenarios and budget templates for midsized projects

Ideally, design and development of pools of questions is teamwork. Based on the fifteen ALTB case studies a set of design and development tasks has been identified. Next, tasks and resources were matched. Ten possible scenarios for carrying out midsized design and development projects were developed. For these scenarios budget templates were developed as well. Chapter 7 describes the tasks, scenarios, budget templates. The chapter also provides some estimates of costs based on experience in the fifteen case studies.

From the experience in the ALTB project, it must be concluded that design and development of closed questions that are regarded as appropriate by the subject matter expert (SME), will cost about two hours per question on average. This implies that a project plan should be based on an expected design and development time of two hours per question. It turns out that this estimate is very counter intuitive for most academics. Most academics regard the design and development of questions to be a task that can be executed in much less time. A principal reason why this effort is higher than expected in general, is that the design and development of questions is a cyclic and concentric process. Usually several versions of questions are made before a version is produced that satisfies the design team. Almost every question goes through several implicit or explicit stages of review and refinement. In addition to that, digital and innovative question formats call for even more effort because they add an extra conceptual, technical and management dimension to the design and development process. On the other hand, budgets allocated for design and development of questions as part of test construction projects aimed at many participants (for example on a national scale) are often much larger.

The primary intended audiences for this chapter are SMEs, ETs but also higher level managers within the university.

	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
7. Scenarios, & budget templates	++	+	++	+	++	++

1.5.6 Insight in the required functionality of future LMSs and CBA systems

The scenarios as presented in Chapter 7 are based on the functionality and modality of the most commonly available LMSs and CBA systems. More advanced LMS and CBA functionality would considerably simplify the scenarios.

Chapter 8 describes how the IMS Question and Test Interoperability specification (IMS, 2006f) can be used as a means to support the design and development of innovative closed questions. The chapter describes the five dimensions of innovation that can be distinguished in closed question assessment items and links them to the functionalities supported by the IMS QTI specification. The chapter shows that the QTI specification offers enough flexibility and supports enough functionality to be used as the basis for innovative closed question items and very interactive structures of multiple individual questions. Furthermore, the chapter describes the functionality requirements for a flexible authoring environment for assessment items which surfaced in the ALTB case studies. Finally the chapter clarifies how these functionalities can be realized based on web services in a service oriented architecture.

This ALTB output is rather technical. The primary intended audiences of this chapter are system developers and educational technologists.

G	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
8. Computer Support	-	-	++	+	++	-

1.5.7 Insight in the possibilities of question and test interoperability

In the ALTB project, the expectations about large scale adoption of the QTI 2 specification were initially high. Actual developments with respect to implementation of authoring and delivery systems that are, at least to a certain extent, conform the QTI 2 specification, have been disappointing and a cause for worry as well. In the ALTB project, the QTI 2 specification has mainly contributed in terms of conceptual insight. However, this conceptual insight was limited to very few team members and was only acquired by many hours of study and involvement in design and development of a delivery system for QTI 2. The initial assumption that the concepts would also support question designers, could not be confirmed in the project. The most commonly used terminology about closed questions is confusing and it will take a real effort to change over to another terminology. At least, educational technologists should be able to map the most commonly used terms onto the concepts defined in the QTI 2 specification and concepts based on the structure of the response as defined in Chapter 2. Against the background of the experience with the QTI 2 specification, it is very important that a few implementations become available soon.

In the ALTB project, about 180 questions have been represented as QTI 2 interactions. These questions and their XML – QTI 2 representations, can be viewed on the ALTB website (Hartog, 2005).

1.5.8 Training materials

The ALTB project has produced some instructional material for ETs and assistants of SMEs. This material can be found on the ALTB website (Hartog, 2005). In particular, a Blackboard course on the use of design patterns can be downloaded from the site. Furthermore, several chapters in this book are intended to be suitable as training material as well. In particular the chapters on design patterns (Chapter 6) and on task structures and resource allocations (Chapter 7) can be used directly for training of ET's, but also for training SME's and their assistants.

1.5.9 Further research

The ALTB project revealed a strong need for authoring and design support in the initial stages of design or in collaborative design in later stages. In Chapter 8, the functionality and workflow support that is really necessary is described. Currently such functionality and workflow support is lacking. The gap between what is really needed and what is available is huge. The task descriptions and suggestions for task allocation and communication in Chapter 7 are to a large extent based on systems that are currently in use.

The ALTB project has provided a wealth of experience. In evaluating the conclusions and results of the ALTB project, it is essential to keep in mind the scope of the project. On the one hand, this scope is wide in the sense that it covers design and development of closed questions in which the questions serve different roles. On the other hand, the scope is limited in several ways. First the scope is limited to small and midsize projects in higher education. Furthermore, the scope is limited to natural, engineering, and social sciences. Thus, fields like linguistics or literature were not within the scope of the ALTB project. Finally, the scope does not include many details of the actual construction of complete tests, execution of tests and analyzes.

1.5.10 The 'cluster of five approach'

For design and development of closed questions that are going to be used in assessment the 'cluster of five approach' was developed.

Obviously, questions designed and developed for the CBA role will have to be measurement instruments and thus a number of requirements related to measurement will apply. Furthermore, in projects for design and development of question pools for assessment, it makes sense to aim directly at equivalent sets of questions. A good approach is to design and develop four additional equivalent questions for this cluster, as soon as the first question has been validated. In this book, this approach is called the 'cluster of five' approach. The 'cluster of five approach' also implies that a project aiming at CBA, will have a minimum size that is considerably larger than a project aiming at the role of activating learning material.

The fact that the 'cluster of five approach' is important for design and development of questions for assessment but not relevant for design and development of questions for activating learning is a starting point in Chapter 5 and in Chapter 7.

1.5.11 Quality

Of course, costs for design and development projects should be linked to a well defined quality level for each relevant quality dimension. The general quality dimensions for the CBA role are validity, reliability and discrimination power of questions. For the ALM role, parameters such as the motivational value and specificity of feedback are of importance. However, many of such 'abstract' parameter values can only be estimated *ex ante*. Given the limitations in budget and student numbers of many projects in higher education these parameters are often not usable in practice. For a number of quality dimensions, there are no clear criteria in terms of a minimum level.

A fundamental problem is that in higher education, quality is primarily related to the extent to which the question – in view of the SME – operationalizes *understanding* of a concept, a procedure, a technique and so on and so forth.

In the ALTB project, explicit quality levels were therefore not defined from the start. As said, this is quite normal in higher education and is therefore an aspect of a methodology for innovative design and development of questions. As a result however, every teacher and educational developer in higher education has his/her own quality standards. For the 'two hours per question' conclusion, this means that the quality level of the questions is defined *implicitly* by the validators in the subprojects through statements such as 'this question is good'. This can be viewed as 'defining quality by example'. In particular, with respect to the mapping of a learning objective to a question, this is a workable option.

Defining quality of innovative closed questions in higher education is a challenge for future research.

1.5.12 Design and Development of closed questions in competency based education

Many institutions in higher education have adopted to some extent a competency based education philosophy. The practice in higher education is that assessment of competencies using closed questions in the initial years of curriculum, is a real challenge for which few satisfactory solutions have been published.

In the ALTB project, it was assumed that linking competency directed education with closed questions can best be approached by developing cases and integrating closed questions in these cases. Information is more meaningful and can be retrieved easier when – in a learning situation – it is presented or embedded in real life professional situations (e.g. Merriënboer et al., 2002). Based on that idea, the professional situation (case) of a graduated professional in a specific domain could be the basis of these questions. For the ALM role, already some successes of such an approach were reported (Aegerter-Wilmsen et al., 2003; Aegerter-Wilmsen et al., 2005; Schaaf et al., 2003; Schaaf et al., 2006). One sub project in ALTB, focused explicitly on competency based education and developed a case, based on the operation of a swimming pool. For this project it turned out that the actual questions could easily be mapped on traditional detailed learning objectives. In two other sub projects in ALTB, cases were used as a foundation for sets of closed questions. In these two other projects the philosophy of competency directed education was less explicit.

All in all, it is now concluded that, given the current status of adoption of competency directed education in higher education and the limitations of current LMSs and systems for CBA, it is difficult to design and develop closed questions that really support the competency directed approach. Chapter 8 shows how QT12 enables the representation of cases. Such cases could be used for competency assessment.

1.6 Methodologies are never complete

The most basic influence leading to changes in any is of course the influence of growing insight. With respect to design and development of digital closed questions, contextual changes are drivers for new insights and aspects. Contextual changes are for example developments with respect to standards, specifications and reference models such as QTI (IMS, 2005) and SCORM (ADL, 2006), but also the influence of de facto standards such as widely used learning management environments and systems for computer-based assessment.

At the start of the ALTB project, the ALTB team assumed that there would be many building blocks for a methodology in literature. The team was surprised how little could be found in standard approaches for instructional design or literature on item writing. With respect to assessment, Anderson et al. (2001, p. 298) wrote : "*Forty Four years after publication of the handbook [...] we could add little that would show any advance in item writing*". Question design and development does call for a specific methodology that takes into account both the ideal design and development strategies, whilst also recognising and addressing approaches for the barriers and limitations that are encountered in actual situations in higher education. Given the evolutionary change in higher education in both content, organization and technology, a methodology will always have to be adapted accordingly.

Table 1: Overview of intended audiences for each of the chapters in the book

	SME	ASME	ET	Researchers Assessment, Learning and Instruction	System Developer	University Administration
1. Intro	++	++	++	++	++	++
2. Concepts & taxonomy	+/-	+/-	++	++	++	-
3. Requirements & scoring	+	+	++	++	++	-
4. Guidelines	+/-	+/-	++	++	-	-
5. How to select adequate guidelines	+	+	++	+/-	-	-
6. Design Patterns	++	++	++	+	++	-
7. Scenarios, & budget templates	++	+	++	+	++	++
8. Computer Support	-	-	++	+	++	-

- ++ Directly useful
- + Indirectly useful
- +/- Maybe interesting
- Not useful

1.7 References

- ADL. (2006). SCORM. Retrieved march 14 2006, from <http://www.adlnet.gov/>
- Aegerter-Wilmsen, T., Bisseling, T., & Hartog, R. (2003). Web based Learning Support for Experimental Design in Molecular Biology: A Top-Down Approach. *Journal of Interactive Learning Research, 14*(3), 301-314.
- Aegerter-Wilmsen, T., Coppens, M., Janssen, F. J. J. M., Hartog, R., & T.Bisseling. (2005). Digital learning material for student-directed model building in molecular biology. *Biochemistry and Molecular Biology Education, 33*, 325-329.
- Alexander, C. (1979). *The timeless way of building*. New York: Oxford University Press.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.
- Bull, J., & McKenna, C. (2001). *Blueprint for Computer-assisted Assessment*: RoutledgeFalmer.
- Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Amsterdam: Addison and Wesley Professional Computing Series.
- Gardner, M., Greeno, J. G., Reif, F., Schoenfeld, A. H., diSessa, A., & Stage, E. (1990). *Toward a scientific practice of science education*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Haladyna, T., M.,, Downing, S., M., & Rodriguez, M., C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. In *Applied Measurement in Education* (Vol. 15, pp. 309-334): Lawrence Erlbaum Associates, Inc.
- Hartog, R. (2005). ALTB website. Retrieved march 20 2007, from <http://fbt.wur.nl/altb/>
- Hartog, R. J. M. (2004). *Actief Leren Transparant Beoordelen*. Wageningen: Wageningen University & CITO.
- IMS. (2005). Question and Test Interoperability. Retrieved august 25 2006, from <http://www.imsglobal.org/question/index.html#version2.0>
- IMS. (2006). QTI : IMS Question and Test Interoperability Specification Version 2.1 - Public Draft Specification Version 2. Retrieved feb 22 2007, from <http://www.imsglobal.org/question/>
- Merriënboer, J. J. G. v., Clark, R. E., & Croock, M. B. M. d. (2002). Blueprints for Complex Learning: The 4C/ID-Model *Educational Technology Research and Development, 50*(2), 39-64.
- Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-Based Testing, Building the Foundation for Future Assessments*. London: Lawrence Erlbaum Associates.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York,: Springer-Verlag.
- Schaaf, H. v. d., Vermue, M., Tramper, J., & Hartog, R. (2003). A design environment for downstream processes for Bioprocess-Engineering students. *European Journal of Engineering Education, 28*(4), 507-521.
- Schaaf, H. v. d., Vermuë, M., Tramper, J., & Hartog, R. (2006). Support of Modelling in Process-Engineering Education. *Computer Applications in Engineering Education, 14*(3), 161 - 168.

2 A Response-Based Taxonomy of Closed Questions

Ignace Latour
Cito

Rob Hartog
Wageningen University

Abstract

In the ALTB project a Conceptual Question Framework (CQF) has been developed for describing closed question types. The framework ultimately supports question design teams in thinking and communicating about goals and content of questions without being distracted by presentation and interaction possibilities and limitations of currently available CBA systems. Furthermore, the framework supports decision making with respect to design and development of future CBA systems. Finally, the framework conceptualizes a set of newly to be developed question types. A response-based taxonomy of closed questions, enclosing both classical and innovative question types, is presented and illustrated with examples.

The need for a conceptual framework was revealed by discussions in the case studies of the ALTB project. These discussions could not be resolved based on existing literature. After years of developments in Computer-Based Assessment (CBA) a consistent set of names and definitions for closed question types (selected-response question types) is still lacking. In different systems for CBA and also in literature the same terms may have different meanings. Furthermore, names and definitions of closed question types often focus on non-essential differences between question types. This obscures the more basic characteristics that question types have in common.

2.1 Introduction

The ALTB project has produced an abstract conceptual framework for describing closed questions. The need for such a framework became apparent in discussions on the selection of question types in several of the case studies of the ALTB project. Educational or instructional technologists in Higher Education are often involved in relatively small scale projects of designing question pools. They encounter problems such as how to advise subject matter experts about making a selection between a hotspot question and a multiple-choice question when in the available system the hotspot question is at the conceptual level actually a multiple-choice question. This chapter will help educational technologists who take part in projects for the design and development of closed questions to structure discussions on the selection of specific closed question types that are supported by the available system for computer-based testing or the available learning management system.

Question authoring or designing is the actual mental creation of question ideas and elaboration of these ideas to the extent that they can be technically realized. Authors (designers) should at the design level not be constrained unnecessarily by practical limitations of the available system. Nor should their attention be misdirected by inadequate terminology. In many design and development contexts, it makes sense to allocate the actual technical realization of a question in the CBA system or in the LMS to someone who can routinely carry out this technical realization. In these contexts question authors (subject matter experts - professors and lecturers) should be able to delegate everything that is not directly subject matter related to someone who implements the question in the CBA system or LMS. Therefore, the framework developed in the ALTB project helps to postpone decision making with respect to question types that are actually available to later stages in the design process and to delegate this decision making to someone who is very proficient with the available systems.

This chapter also aims at researchers in the field of assessment because it highlights ambiguities in existing question typologies and provides an alternative that is also extendible in a natural way. For the same reasons this chapter is also relevant for system developers. The design and development of systems for CBA should be based on a conceptual question framework that is extendible along logical lines. The basic architecture should match a framework that allows separation of the interaction dimension, the presentation dimension, the scoring dimension and the conceptual dimension. The concepts and taxonomy developed in the ALTB project should support developers of new CBA systems and LMS. For this reason the chapter also includes a UML class diagram that represents the question structure.

While the chapter is relevant for subject matter experts the chapter is not primarily aimed at subject matter experts. In teams that design and develop digital closed questions it will be the task of the educational technologist to apply the concepts and the taxonomy 'just in time', i.e. when the need for clear concepts becomes apparent in the team.

2.2 Towards requirements for a conceptual framework

In the last decennia a number of closed question types have been developed, based on new possibilities of information and communication technology (ICT). In literature these question types are called 'innovative' (Parshall et al., 2002). In the ALTB project a closed question is defined as a question where the required response should be based on options or value ranges offered to the respondent. In general closed questions can be evaluated automatically by comparing selected values with intended values. Modern systems for computer-based assessment (CBA) such as Question Mark Perception (QMP) (Questionmark, 2002) as well as assessment modules that are part of learning management systems (LMSs) such as Blackboard (Blackboard, 2006) or Moodle (Moodle, 2006) provide a range of different closed question types.

None of the currently available typologies of questions can be mapped onto a single taxonomy. In currently available typologies many of the innovative question types have names that primarily reveal how the question will be presented on a computer screen, or what type of action is expected from the respondent. Furthermore, the operational semantics of these names are not uniquely defined. For instance the name 'HotSpot question' is being used for questions that request the respondent to point out a specific point in a picture, but also for questions that present several areas in a picture as options. The latter is essentially a graphic version of the multiple-choice question. Likewise, a 'Fill-in-the-blank question' may or may not be a closed question.

Zenisky and Sireci (2002) present a list of question types but the definitions of the question types are based on a mixture of characteristics such interaction (for instance 'drag-and-drop'), domain (for instance 'mathematical expressions'), skills (for instance 'analyzing situations'), response format (for instance 'Essay/Short answer') and response constraints (for instance 'Multiple selection'). They describe a selection of innovative formats of selected response item types but a systematic hierarchy is missing. They discuss the role of the response format and conclude: "... with the advent of certain new formats that involve skills such as ordering information or classifying objects according to some defined dimension, the line between how much is selected and to what extent examinees are generating their own responses becomes blurred." In this chapter it will be argued that these types are closed questions and should also be described in a framework.

Parshall (2002) presents a typology based on five innovative dimensions: Item format, Response Action, Media Inclusion, Level of Interactivity and Scoring Method. While such a typology is useful in order to focus on what is new in comparison to a recent past, the level of detail of the specification of selected response questions is too low.

Scalise and Gilford (2006) present an 'intermediate constraint' question typology where dimensions of classification are the level of response constraints, leading to seven categories, and the 'innovation complexity' leading to four 'iconic' item types per category. Because this typology intends to provide a practical resource for assessment developers on "intermediate constraint" questions, the classification does not need sound and unambiguous criteria.

All in all no framework for systematically describing closed questions that is satisfactory in terms of coherence, terminology, response structure and extendibility of the set of question types has been found.

At a more technical level a typology that is based on the way in which the respondent and the system should interact is defined in the QTI2 specification (IMS, 2005). The QTI2 specification defines 'interaction classes'. One criterion for distinguishing interaction classes is in particular the distinction between in-line and graphic interaction. The QTI2 typology is defined in a technical format in order to prescribe those elements and relationships, which are needed in the transfer of test information between CBA systems. Although the structure of the interaction classes is extensive, the resulting model is too technical and specific to use as a conceptual framework in the day-to-day practice of question design by question authors and designers.

This chapter is based on the experience that question authors and designers need a taxonomy of question types and a corresponding conceptual framework based on the selection mechanism. A first requirement for such a taxonomy is that a designer should not be forced to make decisions on the type of interaction and the type of presentation options during initial design stages. A second requirement is that these latter two decisions can be delegated to people with another expertise, for instance to an educational technologist or an employee specialized in question entry. The third requirement is that the conceptual framework supports the assessment community and in particular developers of future CBA systems with the development of support for new question types by defining placeholders for new closed question formats in the taxonomy.

2.3 Description of the Conceptual Question Framework (CQF)

In this section the Conceptual Question Framework (CQF) that has been developed in the ALTB project will be described in detail. On the one hand there is a need for a formal language to describe the basis of closed questions, on the other hand the language should be used as a vehicle for communication between authors, designers and staff with more technical background. Obvious handles for core concepts are the structure of the requested response and the type and number of the value domains that are made available in the question. In this section, the CQF is represented in a set of definitions, a question structure diagram and a taxonomy of closed questions.

2.3.1 Definitions

The structural concepts needed to describe questions are introduced and defined. The main concepts are : Response, Finding, Fact, Value, Option, Position and Domain.

A closed question asks for a Response, based on one or more sets of Values , presented as Option Domains or , where only ranges are offered, Position Domains. A Response may be Multiple where each answer is to be evaluated independently. (e.g. *select capitals from a list of cities*). Any independent answer (e.g. *'Rome' or 'Paris'*) is called a Finding. Findings may contain one (like *'Rome'*) or more elements where each element is an understandable Fact. When there are more Facts expected in a Finding the relation between these Facts may be Unordered (like *'pen' AND 'paper' to write a letter*) or Ordered (e.g. *'millimeter', 'centimeter', 'decimeter' as units of increasing size*). Facts may be Simple (like *'pen'*) or Composite, where a match of Values from different Option Domains (e.g. *countries and capitals*) has to be made (*'France' - 'Paris'*). Options offered in a Domain may or may not be selected more than once in the Response or an individual Finding, to be defined in Occurrence property of an Option Domain. When a selection of a Value has to be made by positioning between upper and lower limits of a range (like *the percentage of the unemployed in a population*) the Value type is named Position.

The concepts are defined in a more formal language in Table 2.

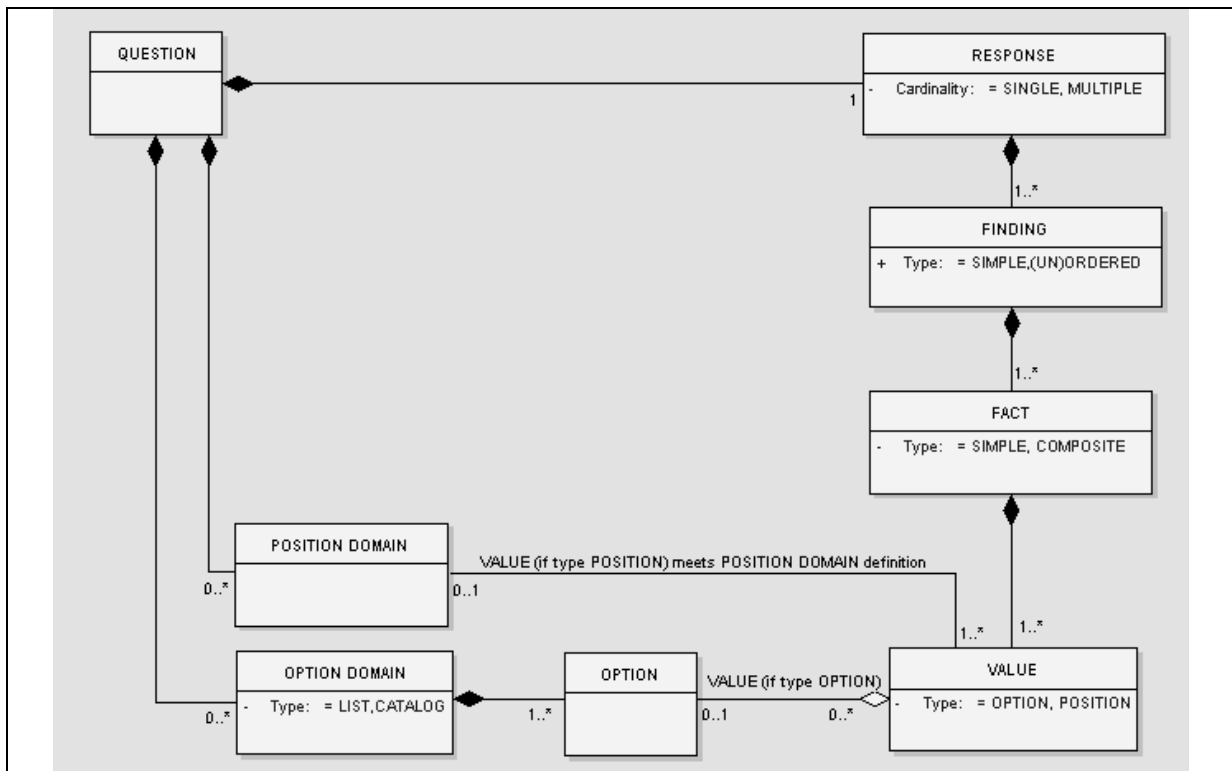
Table 2: Definitions of the basic concepts in the CQF

<i>TERM</i>	<i>DEFINITION</i>
RESPONSE	Response is a set of structured data (deliberately) submitted by the respondent as reaction to a question. A Response is a set of mutually independent Findings.
- SIMPLE Response	The cardinality of Findings in the Response is 1
- MULTIPLE Response	The cardinality of Findings in the Response is more than 1
FINDING	Finding, expressed by the respondent, is the smallest set of related Facts that can be evaluated as adequate reaction to a question.
- SIMPLE Finding	A Finding consisting of only one Fact is named Simple Finding
- UNORDERED Finding	A Finding consisting of more than one Fact is named Unordered Finding when the order of its Facts is irrelevant
- ORDERED Finding	A Finding consisting of more than one Fact is named Ordered Finding when the order of its Facts is relevant
FACT	Fact is the smallest component of a Finding that has a meaning for evaluation
- SIMPLE Fact	A Fact is named Simple when it may contain a Value from just one Domain
- COMPOSITE Fact	A Fact is named Composite when it contains a composition of Values from the different Domains.
VALUE	Value is the smallest entity a respondent can select
- OPTION Value	Option is an explicitly defined Value
- POSITION Value	Position is an implicitly (by boundary values) defined potential Value.
DOMAIN	Domain is a role-specific set of potentially selectable Values
- OPTION Domain	Option Domain is a domain consisting of explicitly named Options
- Small : LIST	List is an Option Domain containing a small set of Options.
- Large : CATALOG	Catalog is an Option Domain that is in practice (much) longer than List, sets like hundreds or thousands of Options may be defined.
- Option OCCURRENCE	In Response: Number of times any Option of a specific Option Domain may occur within a Response In Finding: Number of times any Option of a specific Option Domain may occur within any Finding of a Response Minimum and maximum numbers may be relevant, depending on the question type.
- POSITION Domain	Position Domain is a Domain consisting of implicitly defined potential Values. The ordered set of Positions in a Domain is implicitly defined by upper and lower boundaries

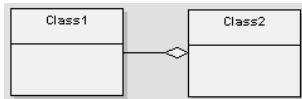
2.3.2 The question structure diagram

The CQF is primarily intended to support the process of designing closed questions by means of a hierarchy of closed questions that enables designers to separate decisions. However for the CQF to be used by designers it is desirable that design will be supported by an integrated design and development environment. Therefore, developers will at least have an object model view of the framework. Figure 1 presents the closed-question structure as a UML class diagram of the most important object types that are needed to support the CQF. A detailed definition of the classes does not fit the scope of this chapter.

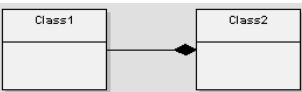
Figure 1: Question Structure Diagram as an UML class diagram



Key to the specific relations:



An *aggregation* is used to depict an element(Class2) which is made up of other component(s)(Class1). Deletion of the aggregation (Class2) does not affect the existence of a component (Class1)



A *composite aggregation* is used to depict an element (Class2) which is made up of other component(s)(Class1). If a composition is deleted, all of its parts are deleted with it; however a part can be individually removed from a composition without having to delete the entire composition

2.3.3 Taxonomy of closed questions

Based on the concepts discussed a taxonomy is to be designed. The first step towards a taxonomy of closed questions is the construction of the Finding, the second step is the construction of the Fact. Next the number and type of Domains creates variety. For the formats in the taxonomy one can formulate the question in a generic manner by the instruction "How to Respond a Finding". For the notation of the Finding a set of elements is proposed.

Executing the first three steps results in the taxonomy of closed questions with a number of domains that has been limited to two results in a taxonomy as presented in Table 3. Currently, the table contains fifteen question types. Elaborations of the taxonomy are an extension based on a larger number of value domains, a distinction between one or more Findings in the Response (Simple, Multiple) and the differentiation of (min. and max.) Occurrence of Options in Response and Finding. This further differentiation is illustrated by examples in the next section.

Table 3: Taxonomy of Closed Questions (limited)

Finding type	Fact type	Domain type(s)	How to Respond a Finding :	Notation of Finding
SIMPLE	SIMPLE	OPTION	State a Fact by selecting an Option	[(o)]
	COMPOSITE	OPTION, OPTION	State a Fact by matching an Option from both Option Domains respectively	[(o,o)]
	SIMPLE	POSITION	State a Fact by selecting a Position	[(p)]
	COMPOSITE	POSITION, POSITION	State a Fact by matching a Position from both Position Domains respectively	[(p,p)]
	COMPOSITE	POSITION, OPTION	State a Fact by matching a Position with an Option	[(p,o)]
UNORDERED	SIMPLE	OPTION	Combine Facts, state a Fact by selecting an Option	[(o) , (o) , (o)]
	COMPOSITE	OPTION, OPTION	Combine Facts, state a Fact by matching an Option from both Option Domains respectively	[(o,o) , (o,o) , (o,o)]
	SIMPLE	POSITION	Combine Facts, state a Fact by selecting a Position	[(p) , (p) , (p)]
	COMPOSITE	POSITION, POSITION	Combine Facts, state a Fact by matching a Position from both Position Domains respectively	[(p,p) , (p,p) , (p,p)]
	COMPOSITE	POSITION, OPTION	Combine Facts, state a Fact by matching a Position with an Option	[(p,o) , (p,o) , (p,o)]
ORDERED	SIMPLE	OPTION	Arrange Facts, state a Fact by selecting an Option	[(o) → (o) → (o)]
	COMPOSITE	OPTION, OPTION	Arrange Facts, state a Fact by matching an Option from both Option Domains respectively	[(o,o) → (o,o) → (o,o)]
	SIMPLE	POSITION	Arrange Facts, state a Fact by selecting a Position	[(p) → (p) → (p)]
	COMPOSITE	POSITION, POSITION	Arrange Facts, state a Fact by matching a Position from both Position Domains respectively	[(p,p) → (p,p) → (p,p)]
	COMPOSITE	POSITION, OPTION	Arrange Facts, state a Fact by matching a Position with an Option	[(p,o) → (p,o) → (p,o)]

Explanation:

- o = Option Value
- p = Position Value
- (...) = Fact
- [...] = Finding
- [(...), (...)] = unordered Finding
- [(...) → (...)] = ordered Finding

2.4 Application of the framework

2.4.1 Some examples

To show the added value of the CQF and its taxonomy of closed questions this section presents a few examples. These examples are meant to illustrate the descriptive power of the CQF and the added value of the taxonomy and not to illustrate how a question which satisfies all possible design requirements should be phrased and presented. The latter would require more context information for the reader and would also distract the attention from the response based structure. On the other hand, the examples are, with exception of the first two examples, based on questions that are really in use in higher education. Note furthermore that, except for example 4, the questions are of types that are not yet supported by available CBA systems or LMSs. This implies that, insofar they are in use in higher education, they are realized in dedicated applications.

Example 1 - Unordered set of Options

This example shows an extension of a common 'Multiple Response' by requesting more than one Unordered Finding.

There are two Findings requested, any Finding consists of an Unordered set of Options. Options may be selected in both Findings.

The question is an artificial example especially constructed for the purpose of this illustration. The necessity to extend the 'Multiple Response' question will become even more apparent in the next chapter, and illustrated with more realistic examples.

Type in the taxonomy:

Finding type	Fact type	Domain type(s)	How to Respond a Finding :	Notation of Finding
UNORDERED	SIMPLE	OPTION	Combine Facts, state a Fact by selecting a Option	[(o) , (o) , (o)]

Figure 2 (a & b): Question example 1 (left) and CQF concepts based on the structure of the response (right)xx

<p>Describe two methods to produce a letter to put in an envelope by selecting the needed attributes. For each method 2 or 3 attributes are needed. Attributes may be used in both methods.</p> <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 10px;"> <thead> <tr> <th style="width: 20%;"></th> <th style="width: 30%; text-align: center;"><i>Method</i></th> <th style="width: 30%; text-align: center;"><i>Method</i></th> </tr> </thead> <tbody> <tr> <td><i>Pen</i></td> <td style="text-align: center;"><input checked="" type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> <tr> <td><i>Printer</i></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input checked="" type="checkbox"/></td> </tr> <tr> <td><i>Paper</i></td> <td style="text-align: center;"><input checked="" type="checkbox"/></td> <td style="text-align: center;"><input checked="" type="checkbox"/></td> </tr> <tr> <td><i>Computer</i></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input checked="" type="checkbox"/></td> </tr> <tr> <td><i>Telephone</i></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </tbody> </table>		<i>Method</i>	<i>Method</i>	<i>Pen</i>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<i>Printer</i>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<i>Paper</i>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<i>Computer</i>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<i>Telephone</i>	<input type="checkbox"/>	<input type="checkbox"/>	<p>CQF concepts</p>
	<i>Method</i>	<i>Method</i>																	
<i>Pen</i>	<input checked="" type="checkbox"/>	<input type="checkbox"/>																	
<i>Printer</i>	<input type="checkbox"/>	<input checked="" type="checkbox"/>																	
<i>Paper</i>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>																	
<i>Computer</i>	<input type="checkbox"/>	<input checked="" type="checkbox"/>																	
<i>Telephone</i>	<input type="checkbox"/>	<input type="checkbox"/>																	

Table 4: Question example 1

'Design Pattern'	<i>Attributes of a Preparation method</i>	
Introduction of the question	<i>To produce a letter (to put in an envelope) there are different methods, any method is supported by the use of attributes. To let the respondent show his ability to prepare he is asked to explicitly select the needed attributes for a specific preparation method. There are two methods. Instruction to the respondent : Choose 2 to 3 attributes you need for a method to prepare a letter, respond in relation to two different methods</i>	
Description	In CQF	In Example
Response	<i>Multiple : max 2 Findings</i>	<i>Max 2 combinations of attributes</i>
Finding	<i>Unordered : max 3 Facts</i>	<i>Max 3 attributes in a combination</i>
Fact	<i>Simple</i>	<i>An attribute is a single option</i>
Domain	<i>Option Max. occurrence in Response: free Max. occurrence in Finding : 1</i>	<i>Letter attributes (options) pen printer paper computer telephone</i>
How to respond a Finding	<i>Combine Facts, state a Fact by selecting a Option</i>	<i>Combine (2-3) attributes, choose a attribute by selecting an option from the list</i>


Example 2 - Catalog

This example shows an extension of an option list by offering a Catalog Domain. Three Findings are requested, an extended index of options is offered. The question is an artificial example especially constructed for the purpose of this illustration.


Type in the taxonomy:


Finding type	Fact type	Domain type(s)	How to Respond a Finding :	Notation of Finding
SIMPLE	SIMPLE	OPTION	State a Fact by selecting an Option	[(o)]


Figure 3: Question example 2

Read the medical case description carefully and determine possible diagnoses. Select a disease from the Alphabetic List by a click on the disease and a click on the corresponding button 

Diagnosis







Alphabetic List of Specific Diseases/Disorders

[Start Page](#)

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

- [A-alphaipoprotein Neuropathy \(see Tangier Disease\)](#)
- [Abdominal Cramps \(see Colic\)](#)
- [Abdominal Delivery \(see Cesarean Section\)](#)
- [Abdominal Injuries](#)
- [Abdominal Pain](#)
- [Abnormalities](#)
- [Abortion, Induced](#)
- [Abortion, Spontaneous](#)
- [Abscess](#)
- [Abscess, Amebic \(see Amebiasis\)](#)
- [Abscess, Pulmonary \(see Lung Abscess\)](#)
- [Abscess, Retropharyngeal \(see Retropharyngeal Abscess\)](#)
- [Acantholysis Bullosa \(see Epidermolysis Bullosa\)](#)

(List retrieved from:: <http://www.mic.ki.se>)

Table 5: Question example 2

"Design Pattern"	<i>Diagnosis selection from an index</i>	
Introduction of the question	<i>Based on the case description(to be delivered) let the respondent choose the possible diagnoses from the alphabetical index to diseases Instruction to the respondent : Choose a disease as a possible diagnosis, respond with three possible diagnoses</i>	
Description	In CQF	In example
Response	<i>Multiple : max 3 Findings</i>	<i>Max 3 possible diagnoses</i>
Finding	<i>Simple : max 1 Fact</i>	<i>Max 1 disease per diagnosis</i>
Fact	<i>Simple</i>	<i>A disease is a single option from the list</i>
Domain	<i>Option Max. occurrence in Response: 1 Max. occurrence in Finding : 1</i>	<i>Alphabetical index to diseases</i>
How to respond a Finding	<i>State a Fact by selecting a Option</i>	<i>Choose a disease by selecting an option from the alphabetic list</i>

Example 3 A - set of Option Domains

This example shows an extension of a common 'Matching' question by offering five Option Domains. Only one Finding is requested, based on only one Fact composed of 5 values, each from a different Domain.

The question is adapted from (Busstra, 2007). The example also shows how questions with more than two Domains are natural when learning goals involve skills to design experiments or other design challenges such as the design of production facilities. In many design challenges one often needs a set of attributes or decisions, each from another collection or aspect. This example illustrates furthermore how question types that cannot yet be described in existing question typologies can be described in the CQF.

Type in the taxonomy:

Finding type	Fact type	Domain type(s)	How to Respond a Finding :	Notation of Finding
SIMPLE	COMPOSITE	OPTION(5x)	State a Fact by matching an Option from all Option Domains respectively	[(o,o,o,o,o)]

Figure 4: Question example 3

Table 6: Question example 3

"Design Pattern"	<i>Possible design by matching options from more than 2 domains</i>	
Introduction of the question	<p><i>To design an experiment choices have to be made. A match of options from five domains defines the experiment design.</i></p> <p><i>Instruction to the respondent:</i></p> <p><i>Design the most useful experiment to answer your research question by choosing one Study object, one Type object, one Treatment, one Measurement and one Technique.</i></p>	
Description	In CQF	In example
Response	<i>Multiple : max 1 Finding</i>	<i>Max 1 experiment design</i>
Finding	<i>Simple : max 1 Fact</i>	<i>Max 1 match per experiment design</i>
Fact	<i>Composite</i>	<i>A solution matches one Study object, one Type object, one Treatment, one Measurement and one Technique.</i>
Domain (1)	<i>Option</i> <i>Max. occurrence in Response: 1</i> <i>Max. occurrence in Finding : 1</i>	<i>Study object</i> <i>Human, Cells, Mice</i>
Domain (2)	<i>Option</i> <i>Max. occurrence in Response: free</i> <i>Max. occurrence in Finding : 1</i>	<i>Type object</i> <i>FAR, PPARalpha, HNF4alpha, SREBP1c</i>
Domain (3)	<i>Option</i> <i>Max. occurrence in Response: free</i> <i>Max. occurrence in Finding : 1</i>	<i>Treatment</i> <i>high/low, fat diet, fasting,fatty acid-injection,</i> <i>obese/non-obese comparison</i>
Domain (4)	<i>Option</i> <i>Max. occurrence in Response: free</i> <i>Max. occurrence in Finding : 1</i>	<i>Measurement</i> <i>proteome, transcriptome,metabolome</i>
Domain (5)	<i>Option</i> <i>Max. occurrence in Response: free</i> <i>Max. occurrence in Finding : 1</i>	<i>Technique</i> <i>anti-body array, 2D-gels</i>
How to respond a Finding	<i>State a Fact by matching an Option from all five Option Domains respectively</i>	<i>Choose an option from all five experiment design aspects</i>

Example 4 - Option Occurrence restricted to 1

This example shows how a common 'Matrix' question is built on two Option Domains and can be restricted by Option Occurrence. A single occurrence of any Option from both Option Domains is compulsory in the response. The Option Occurrence in both Option Domains is 1 (both max. and min.). The question is produced in the ALTB project.

Type in the taxonomy:

Finding type	Fact type	Domain type(s)	How to Respond a Finding :	Notation of Finding
SIMPLE	COMPOSITE	OPTION, OPTION	State a Fact by matching an Option from both Option Domains respectively	[(o,o)]

Figure 5: Question example 4

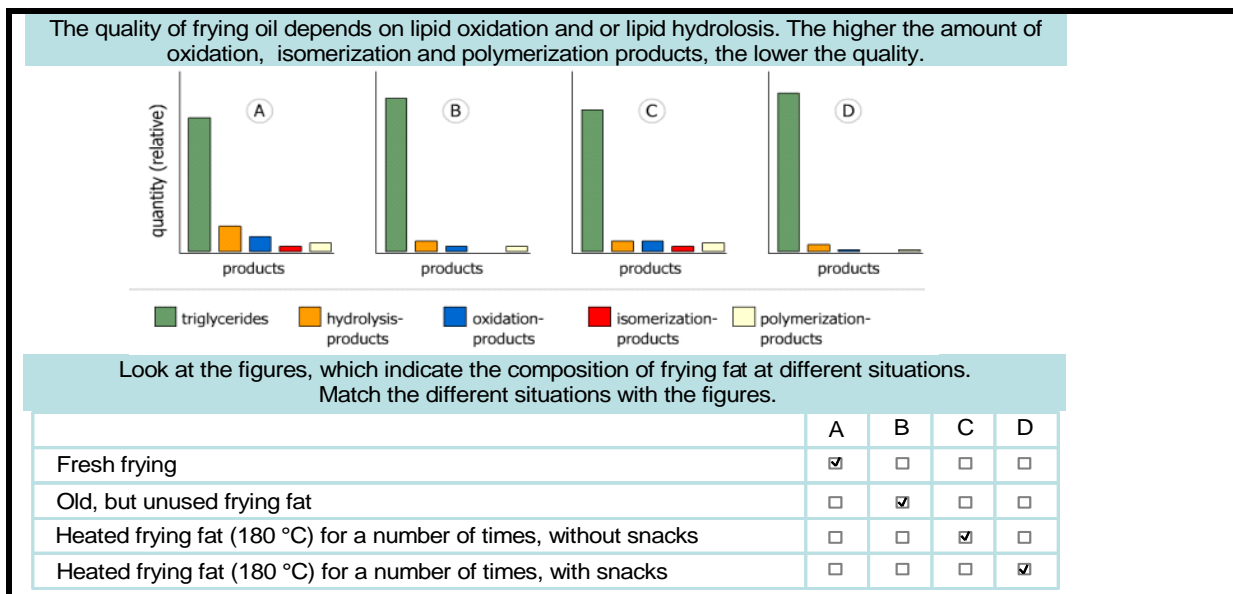


Table 7: Question example 4

"Design Pattern"	<i>Situation and consequence</i>	
Introduction of the question	<i>The quality of frying fat is determined by the fat composition and depends on the frying fat use. Instruction to the respondent : Match any fat composition with any fat use situation</i>	
Description	In CQF	In Example
Response	<i>Multiple: max 4 Findings</i>	<i>Max 4 independent fat composition – fat use relations are to be responded</i>
Finding	<i>Simple: max 1 Fact</i>	<i>Max 1 match of options in a independent fat composition – fat use relation</i>
Fact	<i>Composite</i>	<i>A match is a composite of a fat composition option and a fat use option</i>
Domain(1)	<i>Option Max. occurrence in Response: 1 Min. occurrence in Response: 1 Max. occurrence in Finding: 1</i>	<i>Situation of fat use Fresh frying fat, Old but unused frying fat, Heated frying fat for a number of times (180° Celsius) without snacks, Heated frying fat for a number of times (180° Celsius) with snacks</i>
Domain(2)	<i>Option Max. occurrence in Response: 1 Min. occurrence in Response: 1 Max. occurrence in Finding: 1</i>	<i>Consequence in fat composition - composition diagram 1 to 4</i>
How to respond a Finding	<i>State a Fact by matching an Option from both Option Domains respectively</i>	<i>Indicate a true fat composition – fat use relation</i>

Example 5 - Position Domain

This example shows how a set of Position Domains can be used in a 'Matching' question. Three independent Position Domains are offered. Only one Finding is requested, based on only one Fact composed of 3 Position Values, each from a different Domain. The question is adapted from (Busstra, 2007).

Type in the taxonomy:

Finding type	Fact type	Domain type(s)	How to Respond a Finding :	Notation of Finding
SIMPLE	COMPOSITE	POSITION (3x)	State a Fact by matching a Position from all Position Domains respectively	[(p,p,p)]

Figure 6: Question example 5 (adapted from (Busstra, 2007))

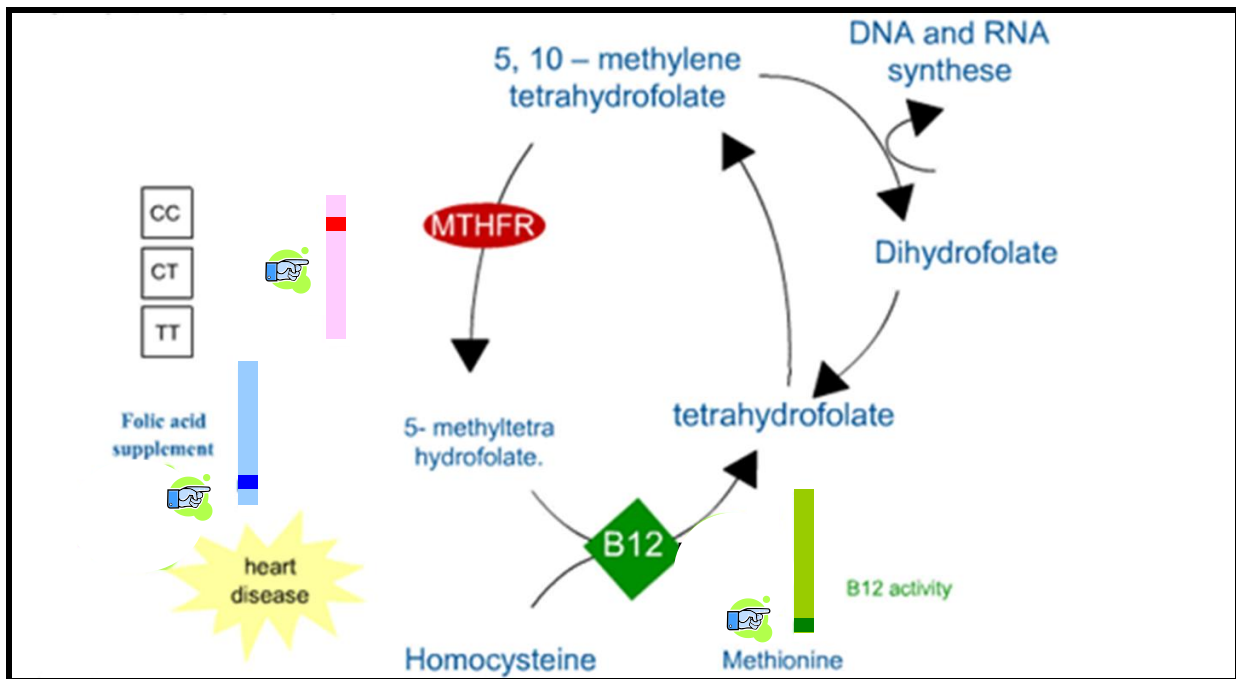


Table 8: Question example 5

“Design Pattern”	<i>The influence of process variables</i>	
Introduction of the question	<p><i>The question is presented here as support for learning. The presentation can also be used as a question to assess how well the respondent can adjust the supplements</i></p> <p><i>Possible Instruction to the respondent:</i> <i>Adjust the 3 sliders related to each other in the optimal positions, the scale is relative</i></p>	
Description	In CQF	In Example
Response	<i>Multiple: max 1 Finding</i>	<i>Max 1 optimal situation</i>
Finding	<i>Simple: max 1 Fact</i>	<i>Max 1 composition</i>
Fact	<i>Composite</i>	<i>The adjustment is a match of 3 related values</i>
Domain (1)	<i>Position</i>	<i>Relative quantity folic acid supplement(blue)</i>
Domain (2)	<i>Position</i>	<i>Relative B12 activity(green)</i>
Domain (3)	<i>Position</i>	<i>Relative concentration MTHFR(red)</i>
How to respond a Finding	<i>State a Fact by matching a Position from all three Position Domains respectively</i>	<i>Position all three sliders as a combined adjustment</i>

2.4.2 Positioning question types of QMP in CQF

For users of existing CBA systems the Conceptual Question Framework will be very different from what they are used to. In this section the relationships between the questions types as defined in QMP (Questionmark, 2002) and the question types as defined in the taxonomy of closed questions will be listed and explained. QMP is widely used and well known in higher education and provides an extensive typology. Because the CQF is based on logical distinctions between imaginable responses whereas question types in QMP are defined based on a mixture of screen presentation differences, interaction differences and other differences the relationships between question types in both frameworks are not straightforward. Table 9 shows the relation between the question types in the CQF and the question types in QMP

Table 9: Positioning QMP question types in the CQF

Taxonomy of Closed Questions			QMP "closed" question types												
Finding type	Fact type	Domain type(s)	Yes/No	Likert scale	Multiple-choice	Matching	Multiple response	Matrix	Numeric	Pull down	Ranking	Select a blank	True/False	Hotspot	Drag & Drop
SIMPLE	SIMPLE	OPTION	✓	✓	✓		✓			✓		✓	✓		
	COMPOSITE	OPTION, OPTION				✓		✓							
	SIMPLE	POSITION						✓							
	COMPOSITE	POSITION, POSITION												✓	✓
	COMPOSITE	POSITION, OPTION													
UNORDERED	SIMPLE	OPTION					✓								
	COMPOSITE	OPTION, OPTION													
	SIMPLE	POSITION													
	COMPOSITE	POSITION, POSITION													
	COMPOSITE	POSITION, OPTION													
ORDERED	SIMPLE	OPTION									✓				
	COMPOSITE	OPTION, OPTION													
	SIMPLE	POSITION													
	COMPOSITE	POSITION, POSITION													
	COMPOSITE	POSITION, OPTION													

Question types of the CQF taxonomy are represented (cell with) by one or more QMP item types or not represented at all (dark cells).

The comparison in Table 9 highlights the insights, which the CQF provides:

- From 15 proposed question types, 6 types (see with) are represented in QMP. When there is a need for one of the other nine, the question has to be simplified to a more implicit question, i.e. the team has to compromise. Application of the directives about postponing design decisions and delegating design decisions described above implies that a question design team should focus on the left three columns in the initial design stage.
- Most of the QMP question types are a representation of the basic CQF question type with one Fact in a Finding and only one Option Domain (the first type in Table 9). The aspects that differentiate between all these QMP types are the number of Findings (Multiple Response more than one Finding) the explicit Options (True/False, Yes/No, Likert) or the interaction (Select a Blank and Pull down).

- With one exception, each QMP closed question type is related to only one of the types in the CQF taxonomy. The exception is the QMP 'Multiple Response' question. The QMP 'Multiple Response' is not explicit in whether it has a Response with more Findings or only one (Unordered) Finding with more Facts.
- The set of question types as distinguished in QMP is actually projected on the response dimensions. This shows that – from the viewpoint of the response structure – several of these question types are essentially the same. This is important for the definition of scoring rules for these question types.

2.4.3 Positioning QTI2 interaction types in CQF

The IMS Global Learning Consortium provides an extensive and detailed specification for question and test interoperability usually referred to as QTI. Core of QTI2 is constituted by interaction types and not by question types. One item can contain more than one interaction. To make these interactions recognizable the classification is also based on the way the interaction is implemented, like graphical, inline or isolated, In contrast for the taxonomy of closed question in CQF the approach is just focussed on the response structure. The mapping of the closed-question related interaction types of the QTI2 on the CQF taxonomy is presented in Table 10.

Table 10: Mapping of QTI2 interaction types onto the CQF

Taxonomy of Closed Questions			IMS QTI 2 "Interaction types"													
Finding type	Fact type	Domain type(s)	choiceInteraction	orderInteraction	associateInteraction	matchInteraction	gapMatchInteraction	inlineChoiceInteraction	hottextInteraction	hotspotInteraction	selectPointInteraction	graphicOrderInteraction	graph.AssociateInteraction	graph.GapMatchInteraction	positionObject	sliderInteraction
SIMPLE	SIMPLE	OPTION	✓					✓	✓	✓						
	COMPOSITE	OPTION, OPTION				✓	✓								✓	
	SIMPLE	POSITION														✓
	COMPOSITE	POSITION, POSITION									✓				✓	
	COMPOSITE	POSITION, OPTION														
UNORDERED	SIMPLE	OPTION			✓								✓			
	COMPOSITE	OPTION, OPTION														
	SIMPLE	POSITION														
	COMPOSITE	POSITION, POSITION														
	COMPOSITE	POSITION, OPTION														
ORDERED	SIMPLE	OPTION		✓									✓			
	COMPOSITE	OPTION, OPTION														
	SIMPLE	POSITION														
	COMPOSITE	POSITION, POSITION														
	COMPOSITE	POSITION, OPTION														

Six question types of the CQF taxonomy can be matched directly with one or more QTI2 interaction types (cell with ✓). The dark cells indicate for which CQF question types no direct match with a QTI2 interaction type can be made

Table 10 highlights that:

- Any QTI2 interaction type is related to just one of the types in the CQF taxonomy (see). This means that the distinction of interaction types in QTI2 is unambiguous in relation to the CQF concepts.
- From 15 proposed question types, 6 types can be directly represented by means of just one interaction from the QTI2 set of interactions. The question if and how each of the other nine types can be realized in conformance with the QTI2 specification does not fit the scope of this chapter.
- In comparison with QMP the QTI2 interaction types are more equally distributed. The main distinction between QTI2 interaction types of the same CQF type is whether the implementation of the interaction is graphical.
- The set of interaction types as distinguished in QTI2 is actually projected on the response dimensions. This shows that – from the viewpoint of the response structure – several of these question types are essentially the same. This is important for the definition of scoring rules for questions based on these interaction types.

2.5 Conclusions and summary

Within the ALTB project a Conceptual Question Framework (CQF) has been developed, based on an analysis of closed questions and typologies for closed question types. The response structure dimension has been isolated from other descriptive dimensions like the question presentation dimension and the user interaction dimension. As a result a set of concepts and a taxonomy of closed questions has been presented that is based on the response structure of the question. The concepts and their relationships are described in the (CQF). The use of the concepts has been illustrated with examples.

It has been shown that the Closed-Question Taxonomy can be mapped on existing typologies QMP4 and QTI2, such that each question type of existing typology can be related to at least one of the CQF types. The CQF closed-question taxonomy proposes also question types which are not yet implemented.

Further research should aim to extend the CQF by relating scoring models, presentation models and interaction models. It is likely that a system design based on the CQF classes and relationships will require little effort to realize interoperability, based on QTI2 or any other interoperability reference model.

Mapping question types that are supported by available LMSs and systems for CBA onto the CQF taxonomy shows that many question types as defined in the CQF have not yet a counterpart in existing systems, though some of these question types are already in use in dedicated applications. Further research is needed to assess the benefits of implementing these not-yet-realized question types in an LMS or a system for CBA. Adoption of the CQF concepts in CBA systems will support this implementation. For this reason a new CBA system is now being developed that takes the CQF class diagram as a starting point.

Because the definitions of question types in the CQF do not yet determine the form of user interaction and the presentation of the question components, it will ultimately be possible to present the same question in different forms to different user groups. The CQF frees content developers from thinking in restricted, traditional ways and instead enables them new or innovative question formats.

2.6 References

Blackboard. (2006). Blackboard. Retrieved march 04 2006, from <http://www.blackboard.com/>

Busstra, M. C., Hartog, R., Kersten, S., & Müller., M. R. (2007). Design guidelines for the development of digital nutrigenomics learning material for heterogeneous target groups. . *Advances in Physiology Education*, 31.

IMS. (2005). Question and Test Interoperability. Retrieved august 25 2006, from <http://www.imsglobal.org/question/index.html#version2.0>

Moodle. (2006). Moodle. Retrieved march 04 2006, from <http://moodle.com>

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York,: Springer-Verlag.

Questionmark. (2002). Questionmark. Retrieved jan 19 2007, from <http://www.questionmark.com/>

Scalise, K., & Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment* 4(6).

Zenisky, A. L., & Sireci, S. G. (2002). Technological Innovations in Large-Scale Assessment. *Applied Measurement in Education*, 15(4), 337-362.

3 Multiple Response Questions in Computerized Testing

Mia van Boxel
Huub Verstralen
Cito

Abstract

This chapter describes design requirements and scoring rules for multiple response questions. In general, the same design requirements apply to multiple response questions and multiple-choice questions. However for multiple response questions additional requirements have been formulated in the ALTB project. These requirements are related to scoring. The conclusions are important for anyone who is involved in the design of questions that are basically multiple response questions or in test construction. However the line of reasoning requires background knowledge of the literature in computer-based testing.

Three different types of multiple response questions are distinguished:

1. Type 1 MR-multiple true/false items with independent correct options.
2. Type 2 MR-combination items with one correct subset of options.
3. Type 3 MR-multi-combination items with more correct subsets of options.

The chapter explains why items with one correct subset of options (type 2 MR-combination) should be scored dichotomously. When more than one subset is requested (type 3 MR-multi-combination) each correct subset is scored dichotomously, the sum of which gives a polytomous score for the item. Items with independent correct options (type 1 MR-multiple true/false) can also be scored with partial credit. For these items a rank order of the options as a response is introduced as an alternative. An analysis of ranks with the so-called Luce model offers substantial advantages in terms of measurement precision of the test result compared to scoring of selection of options. This chapter describes use of multiple response questions for some learning objectives that are common in higher education and illustrates this use with examples.

3.1 Introduction

The ALTB experience indicated that many learning objectives in higher education can be matched in a natural way to multiple response questions. Also certain types of multiple response items that will be introduced below, meet less resistance by subject matter experts than traditional multiple-choice items. The reason for this is that the chance to arrive at the correct answer by means of a pure guess in these types of multiple response questions is very low. At the same time the design and scoring of multiple response items raised many questions in the ALTB project which could not be answered by the literature. There are just a few publications where the multiple response item format is mentioned, but a more in depth treatment is lacking. Nevertheless, multiple response questions are frequently used in summative and formative testing. In response to the questions that arose from practical experience in question design the theory about multiple response items has been extended in the ALTB project.

This chapter will help item design teams and test construction teams with the design process and scoring of multiple response questions. In section 3.2 three different types of multiple response questions will be described. In section 3.3 three new design requirements for multiple response questions will be presented and discussed. In addition some examples are given of the use of multiple response questions in higher education. Section 4 focuses on scoring responses to MR questions, and how to convert these item scores into a test score. Two types of responses will be discussed: selection of one or more subsets of the options, and a rank order of the options. The efficiency of both approaches will be compared. Also the relative merits of the three MR item types are discussed.

3.2 Definition and types of multiple response questions

A multiple response (MR) question consists of a stem which poses the question and three or more options of which two or more options are correct.

The MR question type is also labeled in the literature (see for instance Haladyna, 2004) with the terms "multiple answer" or "multiple mark" or "multiple multiple-choice". The multiple response question type is often used in paper based and computer-based testing. The participant has to select two or more options as the correct answer. In this chapter the rank order of all the options is introduced as an alternative response. The rank order response is restricted to one type of MR items, the MR-multiple true false, to be explained in the sequel.

According to (Parshall et al., 2002) there are five dimensions to describe innovative digital question types: item format, response action, media inclusion, level of interactivity, scoring method. In every day practice those different dimensions are used to categorize different item formats. For the topic of this chapter not all these dimensions are considered relevant. Often item formats with distinct names are essentially "multiple response" questions (like multiple hotspot, some drag and drop formats) or "multiple-choice" questions (like hotspot with predefined spots).

Usually, multiple-choice questions are scored dichotomously. This means that a participant always gets a score 0 or 1. Multiple response questions are scored dichotomously or polytomously. It means that scores other than 0 and 1 are possible, like score 3 or -1. (Note that polytomous scoring and partial credit scoring are the same.)

The technical implementation of polytomous scoring methods in test software is often not very sophisticated. For a further discussion of scoring multiple response questions see section 3.4.

Three different types of multiple response questions with different consequences for scoring and response instruction are distinguished:

- Type 1: MR-multiple true/false
- Type 2: MR-Combination
- Type 3: MR-Multi-combination
 - Type 3a : single subset selection
 - Type 3b : multiple subset selection

Table 11 links these types to the ALTB Conceptual Question Framework that has been introduced in Chapter 2. However, in order to allow readers to read this Chapter 3 independently, this chapter uses the more traditional term 'Multiple Response'.

Table 11: Linking the MR question types to the concepts in the ALTB Conceptual Question Framework

MR itemtypes	1. MR-multiple true/false item	2. MR-combination item	3a. MR-multi-combination item type 3a	3b. MR-multi-combination item type 3b
CQF(*)				
Response	Multiple	Simple	Simple	Multiple
Finding	Simple	Unordered	Unordered	Unordered
Fact	Simple	Simple	Simple	Simple
Domain	Option	Option	Option	Option

Below, these types of MR-questions are discussed in more detail.

3.2.1 Type 1: A MR-multiple true/false has a subset of independent correct options

This MR-type is mostly specified in the literature with the term "multiple true false". Each individual option can be evaluated independently, whether it is a correct response to the posed question. So it is not the particular subset that is correct, but each of the individual options. According to Haladyna the efficiency of presenting many items in a short time is the main attraction of this item format.

Example 1: a MR-multiple true/false item

Select the medicines that can be prescribed as a remedy for a headache?

- Lithium
- Paracetamol
- Haldol
- Asperin
- Antibiotic

Two values are correct: paracetamol and aspirin. Paracetamol is a correct answer to the question, but it is not the only one. Aspirin is a correct answer as well.

One could also have asked for exactly two medicines. This fixes the number of to be selected options to two. Although this may change the response strategy of the participant, such an addition does not affect the independency of the individual options.

3.2.2 Type 2: A MR-combination has one correct subset of interdependent options

Only one particular subset of options is the correct answer. No option from this subset can be left out without compromising the correct answer.

Example 2: a MR-combination item

Which combination of 2 medicines gives severe side effects even in a low dose?

- Aurorix
- Paracetamol
- Haldol
- Seroxat
- Lasix

Example 2 is not a type 1 MR-multiple true/false question, because the options are interdependent. The options in a subset of two or more options are called interdependent precisely then when they together constitute a solution to the problem posed in the question, and no option from this subset can be left out without compromising this solution, nor can one option be joined to the subset.

The essential property of a subset of interdependent options is that it is the particular subset itself that is correct, not the individual options that make up this subset. The participant cannot evaluate interdependent options just by themselves of whether they are correct or incorrect, it is the particular subset itself that has to be evaluated.

Example 2 illustrates what is meant by interdependency. Only the combination of the 2 right medicines (aurorix and seroxat) is a correct answer to this question. Each of these two by itself is not an appropriate response to the question.

Therefore, the item type MR-combination is only correctly answered if the particular subset is selected. If two options are chosen of which one is right and one is wrong it is obvious that the response is false because both options in the correct subset are a necessary ingredient of the correct response.

3.2.3 Type 3: A MR-multi-combination has more than one correct subset of interdependent options

There are more combinations of options possible as a correct answer to the question. The correct subsets may overlap and are not necessarily of the same size. For example the combination (A+C) can be a correct answer to the question, but also the combination (A+B+D). Note that in this definition it is the subset itself that is correct, not the individual options.

Furthermore two variants must be distinguished:

- type 3a: single subset selection
- type 3b: multiple subset selection

In type 3a the participant is asked for only one combination, but in principle the question may ask for more or even all correct subsets, type 3b.
Type 3 MR-multi-combination is hardly discussed in the literature.

Example 3a: a MR-multi-combination item type 3a

Which combination of 3 different medicines can be given to the patient without severe side effects?

- Aurorix
- Paracetamol
- Haldol
- Seroxat
- Lasix
- Fevarin

In the above example there are more correct subsets that partly overlap, but the participant can respond with only one subset.

The next example (from Chapter 2) shows how one can ask for more than one subset. In this example the participant has to submit both possible subsets of attributes for producing a letter.

Example 3b: a MR-multi-combination item type 3b

Describe two methods to produce a letter to put in an envelope by selecting the needed attributes. Per method 2 or 3 attributes are needed. Attributes may be used in both methods.

	Method	Method
Pen	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Printer	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Paper	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Computer	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Telephone	<input type="checkbox"/>	<input type="checkbox"/>

In particular, this example shows that the union of two correct subsets of interdependent options is not a solution to the problem, and consequently incorrect. Another example is the following. If for a particular problem with a patient there are two ways to proceed, of which one must be chosen, and each of these ways can be phrased in a number of options, which possibly overlap, then this can be the core of a MR item with two correct subsets of interdependent options. Clearly it is not the individual option that is correct here but the entire subset: the way to proceed is phrased by means of the entire subset.

3.3 Design requirements and guidelines for MR-questions

Generally, the same design requirements apply for multiple response questions and multiple-choice questions. In the literature there are many checklists with do’s and don’ts (Haladyna, 2004), (CITO, Toetswijzer).

An example of such a checklist is the shortlist of (Haladyna, 2004,Table12). This list is a mixture of requirements and guidelines. In this list the numbers 4,7,8,15,19, 22 are requirements; numbers 2, 3, 26 are guidelines. Whereas design requirements come in at the end of the item design process, design guidelines are helpful at the initial stages of generating ideas, formats, and formulations of items (see also Chapter 4 of this book).

Therefore, a clear distinction between requirements and guidelines is helpful to the design process. For instance, a requirement that the layout of parts of the question should be vertically arranged does not really help the designer in the creative phase of item design. Experience in the ALTB project shows that at least in higher education subject matter experts are not pleased with requirements disguised as guidelines.

This experience but also literature on creativity and design shows that attention should not be directed to requirements in the initial stages of design and development. Guidelines should help the designer and give direction to the design process. This is also what subject matter experts and their assistants ask for.

Table 12: General item writing requirements and 'guidelines' (Haladyna, 2004p 99 table 5.1)

<p>Content Guidelines:</p> <ol style="list-style-type: none">1. Every item should reflect specific content and a single specific cognitive process, as called for in the test specifications (table of specifications, two-way grid, test blueprint).2. Base each item on important content to learn; avoid trivial content.3. Use novel material to measure understanding and the application of knowledge and skills.4. Keep the content of an item independent from content of other items on the test.5. Avoid over specific or over general content.6. Avoid opinion-based items.7. Avoid trick items. <p>Style and Format Concerns:</p> <ol style="list-style-type: none">8. Format items vertically instead of horizontally.9. Edit items for clarity.10. Edit items for correct grammar, punctuation, capitalization and spelling.11. Simplify vocabulary so that reading comprehension does not interfere with testing the content intended.12. Minimize reading time. Avoid excessive verbiage.13. Proofread each item. <p>Writing the Stem:</p> <ol style="list-style-type: none">14. Make directions as clear as possible.15. Make the stem as brief as possible.16. Place the main idea of the item in the stem, not in the choices. <p>Writing Options:</p> <ol style="list-style-type: none">17. Develop as many effective options as you can, but two or three may be sufficient.18. Vary the location of the right answer according to the number of options. Assign the position of the right answer randomly.19. Place options in logical or numerical order.20. Keep options independent; choices should not be overlapping.21. Keep the options homogenous in content and grammatical structure.22. Keep the length of options about the same.23. Avoid negative words such as not or except.24. Avoid options that give clues to the right answer.25. Make distracters plausible.26. Use typical errors of participants when you write distracters.27. Use humor if it is compatible with the teacher; avoid humor in a high-stakes test.
--

Next the requirements that have been developed in the ALTB project and that apply specifically to the different types of multiple response questions will be presented. These requirements primarily emerge from problems related to scoring (see section 3.4 below).

3.3.1 If it complies with the aim of the question the number of options must be fixed by instruction in the stem

In general the number of options to be selected should be fixed by instruction in the stem, because there are two major disadvantages in not giving the number of correct options:

- Letting the participant to decide how many options he will select, introduces variation into the data that is primarily related to personality traits that are not part of the ability we want to measure. One can imagine that a person who is overconfident will be inclined to select larger correct subsets than a less confident participant. A statement like this is always in need of empirical verification, but because empirical research into this question is unknown, it seems wise to be careful in this respect.
- When scoring polytomously, one has to take measures to compensate for wrong or forgotten options. If the number of options is given in the stem and the test taker chooses a different number of options, the answer is not in accordance with the instruction and can be scored with 0.

Because the correct subsets in type 3 MR-multi-combination items may differ in size, the instruction to select a particular number of options is, in general, not applicable with these items. Or one must generalize the instruction to for instance: "choose either 2 options or 4". This primarily tends to create confusion. Therefore, when the reasons to fix the number of options by instruction as given above outweigh the reasons not to fix the number of options by instruction, one should try to avoid MR items of type 3 and to rephrase them in such a way that only one correct subset of interdependent options remains.

Furthermore, when the learning objective or target competency requires that the student can distinguish relevant from irrelevant information or has to be able to select all necessary information or all necessary operations or tools it is better not to give the number of options that need to be selected.

In the discussion of the use of MR-questions in higher education below some more examples will be presented where recognizing the number of correct options is an essential part of answering the question.

3.3.2 If the number of options is fixed by instruction, then the number of correct options must be half the total number of the options or less and the total number of options must be at least four

This requirement only applies to type 1 MR-multiple true/false items. If the number of independent options to be selected is fixed by instruction, the number of correct options must be half the total number of the options or less. Otherwise, with partial credit scoring at least a score 0 is impossible, because even the worst selection necessarily contains a correct option. This requirement implies that the total number of options must be at least four, because the item must contain at least 2 correct options. If the number of options is fixed at one by instruction the question is not a MR question but an MC question.

Note that if the number of to be selected options is not fixed by instruction the number of options may be less than four, and, in principle, all options could be correct.

3.3.3 The maximum score of the question must be indicated in the test

Because there are different methods to score MR-questions there must be clarity to the test taker as to which method is used and what the maximum score will be.

3.3.4 Guidelines for writing MR-questions

A general guideline for writing MR-questions is the advice to aim at large option lists that are already available as a 'natural set' of options, for example a list of formulas or a list of medical treatments. Clearly, as the list of distracters or choices of an MR question grows, it limits drastically the chances of correct guessing (Scalise et al., 2006). When it is difficult to find incorrect options this advice does not imply to include obviously wrong options.

The use of a small number of options comes from the tradition of paper based tests. In computer-based assessment extensive option lists can more easily be used in multiple response questions.

In many practical situations a natural set of options is already existent, for example a list of formulas or a list of medical therapies. A typical example is example 4. The number of to be selected options in this example is fixed by instruction because first the aim of the question is to ascertain whether the participant is able to select a few, correct possibilities, and second, with so many alternatives the influence of personality traits can hardly be avoided with free selection.

**Example 4 : A MR-multiple true/false item with a large list of options
(From : Case et al., 2002)**

a) Calcium, b) Fluoride, c) Folic acid, d) Iron, e) Vitamin A, f) Vitamin B1 (Thiamine), g) Vitamin B6, h) Vitamin B12 (cyanocobalamin), i) Vitamin C, j) Vitamin D, k) Vitamin E

For each child select the appropriate vitamin or mineral supplements

Case 1. A 1-month old infant is brought to the physician for a well-child examination. He has been exclusively breast-fed, and examination shows normal findings. (Select two supplements)
Correct: b,j

Case 2. A 6-year-old girl has cystic fibrosis she has been taking no medications. (Select three supplements)
Correct: e,j,k

Note that various interaction types can be used for the selections. The interaction type is not given here in order to avoid the discussion about interaction types.

3.3.5 Some examples of the use of Multiple Response Questions in higher education

This subsection presents a few typical examples of MR questions and relates them to types of learning objectives that are relatively important in higher education in natural, engineering and social sciences. Note that in all these examples the requirement to fix the number of options was overruled by the requirement that in these questions it is essential that the participant recognizes all the appropriate options, and not just selects the so many best options in his view.

The use of multiple response questions in problem solving

MR questions are perfectly well suited for measuring whether a participant knows or can infer what information is relevant or necessary to find or create solutions to a given problem. This also applies to problems that at first sight seem to be calculation problems. In higher education, most subject matter experts want to focus not so much on the actual calculation problem or the correct numerical outcome to a calculation problem. Rather a participant must be able to identify specific steps in the calculation or inference chain or be able to select all the necessary elements that are needed to arrive at a solution to the problem or to distinguish between relevant and irrelevant information, tools and operations (Biggs, 1999).

The MR-question type is the most basic question type to handle this class of measurement problems. In theory, one can also use drag and drop but in most currently available test software designing drag and drop questions is more expensive and technically complicated than MR-questions. Furthermore, there has to be agreement on the method to score drag and drop items.

**Example 5. A MR-combination item
E. Boer: Course sample and monitoring, WU**

Suppose we would like to test a lot of powdered milk on Salmonella.

- A lot of 20.000 kg powdered milk is produced.
- 15 sample units of 25 g are taken randomly
- A lot is only accepted if all samples are negative
- Suppose you know that 100.000 nests of salmonellae are present in the lot and are homogeneously distributed over the lot.

Which of the following distributions should you use to calculate the probability of accepting the lot?

- Binomial distribution
- Normal distribution
- Lognormal distribution
- Poisson distribution
- Uniform distribution
- Standard distribution

In the example the correct subset is indicated by marks in the checkboxes. In example 5 only one subset is a correct subset. The selection size is not fixed by instruction because it is essential that the participant recognizes the necessary distributions.

In the next example two subsets are correct.

Example 6. An MR-multicombination type 3b item

Adapted from: L. Rietveld: Process technology, Course Drinking Water Treatment, TUDelft

What information is necessary to calculate the concentration of oxygen in water that is in open connection with the outside air?

	Subset1	Subset2
Temperature of the water	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Temperature of the air	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Volume percentage of oxygen in the air	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Valence number of electrons in the oxygen	<input type="checkbox"/>	<input type="checkbox"/>
Molecular mass of oxygen	<input type="checkbox"/>	<input type="checkbox"/>
Partial pressure of oxygen	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Ion strength	<input type="checkbox"/>	<input type="checkbox"/>
Partitioning coefficient H	<input type="checkbox"/>	<input type="checkbox"/>
Molar fraction oxygen	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Atmospheric pressure	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Molecular mass of H ₂ O	<input type="checkbox"/>	<input type="checkbox"/>
The gas constant	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

In this case the two correct subsets must both be given. Thus the question is of type 3b MR- multi-combination multiple subset selection.

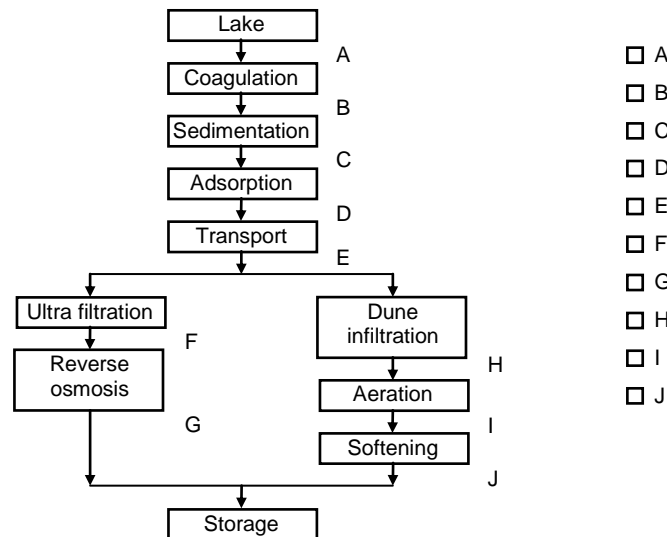
The use of multiple response questions in procedural knowledge

Multiple Response questions are not yet very often used in the field of process knowledge. Drag and drop and ordering question types are better suited to measure knowledge about relationships in a process-model (see Chapter 6). Nevertheless there are some good examples of using MR-questions in measuring the ability of a participant to indicate positions of sub processes in a process diagram.

Example 7. An MR-combination item

L. Rietveld: Process technology, Course Drinking Water Treatment, TUDelft

The figure shows a scheme of a groundwater treatment plant. For each position indicate if in this position a rapid sand filtration functionally can be placed.



Based on the incomplete groundwater treatment plant presented in this example 7, different types of MR questions can be formulated for different purposes. One purpose is to test primarily if the student knows and understands the interface of rapid sand filtration. The question could then be formulated as "...for each position indicate if rapid sand filtration can functionally be placed ...". This is the formulation in the given example. The resulting question is a type 1 Multiple true/false question. The correct answers are in this case C, I, J. In such a question, most subject matter experts in higher education would refrain from fixing the number of correct answers by instruction. The second purpose is to test primarily if the student knows and understands a complete ground water treatment plant. The question could then be formulated as "... indicate a combination of placements of rapid sand filtration such that the total system constitutes a complete ground water treatment plant." Two answers can be correct answers in this case: the combination C, I, J or the combination C, J. Thus the type is type 3a MR-multi-combination. Alternatively a type 3b MR-multi-combination can be formulated for the purpose to test if the student understands and knows about the complete ground water treatment plant: "... indicate two combinations of placements". In this case the correct answer is the combination C, I, J, and C, J.

3.4 The scoring of multiple response questions

When a participant is asked to select for each item the options he judges correct, the raw response to the test is a series of subsets of options. Traditionally, the way to convert this series of subsets into a single numerical judgment as the test result is regarded to consist of two parts. The first part of the problem is to devise a reasonable way of scoring the response per item, i.e. to convert the chosen subset of options into a numerical item score. The second part of the problem is to combine these item scores into a single test score. Although this is far from unproblematic, most of the times simple addition of item scores is adopted as the only possible and reasonable choice for this second problem.

As a new development a response type other than choosing a subset of the options is introduced for type 1 questions (MR-Multiple true/false), viz. giving a rank order of the options, the subjectively most likely first.

The following conclusions will emerge:

- Rank orders allow for better measurement precision than selection of a subset.
- It is inefficient to instruct participants to select less options than the number correct.
- With fixed selection size an item should not contain more correct options than half the total number.
- With free selection size the number of correct options can be anywhere between two and the total number of options.

3.4.1 Types of item scores for subset selection

The discussion of item score types for subset selection is based on the distinction between the three types of MR items from the previous section.

They were:

Type 1. MR-multiple true false, i.e. the item contains a subset of independent correct options

Type 2. MR-Combination, i.e. one subset of interdependent options is correct

Type 3. MR-Multi-combination, i.e. more than one subset of interdependent options is correct

Note that in Type 1 the phrase 'correct options' is used whereas in the other two types of items the phrase 'correct subset' is preferred. A major distinction between Type 1 and the other two types is the presence or absence of interdependency within a subset of correct options. Therefore, it will now be discussed and made more precise what is meant by a subset of *interdependent* options versus a subset of *independent* options. Next, it will be pointed out what the consequences are for instruction and scoring if the third item type is part of the test.

The options in a subset of two or more options are called interdependent precisely then when they together constitute a solution to the problem posed in the question, and no option from this subset can be left out without compromising this solution, nor can one option be joined to the subset. In particular this means that the union of two correct subsets of interdependent options is not a solution to the problem, and consequently incorrect. The essential property of a subset of interdependent options is that it is the particular subset itself that is correct, not the individual options that make up this subset. The participant cannot evaluate interdependent options just by themselves of whether they are correct or incorrect, it is the particular subset itself that has to be evaluated (see e.g. Example 6).

Within an MR-multiple true/false item with a subset of independent correct options, each nonempty subset of the correct options is (partly) correct. In particular each correct option by itself is a solution to the stated problem. This cannot be said of the correct subsets of interdependent options. If, for instance, in response to a sleeping problem of a patient, one has to offer him a. a somniacelesta sleeping pill, and b. a glass of water to swallow it, it would be preposterous to claim that selecting only b. is partly correct.

In general, literature on MR question types and scoring of MR questions is scarce. For instance, in (Brennan, 2006) the term Multiple Response item is not found in the index. An exception is Lampe and Eggen (2003) who discuss a number of item scoring rules for MR items.

They define a question type with the following characteristics:

- more than one of the options is correct
- a participant has to select more than one correct option to produce a completely correct response
- the completeness of the response determines the item score.

This definition and the examples of MR items in their treatment are described with type 1 MR-multiple true/false items. Moreover, some of their scoring rules give partial credit. In the sequel it will be shown that partial credit scoring only make sense for item type 1 MR-multiple true/false.

3.4.2 Types of item scores for subset choices of fixed size by instruction

Lampe and Eggen argue that dichotomous scoring should be preferred over polytomous scoring. In this section it will be shown that their arguments are inconclusive. Here it is argued that polytomous scoring is more informative than dichotomous scoring, and, therefore, should be opted for. Moreover, it is shown why items in the category of this section should not have more correct options than half the total number of options.

First item scores are discussed with an instruction that specifies the number of options to select. This excludes type 3 items.

The following scoring types are distinguished by Lampe and Eggen (o.c.):

- A. dichotomous scoring: the response is considered correct only if the participant selects the indicated number of options and all selected options are correct. In this case the score is 1 otherwise 0.
- B. partial credit scoring: if the participant selects the indicated number of options, his score equals the number of correct options chosen, otherwise the score equals 0.
- C. partial credit scoring with deduction:
 - C1. from the number of chosen correct options one deducts the number of chosen incorrect options
 - C2. to the number of chosen correct options one adds the number of not chosen incorrect options

Lampe and Eggen show that partial credit scoring with deduction types C1 and C2 are equivalent. If participants follow the instructions and select the indicated numbers of options Lampe and Eggen show that also their scoring types B (partial credit) and C (partial credit with deduction) are equivalent. Therefore, essentially, only scoring types A (dichotomous) and B (partial credit) remain.

Lampe and Eggen argue that, with some exceptions, dichotomous scoring is more reliable than partial credit. However, their approach in showing the alleged superiority of dichotomous scoring invokes the following doubts. In the first place they obtain this result only for a very limited selection of populations, a selection that is not based on any analysis whatsoever. Secondly, the first illustrative item they use for their calculations is a 5- option item with 3 options correct, where exactly 3 options have to be chosen. Consequently, in that particular example, partial credit scoring is severely hampered because score 0 cannot occur, because each subset of three options necessarily contains at least one correct option, because there are but two incorrect options. In their second example they use a 7-option item of which 4 are correct and 3 have to be selected. Here score 0 is possible, but only if exactly the one unique incorrect subset is selected. Even with complete ignorance the probability to select this subset equals 1/35. This again severely hampers the partial credit method of scoring. Last but not least, the classical reliability is an obtrusive measure for this problem. This becomes clear if one applies dichotomous scoring at the level of a complete test. Suppose one would, for instance, with a test of forty dichotomously scored items, introduce the dichotomous test score as follows: less than twenty items correct results in a test score 0, and twenty or more correct yields test score one. A pass/fail decision is such a dichotomous test score. The classical reliability argument would lead to the conclusion that the complete score range from zero to forty is less reliable than the dichotomous test score, because the relative amount of error variation in the first case is larger. Nevertheless, it is commonly felt that this argument is not compelling to prefer the dichotomous test score over the score range from zero to forty. In particular, if one has passed the test, the score gives extra information on how well the test was passed.

These arguments make clear that the discussion on the preference of dichotomous scoring over partial credit scoring of MR items has not been conclusively settled by the discussion in Lampe and Eggen. Notice that in dealing with this problem the arguments in this subsection also indicated why it is not advisable to have more than half the number of options of an item correct. (see design requirement 3.3.2)

3.4.3 Types of item scores for subset choices of free size

When the participant has a free choice as to how many options to select, partial credit scoring and partial credit scoring with deduction are not equivalent. This can easily be understood by pointing out that by selecting all options one obtains invariably the perfect score with partial credit scoring, but not so with partial credit scoring with deduction, unless all options are correct. This observation indicates that partial credit scoring is not compatible with free choice of selected subset size. Therefore, in this case we are only left with dichotomous scoring and partial credit scoring with deduction. The two subtypes of partial credit scoring with deduction remain equivalent.

For item types 2 (MR-Combination) and 3 (MR-Multi-combination type 3a single subset selection), only dichotomous scoring can be reasonably defended, because only a complete subset is correct, and one can only select one subset, also with type 3 MR-Multi-combination items. For MR items of type 3a (single subset selection) an additional argument that only dichotomous scoring is applicable can be put forward. An example suffices to clarify this. Suppose that partial credit scoring with deduction subtype C1 (From the number of chosen correct options one deducts the number of chosen incorrect options) is adopted, and that there is an item with 6 options with the interdependent correct subsets {1,2} and {1,3,4}. Imagine that the participant selects subset {1,3}. If it is supposed that the participant made an error in selecting option 2 in {1,2} he obtains score 0. If, on the other hand, it is supposed that the participant opted for {1,3,4} but missed option 4, he obtains score 2. Because in the data there is no information about which of the two subsets was leading in his selection, the item score can not be decided on. Only dichotomous scoring is unequivocal in this case.

The above is not true for the MR-Type 3b where more than one subset is asked for (multi subset selection, see example 3b). This variant can be scored polytomously: one point for every correctly selected subset. Evidently, this is equivalent to a series of dichotomous scores, as if more questions of type 2 (MR-Combination) are involved, and scored dichotomously. Of course, dichotomous scoring of a MR-Type 3b question is also a possibility, for instance, the participant obtains a score 1 only if all selected subsets are correct. However, this is not to be preferred because it leaves information in the response about the ability of the participant unused. With a free choice of the number of selected options, the rule that the number of correct options should not exceed half the total number of options no longer applies. Even to have all, or zero options correct would be perfectly appropriate.

Table 13 summarizes the dependencies between item types, instruction types and score types as derived in this section.

Table 13: Scoring type dependencies derived in this section

MR-Type	Instruction	Score Type
1. MR-Multiple True/False	Fixed	Dichotomous, Partial Credit = Partial Credit with deduction*
	Not Fixed	Dichotomous Partial Credit with deduction
2. MR-Combination		Dichotomous
3a. MR-Multi-Combination	Single Subset Selection	Dichotomous
3b. MR-Multi-Combination	Multiple Subset Selection	Dichotomous Partial Credit

*In this case the Partial Credit with deduction score s_3 can be obtained from the Partial Credit score s_2 via a simple transformation ($s_3 = 2s_2 + c$, for some constant c) and vice versa. This means that in this case the two score types result in exactly the same rank order of the participants.

3.4.4 Combining item scores to a test score

After having chosen how to score the individual item responses the item scores have to be combined into a final test score.

We can distinguish two main approaches:

1. Classical methods
2. Latent trait methods, also called Item Response Theory (IRT) methods

Classical methods are typically restricted to weighted or unweighted addition of item scores, In the case of weighted addition one assigns to every item a weight, and multiplies the item scores with their weight before they are added. Giving weights to items often happens for psychometrically irrelevant reasons. Moreover, usually it only marginally influences the rank order of the test scores. This means that the rank order of the participants scored in a weighted fashion shows only minor changes in comparison with the rank order of the participants with unweighed addition of item scores. Of course it is assumed that the items tap a common ability, which should be the intention of the test constructor. Therefore, weighing items on subjective grounds cannot be advised.

This leaves us with simple unweighed addition of item scores. When more tests are involved in an investigation, and equating is an important issue, it is a disadvantage of this approach that one is limited to classical methods of equating. According to (Lord, 1980, p. 198) , classical equating can only be accomplished when it is superfluous, that is when the equating transformation implies that the score remains unchanged. Another disadvantage of classical scoring emerges when one pays a large investment in time and money to obtain norms in a population, and several of its subpopulations.

If after the investigation one wants to change one or more items, because they cannot meet newly emerging quality standards, the norm-study has to be redone.

Within the IRT tradition several methods combine the item scores into a final test score, called an ability estimate. Dichotomous scores can be analyzed for instance with the Rasch model, the OPLM, or the Birnbaum two and three parameter models. Partial credit scores can be analyzed with the OPLM, the GPCM, or the GRM. These are well known IRT models that can be found in most text books on psychometrics, such as (Linden et al., 1997). The two disadvantages mentioned with the classical approach are in principle solved within the IRT framework, or, as with the norms, can be repaired at much lower cost.

3.4.5 Item scoring types for rank orders

The discussion will now focus on type - 1 questions (MR-Multiple true/false). As a new development in this field we will introduce a response type other than choosing a subset of the options for type 1 questions (MR-Multiple true/false), viz. giving a rank order of the options, the subjectively most likely first. It will be shown that rank orders allow for better measurement precision than selection of a subset. Moreover, the discussion will also clarify that it is inefficient to fix the selection size to less than the number of correct options.

In choosing a rank order of the options of an MR-Multiple true/false item the participant is asked to evaluate the subjective correctness of each option and to order the options according to this correctness. The option that in his opinion is the most correct is to be given the first rank. Note that with increasing rank the subjective correctness decreases. Because each option is supposed to be individually judged on its correctness, this type of response is compatible only with independent options as in item type 1.

Given rank orders as a response, it is easy to infer the selection of options had the participant been given the instruction to select a subset of correct options of a given size. That is, if one is willing to admit that it would be weird for a participant to select a particular order, but on the request to indicate e.g. the k options of his choice would not pick the first k options of his rank order. Therefore, rank order data admit the same item and test scoring methods, be they classical or IRT, as discussed above with selection of option subsets of fixed size. This observation implies that ordering responses carry at least as much information on the ability of a participant as selection of a subset of fixed size.

Apart from these approaches rank order data can also directly be analyzed with an IRT model, without first converting the rank order to an item score. This model is called the Luce model, and models rank order data as described in (Verstralen et al., 2007). Therefore, with this approach converting the response into an item score can and must be omitted. The analysis with the Luce model directly converts the rank order data on the items in a test into a test score, called an ability parameter.

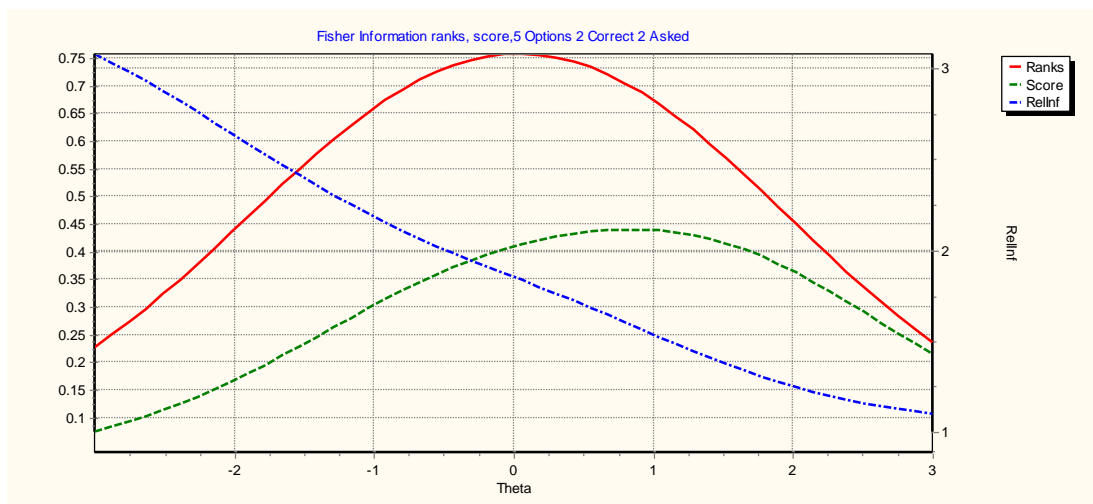
To evaluate the advantages of the Luce model for rank order data in comparison with the scoring approaches discussed earlier, we need to be aware of the statistical nature of participant responses to a test. Imagine that we have two tests that are identical in all respects except that their items differ. If these two tests are administered the one after the other, in general, participants do not obtain the same score on both tests. Nevertheless, when all things go well the two test scores will show a high correlation. This means that when a participant obtains a high score on the first test there is a very substantial chance that his score on the second test is also high, not necessarily identical, but high. So test results are not completely precise. The greater the precision, the better the test informs us about the ability of the participant. To quantify precision we use a statistical measure called the Fisher information of a response. The Fisher information of a response as a random variable equals minus the expectation over all possible responses of the second derivative of their log likelihood with respect to the person parameter. Its inverse equals the asymptotic

variance of the maximum likelihood estimator of the person parameter. The higher the Fisher information the greater the precision of the test score. It is an important property of the Fisher information that the Fisher information function of the test is simply the sum of the Fisher information functions of the items in the test.

Let us return to the Luce model. If an item score is implied by the rank order, such as just explained, the Fisher information of the rank order and the item score can be compared. If, for instance the Fisher information of the rank order is twice as large as the Fisher information of a particular item score, a twice as long test using the item score is needed as when using rank orders to obtain the same precision of the test score. Unfortunately, to complicate matters the Fisher information depends on the ability of the participant. Therefore, various levels of ability need to be differentiated, such as high, medium and low (in relation to the item), in statements about Fisher information of rank order and an implied score. The results of the calculations will also be shown in graphical form.

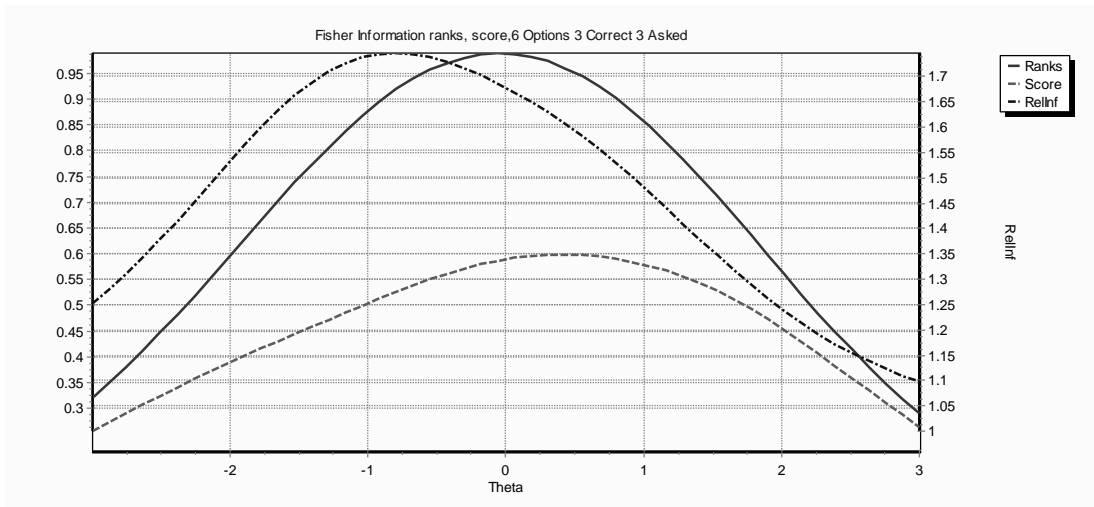
Imagine a hypothetical item with five options of which two are correct. The Fisher information of the rank order and of the partial credit score for the selection of two options is compared in Figure 7. On the horizontal axis the test score that could be obtained from a theoretical test of maximal precision is shown: this test score is the true ability parameter θ . The left vertical axis shows the Fisher information represented by the two bell shaped curves. The higher bell shaped curve is associated with rank order data, the other with the partial credit score. The decreasing curve, from left to right, shows how much more information is obtained with rank order data compared to the score data, and its value is given in the right vertical axis. This is also known as relative information (of the rank order with respect to the partial credit score). As can be seen the rank order response gives at least twice as much precision for the lower range of the ability as the partial credit score. For the higher abilities the size of this advantage diminishes, to become negligible for the truly high abilities.

Figure 7: Fisher information of rank order data and the selection partial credit score of an MR-item with 5 options of which 2 correct, and 2 to be selected for the score



The second example concerns an item with six options of which three are correct and in case the partial credit score would be used three options are to be selected. The results are shown in Figure 8.

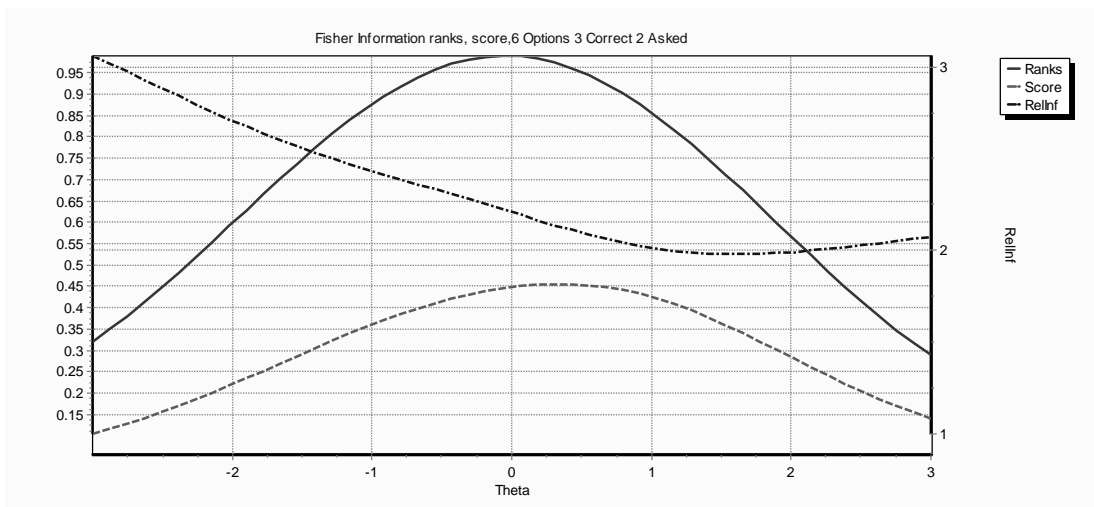
Figure 8: A similar graph for an MR item with six options of which 3 correct and 3 to be selected for the partial credit score



In this picture the relative information is also bell shaped. It is the left most curve at the left side of the picture. Here, as well in the previous picture the rank order information is everywhere higher than the partial credit information. But the advantage of the rank order is highest here for the medium level abilities, where the rank order is more than one and a half as precise as the partial credit score

The final example to be discussed is an MR item with six options 3 correct but only 2 options to be chosen for the partial credit score.

Figure 9: Six options of which 3 correct and 2 to be chosen for the partial credit score



What immediately strikes the eye in viewing Figure 9 is that the rank order is more than twice as precise as the partial credit score over the entire range of the ability. The Fisher information of the rank order is, of course, identical as in Figure 8. This also implies that if one opts for the combination of subset selection of fixed size and the partial credit item score it is a waste of resources, an avoidable loss of precision, to ask for a subset size that is less than the number of correct options.

3.5 Conclusions

In this chapter three different types of multiple response questions have been distinguished. These types are: type 1 items with independent correct options, type 2 items with one and type 3 items with more correct subset(s) of options. For type 3 items two variants have been distinguished: type 3a – select one correct subsets and type 3b select a given number of correct subsets where the given number is larger than one.

Generally, the same design requirements apply to multiple response questions and multiple-choice questions. For multiple response questions it has been argued that the following additional requirements apply: (1) With some exceptions the number of to be selected options should be fixed by instruction. The technical implementation in test software should support the possibility to enforce the actual number of selections that the participant can and must make. A typical exception is a question that measures if the student can distinguish relevant from irrelevant information. (2) With a fixed selection size the number of correct options must be half the total number of options or less. (3) The participant must be informed as to the score he gets by correctly answering the question. (4) Furthermore it is advised to reduce the probability of correct guessing by using large option lists.

The use of multiple response questions for some categories of learning objectives that are common in higher education has been illustrated. For this description we used some examples of design patterns.

Multiple response questions are well suited for measuring whether a participant knows what information is relevant and irrelevant in the area of problem solving. In calculation problems multiple response questions are mainly used to identify mistakes in a given calculation or to select all the necessary elements that are needed to make a calculation. Finally we gave an example of the use of multiple response questions in the field of procedural knowledge. The example showed that multiple response questions are suited for measuring the ability of a participant to indicate positions of sub processes in a process diagram.

We argued that items with a correct subset of options (item types 2 and 3a single subset selection) can only reasonably be dichotomously scored. Items with independent correct options (item type 1) can also be scored with partial credit. In the latter case, when the number of to be selected options is fixed by instruction one only needs to count the number of correct options in the selection. When this number is not fixed, and the size of the selection is up to the participant, either dichotomous scoring or partial credit scoring with deduction must be applied. If applicable polytomous scoring should be preferred over dichotomous scoring. Type 3b (multiple subset selection) can also be scored with partial credit, by granting a score 1 for every correctly selected subset. This latter approach has our preference because it uses the information in the response about the participants' ability more optimal.

Finally, we introduced rank orderings of options as data for type 1 items, and showed that analysis with the so-called Luce model offers substantial advantages in terms of measurement precision of the test result compared to partial credit scoring of selection of options. Moreover, it was shown that if the size of the selected subset is fixed by instruction one loses precision if this fixed size is less than the number of correct options. However, until now, there is little experience in gathering rank order data for MR-questions.

3.6 References

- Biggs, J. B. (1999). *Teaching for quality learning at university* Buckingham: Open university Press.
- Brennan, R. L. (2006). *Educational measurement* . (4 ed.): American Council on Education.
- Case, S. M., & Swanson, D. B. (2002). *Constructing Written Test Questions For the Basic and Clinical Sciences Third Edition (Revised)*: NBME
- CITO. Toetswijzer. Retrieved march 20 2007, from <http://toetswijzer.kennisnet.nl>
- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (Third ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Lampe, T., & Eggen, T. (2003). *Innovative Item Types in Computer-based Testing : Scoring of Multiple Response Items*. Paper presented at the IAEA 2003, Manchester, UK.
- Linden, v. d., W.J. , & Hambleton, R. K. (1997). *Handbook of modern item response theory* Heidelberg: Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Erlbaum.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York,: Springer-Verlag.
- Scalise, K., & Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment* 4(6).
- Verstralen, H., Maris, G., & Bechger, T. (2007). *Modeling ordered options of closed questions* (in preparation). Arnhem: Cito.

4 Guidelines for the Design of Digital Closed Questions for Assessment and Learning in Higher Education

Published as "Draaijer, S., & Hartog, R. J. M. (2007). Guidelines for the design of digital closed questions for assessment and learning in higher education. e-Journal of Instructional Science and Technology (e-JIST), 10(October)."

Silvester Draaijer
Vrije Universiteit Amsterdam

Rob Hartog
Wageningen University

Joke Hofstee
Cito

Abstract

Systems for computer-based assessment as well as learning management systems offer a number of innovative closed question types which are used more and more in higher education. These closed questions are used in computer-based summative exams, in diagnostic tests, and in computer-based activating learning material. Guidelines focusing on the design of closed questions were formulated. In the ALTB project the use of these guidelines was evaluated in fifteen case studies in higher education. This chapter focuses on the rationale for each of the guidelines and the evaluation of their actual use in the ALTB project. The overall conclusion is that guidelines are useful, but should be applied in a broad approach that is preferably supported by educational technologists. The next chapter provides the approach that has been developed in the ALTB project.

Given this overall conclusion the primary intended audiences for this chapter are educational technologists who support question design teams and researchers in the field of instructional design and assessment design in higher education.

4.1 Introduction

During the last decade, a range of selected response format questions and other formats that allow for automatic scoring, have emerged in computer-based testing software (Bull et al., 2001; Mills et al., 2002; Parshall et al., 2002) and Learning Management Systems (LMSs) such as Blackboard or WebCT. Examples of such questions are 'multiple response', 'drag-and-drop', 'fill-in-the-blank', 'hot spot' and 'matching'. For reasons of readability, from now on the term 'closed question' will be used. In higher education such closed questions are used in summative tests (exams), in diagnostic tests but also in activating learning material (ALM). ALM forces the student to actively engage with the learning material by making selections and decisions (Aegerter-Wilmsen et al., 2005; Diederens et al., 2003).

As any design endeavour, the design of sets of closed questions is likely to benefit from a design methodology. The ALTB project (Hartog, 2005) aims to develop such a methodology for the design and development of closed questions for summative exams (SE) and activating learning material (ALM) for engineering and life sciences in higher education. This methodology is expected to consist of design requirements, design guidelines, design patterns, components, and task structures.

The research question of the ALTB project is essentially: 'How and under what conditions is it possible to support the design and development of digital closed questions in higher education? The answer should support the rationale for the methodology. This chapter focuses specifically on the development and evaluation of design *guidelines*.

4.1.1 Limitations in current literature on design guidelines

Literature on the design of questions with a closed format is mainly restricted to the design of summative tests that consist of 'traditional' multiple-choice questions. This literature, for example Haladyna et al. (2002), usually presents a large set of design requirements i.e. constraints that must be satisfied by the questions that are output of the design process. An example of such a constraint is the rule that every choice in a multiple-choice question should be plausible. A constraint like this helps to eliminate a wrong or poorly constructed question, but it does not help to create a new question or better distractors. Only certain requirements can be regarded as direction giving requirements rather than as constraints, but many requirements are not useful for directing and inspiring question designers.

Nevertheless, in literature on the design and development of questions and tests, requirements are often denominated as 'guidelines'. The use of the term 'guideline' for 'requirements' obscures the lack of real design guidelines i.e. rules that open up creative possibilities for question design and support the designer(s) during the design process.

Insofar literature does provide inspirational guidance for designers and developers of closed questions - as for example by Roid and Haladyna (1982), Haladyna (1997) or Scalise and Gifford (2006) - these sources are in the form of quite elaborate texts or research reports and more suited for secondary or vocational education. Given the limited time for training or study available to SMEs in higher education, SMEs do not use these sources and do not feel that they are appropriate.

For that reason, it is assumed that more 'compact' and easily accessible guidelines, preferably in the form of simple suggestions, can be more useful in practical situations in higher education. Based on that idea, in this project, a set of 10 categories and direction giving requirements was formulated and made available in the form of an overview table and brief explanations. In practice in higher education, the same technology and the same question types are used for both summative exams as for activating learning material. Therefore, at the outset of the project, it was the intention to develop guidelines that were suitable for the summative role and the activating learning role.

4.2 The guidelines: dimensions of inspiration

In this section a set of guidelines for the design of closed questions and the rationale of these guidelines will be described. The intention was that the guidelines should serve as an easy to use and effective support for SMEs, assistants and ETs for the design and development of questions and tests.

In order to arrive at a set of potentially useful guidelines, the ALTB project team formulated a set of guidelines. These guidelines were partly derived from literature and partly from experience of the project team members. Some guidelines are quite abstract, other guidelines are very specific, some guidelines refer to methods. The guidelines were grouped into specific categories each of which was intended to define a coherent set of guidelines. The list comprised ten categories: seven categories consisted of guidelines that tap into the use of experiences and available resources for question designers, three categories were essentially traditional requirements. However, those were requirements that also give direction and inspiration to the design process

Guidelines for question designers:

- A. Professional context
- B. Interactions and Media
- C. Design Patterns
- D. Sources
- E. Learning Objectives
- F. Students
- G. Sources

Traditional requirements, which give direction and inspiration:

- H. Motivation
- I. Validity
- J. Equivalence

These categories were subdivided in more specific guidelines, resulting in a total of 60 guidelines. In the following sections, the guidelines are described in more detail.

4.2.1 A: Professional context

This category of guidelines makes question designers focus on the idea that information is more meaningful when it is presented or embedded in real life professional situations (e.g. Merriënboer et al., 2002). Based on that idea, the professional context of a graduated professional in a specific domain could be the basis of these questions. To cover multiple aspects of such cases, more than one question should be defined. An obvious source for such authentic situations can be the professional experience of the question designer himself.

In a more systematic way, question designers can use explicit techniques for constructing and describing cases, for example in the form of professional encounters or problem situations vignettes (Anderson et al., 2001), or item sets that can serve various variations of questions for specific topics or topic variants (Haladyna, 2004; LaDuca et al., 1986; Roossink et al., 1992). A second source that thrives on professional knowledge and experience is to tap into 'Eureka' experiences the professional has had in his own learning and professional development. More specifically these types of situations were worked out in tips and tricks, surprising experiences, counter-intuitive observations and natural laws, relevant orders of magnitude, typical problems and best first steps for tackling them.

Finally a guideline that often pops up in the practice of instructional design projects is the advise to collect all kinds of material (interviews, documentaries, descriptions, journal clippings, broadcast video and audio), that can be used to construct or illustrate cases.

Professional context	
A1	Develop cases with authentic professional context and multiple relevant questions.
A2	Develop vignettes using an item-modelling procedure: split up authentic cases in various components and develop new content for each component and combine them into questions.
A3	Investigate your own professional experience. Make lists of:
A3.1	Tips and tricks.
A3.2	Surprising experiences.
A3.3	Counter-intuitive observations and natural laws.
A3.4	Relevant orders of magnitude.
A3.5	Typical problems and the best first steps.
A4	Collect interviews, documentaries, descriptions (in text, audio or video) of relevant professional situations. Use these for question design.

4.2.2 B: Interactions

The introduction of the computer in learning and assessment makes a new gamut/scale of question types and interactions possible. The ALTB project team anticipated that when question designers play with assessment software and study the accompanying examples, they become inspired. To guide question designers more specifically on the dimension of digital media inclusion, guidelines were formulated that take *specific* digital media types into mind which would lead to more appealing questions or that would measure the intended attribute of interest more directly: pictures and photos, video's, audio, graphs, diagrams, process diagrams.

Interactions	
B1	Play with available assessment software. There is a variety of assessment systems on the market. For inspiration on asking new questions and test set-ups: try out the interactions in the system that is used in one's own organization.
B2	Scan the IMS-QTI interaction types on usability.
B3	Collect material for media inclusion:
B3.1	Pictures / photos.
B3.2	Video clips.
B3.3	Sounds / audio fragments.
B3.4	Graphs.
B3.5	Diagrams.
B3.6	Process diagrams.

4.2.3 C: Design patterns

The term design patterns is introduced by (Alexander, 1979) in the seventies of the last century as a concept in architectural design. In design in general, reuse of components as well as reuse of patterns is beneficial because it usually is efficient but also because reuse of components and/or patterns increases the probability that errors or disadvantages will be revealed. An experienced designer is supposed to have many patterns in his mind. "It is only because a person has a pattern language in his mind, that he can be creative when he builds" (Alexander, 1979, p. 206). Because design patterns for digital closed questions were not readily available, the ALTB team adopted a simpler approach, using types of directions that could be indicative for design patterns. A few guidelines were presented that could be viewed as preliminary versions of design patterns or families of design patterns.

The first pattern was taken from Haladyna (2004, p. 152). This pattern, presented as a guideline, advises question designers to use successful 'starting sentences' that can easily result in interesting and relevant questions. A similar guideline by Haladyna (2004, p. 153) advises question designers to take successful items, strip the items of specific content, however leaving the systematic of the question unaltered, and then systematically design questions based on variations of content. This can be regarded as a generic advice to use design patterns. Another set of design patterns direct question designers toward questions that ask for completion of statements or calculations, to identify mistakes in reasoning or calculations, and to identify the best descriptions or key words for presented texts. The last guideline is based on ideas by Wilbrink (1983). Wilbrink suggests that – especially for designing True/False questions – it is a worthwhile technique to relate different (mis)concepts, to use (in)correct causes and (in)correct effects of concepts as a starting point for questions.

Design Patterns	
C1	Items shells I: Use a list of generic shells. Examples: <ul style="list-style-type: none"> • Which is the definition of ...? • Which is the cause of ...? • Which is the consequence of ... ? • What is the difference between ... and ...?
C2	Item shells II: Transform highly successful items into item shells.
C3	Collect chains of inference and calculations as a basis for a completion question. The completion question requests to fill in the missing rule in an inference chain or calculation
C4	Use design pattern "Localize the mistake": introduce a mistake in a text (paragraphs), photo, diagram etc. and use this as the stem. (Collect texts, photo's and so on.)
C5	Use design pattern "Select the (3) best key words" to a text. (Collect texts)
C6	Use design pattern "select a title" to a text. (Collect texts)
C7	Develop implications of statements.

4.2.4 D: Textbooks

In many courses in higher education, the dominant instructional sources are publishers' textbooks or the course syllabus. These books hold the core of the subject matter for a given course. For question design, a guideline is to use the content of these books not at random, but systematically. Whilst it was anticipated that a large number of question designers could feel that such a guideline was too 'simplistic', pointers that are more specific were added to guide question designers more precisely. The pointers were categorized into the use of media such as photos, graphs, and diagrams on the one hand and statements, contradictions, conclusions, exceptions, examples, abstract concepts, and course specific content emphasis made by the instructor on the other hand.

Textbooks	
D1	Walk systematically through the textbook (paragraph by paragraph) and look for:
D1.1	Photos.
D1.2	Diagrams.
D1.3	Graphs.
D1.4	Statements.
D1.5	Contradictions.
D1.6	Conclusions.
D1.7	Exceptions.
D1.8	Examples.
D1.9	Abstract concepts.
D1.10	What paragraphs and concepts hold key information and which do not.

4.2.5 E: Learning Objectives

Course goals and learning objectives are essential ingredients in instructional design (Dick et al., 1990) and for the design and development of tests and questions. Clear learning objectives are the basis for establishing valid assessment and test objectives: what will be assessed in what way, at what level (often resulting in a test matrix). However, detailed learning objectives are not well specified in advance in many design and development situations. Often it is not possible to formulate a detailed learning objective in terms of natural language phrases. In fact, often the first question for a detailed learning objective is itself a specification of this learning objective. In such situations, making questions without first specifying the detailed learning objectives is a realistic procedure.

Furthermore, a question designer could analyse and categorise the questions that are already available in previously designed assessment material thus raising the objective formulation to a higher level of abstraction. Based on the assumption that previous assessments reflect the knowledge and skills the instructor finds important for a course, this categorisation can be used to design new questions.

Categorisations as described above, will often be formulated in terms of domain specific knowledge and skills that need to be acquired. Taking a top down approach however, questions designers are advised to start with using more abstract formulations of the types of knowledge and types of cognitive processes that need to be assessed with the support of a *taxonomy* or *competency descriptions*. Literature provides several taxonomies. but an often proposed taxonomy is Bloom's taxonomy (1956) or the taxonomy as proposed by Anderson and Krathwohl (2001).

Learning Objectives	
E1	Use an existing list of very specific and detailed formulated learning objectives.
E2	Make a list of very specific and detailed formulated learning objectives.
E3	Analyse educational objectives using a taxonomy of objectives.
E4	Use the competency description of a course as a starting point to design questions.

4.2.6 F: Students

The students' mind set, experiences and drives should – at least for learning materials – be a source of inspiration for the question designer (Vygotsky, 1978). Four guidelines were formulated that express this point of view.

The first guideline directs the question designer towards imagining prior knowledge of the student; specifically when it relates to the subject matter or the learning objectives of the course. Thus, questions relating to for example food chemistry, should build on students experiences with their chemistry knowledge as acquired at secondary education. The second guideline directs the question designer in thinking of the more daily experiences that students have. In the food chemistry case study, questions could start by using examples of food that students typically consume.

The third guideline asks question designers to use facts, events, or conclusions that can motivate and inspire students. Again, for food chemistry, students in certain target populations are motivated for example by questions that relate to toxic effects or environmental pollution. Experience with these first three guidelines has been described in (Diederer et al., 2005) . Finally, it makes sense to use a common error or a common misconception as starting point for the design of a question. This method is elaborated in detail by Mazur (2001) with his ConcepTest approach.

Students	
F1	Imagine and use prior knowledge of the student.
F2	Imagine and use the experience of the student.
F3	Imagine and use the things that motivate and inspire students.
F4	Collect errors and misconceptions that students have.

4.2.7 G: Sources

In a wider perspective than already proposed in set A (Professional context) and set D (Textbooks), a set of guidelines was formulated to stimulate the systematic use of every possible information resource for inspiration. Five specific guidelines were formulated.

The first two guidelines call upon question designers to get informed by interviewing colleagues at the educational institution and professionals working in the field of the domain. A third guideline asks question designer to get informed by, or work with, educational technologists. They can inspire question designers not so much on content related aspects, but much more on the rules and techniques to design questions in general. A fourth guideline suggest that question designers should set up brainstorming or brain writing exercises and the like (Paulus et al., 2003). The goal of such a session is to come up with as much as possible questions and pointers towards possible questions without being restricted too much by all kinds of requirements, impracticalities, or even impossibilities. Restriction and convergence is dealt with in a later stadium. A fifth guideline proposes question designers to systemically collect as much as possible relevant information from sources outside their institution and outside their own social and professional network and in particular from sources that can be accessed over the internet.

Sources	
G1	Question colleague instructors of the faculty.
G2	Question professionals working in the field of the subject matter.
G3	Question educational technologists.
G4	Set up and execute brainstorm sessions.
G5	Collect information from various sources such as news papers, the internet, news broadcasts.

4.2.8 H: Motivation

Attention is a bottleneck in learning (Simon, 1994) and motivation is essential for effective and efficient learning. Keller (1983) formulated four variables that are important for motivation. Based on the variables 'direction giving requirements' are formulated that could inspire question designers. These requirements conform Keller's ARCS model (A: the question should captivate the Attention of the student, R: the question should be perceived as Relevant by the student, C: the question should raise the level of Confidence of the student and S: the question should raise the level of Satisfaction of the student).

So, motivation is regarded as a separate inspirational category. A question designer should try to design questions that meet the requirements given in this category. Only afterwards, it can be established whether a question meets the requirement.

Motivation	
H1	The question focuses the attention of the student for a sufficient amount of time.
H2	The question is experienced as relevant to the student.
H3	The question raises the level of confidence by the student.
H4	Answering a questions yields satisfaction by the student.

4.2.9 I: Validity

Validity in assessment is an important requirement. Tests and questions should measure what they are intended to measure and operationalise the learning objectives (criterion referencing). Because of their relation with learning objectives, validity requirements also give direction to the design process. Three direction giving validity requirements were formulated.

The first guideline reflects the requirement that questions need to measure the intended knowledge or construct that should be learned. The second guideline advises question designers to think more in terms of sets of questions to measure knowledge and skill than solitaire questions. The third guideline is actually a requirement to the test as a whole: in a test, the weight of a learning objective should be proportional to the number of questions measuring the knowledge and skills involved in that objective.

The scope of the ALTB project was limited to question design and not to design of complete assessments. Nevertheless, some of the guidelines clearly interface with design of complete assessments. Guidelines that tap into designing valid assessments and test are formulated in D (Textbooks) and E (Learning Objectives). These guidelines direct the question designer to layout the field of knowledge and skill to be questioned so that a good coverage of the learning material can be achieved.

Validity	
I1	The question is an adequate operationalisation of the learning objectives.
I2	The question itself is not an operationalisation of the learning objectives, but the set of questions is.
I3	Within a test, the weight of a learning objective is represented in the number of questions that operationalise that learning objective.

4.2.10 J: Equivalence

In higher education in general, tests and questions for summative purposes cannot be used again when they have been deployed. The reason for this is that assessments and test questions in general cannot be secured sufficiently and that subsequent cohorts of student would be assessed non-equivalent if they already have been exposed to the questions. Consequently, instructors need to design equivalent assessment and test questions to ensure that every cohort of students is assessed fairly and comparably. Four equivalence requirements were expected to function as not only a filter on questions but also as beacons that could direct the design process. These were equivalence with respect to content (subject matter), interaction type, cognitive process and finally also to scoring rules.

Equivalence	
J4.1	Equivalent in relation to subject matter.
J4.2	Equivalent in relation to interaction type.
J4.3	Equivalent in relation to level of difficulty and cognitive processes.
J4.4	Equivalent in relation to scoring rules.

4.3 Case studies to investigate the appropriateness of the developed guidelines

The use of the guidelines, has been observed in fifteen case studies. An overview of the case studies is presented in Appendix 1. Most case studies had a lead time of less than half a year. The case studies overlapped in time. Later case studies could make use of experience in earlier case studies. The cases mostly consisted of design projects for university level courses in which SMEs, their assistants and sometimes ETs, designed and developed digital closed questions to be used as summative exam material or activating learning material.

The question designers or teams of question designers (SMEs, assistants, ETs) were introduced to the guidelines in an introductory workshop. The function of the guidelines (i.e. inspire the question designers) was emphasized during these introductions, the how and why of the categories was explained and the guidelines were briefly discussed and illustrated with some additional materials. In the first workshop, the teams exercised in question design using those guidelines. Later on, during the execution of the projects, an overview sheet of the guidelines was at the disposal of the SMEs and assistants, any time they felt they wanted to use it.

The set of guidelines was formulated while the case studies WU1 and WU2 and the first part of TUD1 were running. The direction of the literature search for design guidelines was partly determined by projects on the design of digital learning materials that gave rise to the ALTB project and partly by these first three case studies.

Once the set of design guidelines was considered complete, all designer teams in the ALTB project were asked to start using the guidelines in all question design and development activities and to provide two reports.

For the first report the procedure was:

- Design and develop 30 closed questions as follows:
- For each question do:
- For each design guideline/direction-giving-requirement do:
- Record if it was useful;
- Record if its use is recognizable in the resulting question.

It was expected that this procedure would demand considerable discipline from the designers. Therefore, the number of questions that would be subjected to this procedure was limited to 30. The second report would be a less formal record of the experience of working with the guidelines for the remaining questions. A short report was made of every case. For most cases, data were recorded on the execution of the process and use or non-use of guidelines. In the Appendix 2, the major findings per case are listed.

In case studies VU1, VU2, TUD2, WU9 and WU10 – partly based on preliminary versions of both reports – ETs tried to support the designer teams in using the guidelines and described their experience.

4.3.1 Criteria for assessing the value of the guidelines

The research question of the ALTB project as stated in the introduction, can be mapped onto a research design consisting of multiple cases with multiple embedded units of analysis (Yin, 2003). A small set of units of analysis was identified. These units of analysis are: a set of design requirements, a set of design guidelines, a set of design patterns, a set of interaction types, a task structure, and resource allocation. As said, this chapter focuses on the development and evaluation of set of guidelines. What are the useful criteria to establish whether guidelines are a worthwhile component of a methodology?

First, within a methodology, guidelines form a worthwhile component if, for any given design team, the set of guidelines includes at least *five* guidelines the team can use. It is expected that the value of specific guidelines will depend on the specific domain, the competency of the question designers, and so forth and so on. However, a general finding that guidelines can support the design and development process must be answered positively.

Second, the ALTB team wanted to investigate how the development teams would and could work with the complete set of guidelines in practice. Is a team willing and capable of dealing with a fairly great number of guidelines and able to select the guidelines that are most useful for them? If teams are not able or willing to use such a large set, the guidelines themselves can still be useful. In that case however, suggestions should be put forward on how to present subsets of the guidelines to make guidelines a functional instrument.

Third, a methodology for the design and development of closed questions must in principle be as general applicable as possible. As closed questions are used in both summative tests and activating learning material, it is worthwhile to examine the assumption that one set of guidelines can be used equally well for both roles. Maybe however, given the intended role for question design, different sets should be offered upfront in a development project.

4.3.2 Observations

Execution of the method

One team of question designers declined to work with the set of design guidelines. This team was involved in a transition from learning objective oriented education to competency directed education. The goal for this team was to design and develop diagnostic assessments. The team argued that the guidelines had a too narrow focus on single questions instead of on clusters of questions. Furthermore, this team expected that the guidelines would prevent creativity instead of boosting creativity. This team proposed to start developing questions without any guideline and abstract later from their behaviour a set of guidelines. De facto, it turned out that this team focussed completely on guideline A1. The resulting questions however did not reflect their efforts in developing cases. Furthermore, the questions did not reflect the philosophy of competency based education. A number of questions had feedback that consisted of closed questions. No other guidelines came out of this case study.

All other teams were initially positive about performing the two tasks. However, it soon turned out that rigorous following the procedure was more difficult than expected.

Two teams (VU1 and VU2) tried to execute the procedure but got entangled in a discussion on the appropriateness of the guidelines. This caused them to lose track of the procedure. As a result no careful record was produced. However, these two teams did produce a number of closed questions on the basis of the guidelines. All the other teams produced a record of the thirty-question-procedure.

A final general observation, which will be discussed in more detail in Chapter 5, is that budget estimations were too low for all cases. The design and development of questions took three to four times the amount of time that was budgeted based on previous reports.

Use of the guidelines

The developed set of guidelines was actively used by all teams but one. Browsing through the guidelines and discussing them made SMEs and assistants aware of multiple ways to start and execute the conception of closed questions. Within the set, there were always four or five guidelines available that in fact helped question designers to find new crystallization points for question design they had not thought of before.

In VU1, VU2, TUD2, WU9, WU10, SMEs were of the opinion that categories B (Interactions) and C (Design Patterns) often resulted in questions that were new for the intended subject matter. Example questions, presented by the ET (often devised by the ET on the basis of preliminary information, textbooks or identified within other sources such as the internet), or questions stemming from previous developed tests, quickly invoked conceptual common ground between SME, assistant and ET. This common ground enabled the assistant to apply the core idea of the given example to questions within the intended domain. It was also noted that this effect was the strongest when the example questions were as closely as possible linked to the intended domain.

The guidelines to use digital media (B3x, D1.1, D1.2 and D1.3) in the form of photos, graphs, diagrams, and chemical structures and so on, turned out to be a worthwhile guideline for the majority of teams. Systematic focus in the design process to use such media was regarded as useful and led to new questions for the teams.

For the design and development of summative exams, category J (Equivalence) turned out to be a dominant guideline. This is due to the fact that for summative exams a representative coverage of a larger number of detailed learning objectives is necessary and that re-exams should be as equivalent as possible as long as the learning objectives do not change.

Given the observation that the guidelines in category J were not tangible enough, a new guideline for that role was formulated. This guideline advises question designers to aim directly at a cluster of five equivalent questions for each detailed learning objective, textbook paragraph or image by making variations on one question. This guideline is phrased as: *design and develop clusters of five equivalent questions*. Making slight variations on one question (paraphrasing, changing responses orders, splitting up multiple-choice question in variations of 2, 3 or 4 alternative questions, using different examples, questioning other aspects of the same concept, varying the opening sentences) will cost relatively little effort as compared to designing and developing a new question.

General critique in the case study reports regarding the set of guidelines

Many question designers were of the opinion that the presentation of the complete set of design guidelines made them see the wood for the trees. SMEs and assistant repeatedly called for "Give me only the guidelines that really can help me". Presenting the complete set resulted in a lower appreciation for the guidelines as a whole.

At the same time, a number of guidelines were regarded as 'too obvious' by SMEs and assistants or were regarded as variations of the same guideline. This counts especially for guidelines Professional context (A), Textbooks (D), Learning Objectives (E), Validity (I) and Equivalence (J). Of course, the perceived usefulness of a guideline is in practice related to the extent to which a guideline is new for a designer/developer. However, declaring any guideline that is well known, as useless, is in our opinion not a valid reason to exclude it from the set of guidelines. However, this perception of the guidelines by SME's and assistants also results in a lower appreciation for the guidelines as a whole.

Limitations regarding specific guidelines

Often the SMEs and assistants could formulate why they had not used a *specific* guideline. The first general reason for that was that it was unclear how a specific guideline operates. SMEs and assistant simply did not always see *how* to use certain guidelines. For instance H1, the directional requirement to capture and hold the attention of the student, induced the designers to ask: "Yes but how?"

With respect to categories B (Interactions) and C (Design Patterns), the case studies supported the idea that common available question examples (stemming from secondary education) lead SMEs and assistants too quickly come to conclude that "such questioning is not suitable for use in higher education". The content and perceived difficulty of such questions make it explicitly necessary to discriminate between the actual example and the concept underlying such examples to see their potential for use in higher education. That calls for extra mental effort and time, which often is not available in practice.

Once new design patterns became available, the case studies in the last stages of the project revealed the value of design patterns: design patterns can have a greater impact on the conception of innovative digital questions than general guidelines and therefore should receive more attention in the methodology.

Secondly, certain guidelines were perceived as incurring additional costs, which were not balanced by the expectation of additional benefits. For instance, developing a case or a video and using it as the foundation for a question was said to involve too much effort in comparison to the expected benefits. This effect was increased by the fact that most project budgets were underestimated which sometimes was given as a reason to restrict design and development to the more simpler question formats (simple, text based Multiple-choice (MC) questions) and not actively work on more elaborate design activities (such as A2, E3 or G), question types and media use. At the same time the formulation of distractors for traditional text based MC questions was in some case studies reported as being very time consuming in comparison to other design and development tasks and guidelines to avoid having to develop distractors were called for.

Thirdly, in a number of case studies, the SMEs and assistants were of the opinion that a specific guideline was not relevant given the subject matter or that a certain guideline 'did not fit the purpose of the exam'. For example, physiologists stated that contradictions in their subject matter 'do not exist' (though of course they could design questions that use contradictions as foil answering options for example).

Fourth, in a number of case studies, the SMEs and assistant were of the opinion that the role of the question (summative or activating) did not allow to use a specific guideline. In particular, for summative exams, Category B (Interactions) invoked, in a number of case studies, discussion on the scoring models of specific question types. How should questions involving multiple possible responses (such as Multiple Answer question, Matching questions, and Ordering questions) be scored? This uncertainty made SMEs and assistant decide not to pursue the design of such questions.

Summarizing: *specific* guidelines were perceived to have different value depending on the subject matter, the role of the questions, time constraints and the competencies of the designers. Reasons not to use a specific guideline can be categorized under the following labels:

- Directions on how to use the guideline are lacking given the available team knowledge and skill.
- Cost-Benefit estimations of using the guideline were too high given the project conditions.
- The guideline is not relevant given the subject matter.
- The guideline is not relevant given the role of the questions. The guideline cannot be used until the question about transparent scoring is resolved.

Intervention and input of the educational technologist

In case studies VU1, VU2, WU9, WU10 and TUD2, an ET helped the SME and assistants to gain more benefit of the guidelines by extra explication and demonstration and by selecting guidelines that could be most beneficial given the project constraints. Moreover, the ET could actually take successful part in the idea generation process when sufficient and adequate learning materials were available. In particular, the incorporation of various media in question design could be stimulated by the ET. When insufficient learning materials were available, it was very difficult for the ET to contribute to the design and development process. Thus, the actual involvement of the ET with the subject matter and the availability of learning materials is an important context variable for a successful contribution of an ET.

4.3.3 Evaluation of the set of guidelines

As said, this chapter focuses on the development and evaluation of set of guidelines for question design.

The case studies have confirmed that for the majority of teams, four to five guidelines are used and are perceived as worthwhile. Given the criterion that for a methodology, for any given team, a minimum of five guidelines must be useful, it is fair to conclude that the set of guidelines is a useful component within a methodology.

Second, the ALTB project wanted to investigate if question development teams can work with the complete set of guidelines in practice. From the case studies it becomes evident that this is not the case. Simply presenting a set of guidelines had only very limited effect on the process. Offering some modest training and support increased the effect, but not substantially. It truly calls for a considerable effort by the team members for the guidelines to really have an impact on the quality of the design process and the quality of the questions that are developed. Most teams wanted a preselected set of three to five guidelines exactly targeted to their situation without having to select those themselves.

The third criterion that most of the guidelines would be applicable, irrespective of the intended role of the questions (summative or activating), is not met by the set of guidelines. Designing questions for the specific roles calls upfront for different sets of guidelines. A major discriminating factor for this is that for summative exams there is a lack of clear scoring rules for innovative question types and that emphasis is put on effective ways to develop multiple equivalent questions. For activating learning material, transparent scoring is less important and more emphasis must be put on engaging the learner more with the subject matter. In that respect, it is actually beneficial to use a wide variety of innovative closed question types.

4.4 Conclusions

Literature provides little guidance for the initial stages of design and development of digital closed questions. This is an important reason to conduct research in these stages and develop specific tools to support the initial design process. One tool that is developed in the ALTB project is a set of guidelines focussing on the initial stages of design and development in order to boost creativity. This set of guidelines was presented to question design teams and used in 15 case studies. These case studies have been described and summarized in this chapter.

4.4.1 A set of guidelines is an inspirational source for question design but must be embedded in a broader approach

The developed set of guidelines offers inspiration to the majority of teams. There are always four or more guidelines available in the set that help question designers to find inspiration for question design. Within a broader methodology, the guidelines will certainly be appropriate. From the case studies it is concluded that different set of guidelines should be compiled for the summative role or the activating role of questions. In the future, more and different guidelines will without doubt emerge for the specific roles.

Furthermore, it has become clear that guidelines cannot function on their own. Design and development of digital closed questions requires specialized knowledge and skills. That can only be acquired through thorough study and practice. SMEs and assistants need support to interpret and use the guidelines effectively. In particular SME's and assistants need help in selecting those guidelines which are most useful for them in their situation. Without such help, they loose focus and become frustrated.

4.4.2 Design patterns have the potential to be a powerful aid

The case studies revealed the value of design patterns: design patterns can have a great impact on the creative design of digital questions. They can be more effective than general guidelines or too general question examples. In Chapter 6 a more detailed description of the concept of design patterns and a number of design patterns are presented.

4.4.3 A question design methodology must be geared towards educational technologists

Given the observed intricacy of question design and development, the conclusion is drawn in the ALTB project that a methodology must be geared specifically towards ETs. They must be able to use guidelines and design patterns in a variety of situations and domains to support SMEs and assistants. A methodology should help an ET to select a few specific guidelines and a number of adequate design patterns in order to produce quick and effective results when working with SMEs and assistants. The question of what procedures ETs can best act upon to perform that task is a matter for further research.

4.5 References

- Aegerter-Wilmsen, T., Coppens, M., Janssen, F. J. J. M., Hartog, R., & T.Bisseling. (2005). Digital learning material for student-directed model building in molecular biology. *Biochemistry and Molecular Biology Education*, 33, 325-329.
- Alexander, C. (1979). *The timeless way of building*. New York: Oxford University Press.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.
- Bloom, B. S. (1956). *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain*. New York: McKay.
- Bull, J., & McKenna, C. (2001). *Blueprint for Computer-assisted Assessment*: RoutledgeFalmer.
- Dick, W., & Cary, L. (1990). *The Systematic Design of Instruction*. (Third ed.): Harper Collins.

- Diederer, J., Gruppen, H., Hartog, R., Moerland, G., & Voragen, A. G. J. (2003). Design of activating digital learning material for food chemistry education. *Chemistry Education: Research and Practice*, 4(3), 353-371.
- Diederer, J., Gruppen, H., Hartog, R., & Voragen, A. G. J. (2005). Evaluation of computer-based learning material for food chemistry education. *Chemistry Education Research and Practice*, 6(2), 64-82.
- Haladyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*: Needham Heights MA, Allyn and Bacon.
- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (Third ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education* 1, 15 (3), 309-334.
- Hartog, R. (2005). ALTB website. Retrieved march 20 2007, from <http://fbt.wur.nl/altb/>
- Keller, J. M. (1983). *Development and Use of the ARCS Model of Motivational Design*. (No. IR 014 039). Enschede.: Twente Univ. of Technology.
- LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modelling procedure for constructing content-equivalent multiple-choice questions. *Medical Education*., 20(1), 53-56.
- Mazur, E., & Crouch, C. H. (2001). Peer Instruction: Ten Years of Experience and Results. *American Journal of Physics*., 69(9), 970-977.
- Merriënboer, J. J. G. v., Clark, R. E., & Croock, M. B. M. d. (2002). Blueprints for Complex Learning: The 4C/ID-Model *EIR\$D*, 50(2), 39-64.
- Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-Based Testing, Building the Foundation for Future Assessments*. London: Lawrence Erlbaum Associates.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York,: Springer-Verlag.
- Paulus, P. B., & Brown, V. R. (2003). *Enhancing ideational creativity in groups: Lessons from research on brainstorming*. . Oxford: Oxford University Press.
- Roid, G. H., & Haladyna, T. M. (1982). *A Technology for Test-Item Writing*. Orlando, Florida: Academic Press.
- Roossink, H. J., Bonnes, H. J. G., Diepen, N. M., van, & Moerkerke, G. (1992). *Een werkwijze om tentamenopgaven te maken en tentamens samen te stellen* (No. 73): Universiteit Twente.
- Scalise, K., & Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment* 4(6).
- Simon, H. A. (1994). The bottleneck of attention: connecting thought with Motivation. In W. D. Spaulding (Ed.), *Integrative views of motivation, cognition and emotion*. (Vol. 41, pp. 1-21). Lincoln: University of Nebraska Press.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wilbrink, B. (1983). *Toetsvragen schrijven* (Vol. 809). Utrecht/Antwerpen.
- Yin, R. K. (2003). *Case Study Research : design and methods* (3rd ed.): Thousand Oaks (CA) Sage.

APPENDIX 1 - OVERVIEW OF CASE STUDIES

Case	Course Level	Course Subject	Role of the questions	Software	Development team
WU1	Master	Food Safety (Toxicology/ Food Microbiology)	summative	QM	SME and assistant
WU2	Master	Food Safety Management	activating	Bb	SME and ET
VU1	2nd year	Heart and Blood flow (physiology, ECG measurement and clinical ECG interpretation)	diagnostic and summative	QM	SME and ET
VU2	3rd year	Special Senses (vision, smell, hearing, taste, equilibrium)	summative	QM	SME and ET
TUD1	3rd year	Drinking water treatment	activating	Bb	SME and assistant
WU3	Master	Epidemiology	summative (open book)		SME and assistant
TUD2	3rd year	Sanitary Engineering	activating	Bb	SME and assistant and ET
WU4	Master	Food Toxicology	summative	QM	SME and assistant
WU5	Master	Food Micro Biology	activating	Bb	assistant
WU6	Master	Advanced Food Micro Biology	activating	Bb	assistant
WU7	Master	Food Chemistry (general introduction module for candidate students)	diagnostic	QTI delivery	SME = ET
WU8	Master	Food Toxicology	diagnostic	QM	SME and assistant
WU9	Master	Sampling and Monitoring	diagnostic (self -)	Flash	SME and Assistant and ET and Flash programmer
WU10	Master	Food Safety Economics	summative (not open book)	Bb and on paper	SME and assistant and ET
FO1	1st year	Curriculum: General Sciences	Diagnostic-'plus'	N@t-school	SMEs and question entry specialist

The numbering of the case studies is an indication of the point in time when the case studies were carried out. Column two represents the institution in which the case took place. Column three indicates the course level and column four the course subject. The fifth column depicts the role of the questions within the course: summative, (formative) diagnostic or (formative) activating. Column six lists the authoring software that was used in the last column lists the main actors within the development team.

WU = Wageningen University, VU = Vrije Universiteit Amsterdam, TUD = University of Technology Delft, FO = Fontys University of Applied Science, QM = Questionmark Perception, Bb = Blackboard LMS, QTI = Question and Test Interoperability 2.0 format, N@tschool = N@tschool LMS, SME = Subject Matter Expert such as lecturer, professor, instructor, ET = Educational technologist, Assistant = recently graduated student or student-assistant

APPENDIX 2 - OVERVIEW OF CASE AND THE USE OR NON-USE OF GUIDELINES

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
WU1	<p>Role : summative</p> <p>Team : SME Assistant</p>	<ul style="list-style-type: none"> • Toxicology Part • Lecture notes • Handouts of Presentations • Detailed learning objectives in natural language • Food Microbiology Part • Handouts of Presentations • Articles 	<ul style="list-style-type: none"> • E1, C1 	<ul style="list-style-type: none"> • Given the intended role and task of the designer the need of guidelines for design became very apparent. • A comprehensive overview of guidelines which are useful in the domains of the ALTB project at the level of higher education could not be found. • For summative testing the contour of a new guideline became visible: next to designing one question, design 4 equivalent questions using the guidelines for 'parallel design and development' • Useful guidelines for 'parallel design and development' are <ul style="list-style-type: none"> ○ E1 Use a list of detailed learning objectives ○ C1 Use a list of generic item shells • Remarks: • The guidelines E1 and C1 came available during the introductory workshop that the assistant attended.

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
WU2	Role : activating Team: SME assistant	<ul style="list-style-type: none"> • Documents and reports • Examples of Cases and questions in Blackboard • Experience in the team with guidelines for activating learning materials • Literature on guidelines for the design and development of activating learning materials • Guidelines A1, C1, C2, (design patterns), G (scan sources) 	<ul style="list-style-type: none"> • No conscious use of guidelines • Implicit use of A1 	<ul style="list-style-type: none"> • Designer/developer gave most attention to development of cases and to formulation of extended feedback. • The most pressing need felt by the designer/developer was not the need for design guidelines • The designer/developer needed more and better sources on more subject matter knowledge and input with respect to professional experience • The bare availability of guidelines is not sufficient to induce the use of guidelines.

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
VU1	<p>Role : diagnostic and summative</p> <p>Team: SME ET</p>	<ul style="list-style-type: none"> • During the inspiration session, no material was available. • Later on, material was available in the form of: • Previous Exams • Physiology textbook • Complete set of guidelines was available 	<ul style="list-style-type: none"> • The following guidelines were not used: A2, C3, C5, C6, F. • All other guidelines were used. 	<ul style="list-style-type: none"> • All guidelines were systematically discussed and 'forced-fitted' to use in two rounds of 'inspiration sessions' in which an ET guided a question design session. • The subject matter and the learning objectives allow for the definition of authentic cases and authentic 'what to do' questions. Thus, the instructor was already used to apply guideline A1. Guideline A2 was evaluated as too labour intensive to execute and not appropriate for the course. The SME was of the opinion that guidelines A3.1 to A3.2 actually defined instructional content and should not define exam content. Guidelines A3.4 and A3.5 provided some inspiration. Guideline A4 could be used. • Guidelines B1, B2, B3.1 really invoked enthusiasm. Example questions presented by ET resulted in ideas on new questions. However, problems with unclear scoring rules diminished enthusiasm. • C1 was felt to be very useful too, but so straightforward that it was not used during the inspiration session. C2 looked promising but turned out to be difficult to handle. C3, C5 and C6 were not regarded as useful because it was felt to be difficult to develop univocal problems and answer sets. However, if the questions were intended for active learning, the SME was of the opinion that they were very useful. C4 offered opportunity for question generation. G (search for extra sources on the internet) was very worthwhile for the instructor, based on the extra source the Educational technologist retrieved for him). It resulted in a collection of pointers to useful cases, graphics and multimedia elements. • Guidelines F (take mindset of student as starting point) were not used because the instructor was of the opinion that any assumption about the mindset of the students would apply to a very limited part of the student population and would introduce bias. • Directional requirements H were not used. They were considered relevant, but not helpful. ("aim for attention – yes but how") • Guidelines D (textbooks) was considered an 'too obvious' ("how else can you start developing questions") • Directional requirements E (learning objectives), I (validity), J (equivalence) were felt to be 'too obvious' also. They were used all the time but were not considered to provide inspiration. • G3 and G4 were used in the form of the 'inspiration session'.

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
				<ul style="list-style-type: none"> • The instructor preferred to be offered a much smaller dedicated selection of guidelines. Also the overlap between guidelines should be avoided. • Bottom line: • Offering guidelines to question designer in an intensive inspiration session results in questions of types that are new for the course and for the SME • Especially discussing example questions is considered worthwhile. • The ET is an enabler for a greater divergence of questions conceived

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
VU2	<p>Role : summative</p> <p>Team: SME ET</p>	<ul style="list-style-type: none"> • During the inspiration session, no material was available. • Later on, material was available in the form of: • Previous Exams • A course website with digital materials and cases. • The complete set of guidelines was available. 	<ul style="list-style-type: none"> • The following guidelines were not used: A2, C3, C5, C6, F4 and H. • All other guidelines were used. 	<ul style="list-style-type: none"> • All guidelines were systematically discussed and 'forced-fitted' to use in two rounds of 'inspiration sessions' in which an ET guided a question design session (see also case VU1) • Guidelines result in new types of questions as in case VU1. • Comments about the use of authentic cases as in case VU1. This SME normally develops cases as follows: medical specialists deliver questions; the SME edits them and combines them in such a way that a case is the result. • B1, B2, B3 were felt useful, but would not be used by the instructor unless she could rely on the sustained support and input of the ET. • The assessment of the guidelines C, G, D, E, G and I and J was similar to that of case VU1. • With respect to F (students' mind set): The instructor was already used to design questions that relate to students daily life and experiences • The instructor felt that requirement H (motivation) was not really necessary, though in practice she actually used it to 'spice up' the final exam (and that is guideline F).

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
TUD1	<p>Role : activating</p> <p>Team : SME assistant</p>	<ul style="list-style-type: none"> Textbook with many photos, graphs, diagrams, examples, explicit calculations, exam questions with answers Hand-outs of Presentations Hand-outs of Lecture Notes The complete set of guidelines was available 	<ul style="list-style-type: none"> H1, H2, C4 and D1.8 B2 and D1.2 were used most by the assistant. A1, B3.5, D1.10, and I3 were used most by the SME. 	<ul style="list-style-type: none"> Guidelines A* were not used by the student assistant because she did not have sufficient professional experience and because the SME could 'take' the tasks that are related to these guidelines. Guidelines A1* on cases were not used because the SME wanted to cover all subject matter The main determinants for the use of specific guidelines were <ul style="list-style-type: none"> the role of the questions, the extent of professional experience, the characteristics of the subject matter. The use of a number of guidelines can be recognized but the case study did not provide positive evidence about any added value of presenting a set of guidelines to the designers/developers. Bottom line: Many guidelines were considered 'too obvious' For almost every guideline that was not used there was a good reason not to use that guideline. Guidelines that cannot be used in a specific design and development project for a good reason should not be offered in that project. Systematically scanning inspirational dimensions did not work

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
WU3	<p>Role : summative (open book)</p> <p>Team : SME assistant</p>	<ul style="list-style-type: none"> • Textbook • Hand-outs of Presentations • A large set of MC questions, mostly based on 2 propositions • The complete set of guidelines including initial experience with the guidelines 	<ul style="list-style-type: none"> • J 4.1 • C1, C2,C3,C4, C7, D1i, G5 	<ul style="list-style-type: none"> • The directional requirement to design a set of equivalent questions for each detailed learning goal was considered to be crucial. • Textbook (guidelines D) and other sources like internet and journals (guideline G5) were scanned for inspiration. • Guidelines C3 and C4 were relatively useful for design and development of questions of a different format. • Guideline I was used unconsciously whenever the questions were discussed with the SME. • Main conclusion: • The guidelines do hardly result in new question types for the course/instructor • The guidelines do hardly result in quicker or more efficient design of questions • Remark: The summative test is an open book exam, which made it more difficult to design questions. Developing questions which are directly based on text of the book is not an option; questions needed to be formulated in a different way or should test application.

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
TUD2	Role : activating Team : SME assistant ET	<ul style="list-style-type: none"> Textbook with many examples, graphs, open questions. Exam questions, answers to questions The textbook was authored by the chair group sanitary engineering. Also the pictures in the textbook were available electronically Additional handouts of presentations Lecture notes Relevant Websites The complete set of guidelines was available. 	<ul style="list-style-type: none"> C2, C3, C4, and Ci where i denotes any new design pattern that was not yet listed Di where i denotes any of the textbook components or questions inspired by textbook components E was used implicitly as the textbook covered E. G3 (ET) 	<ul style="list-style-type: none"> Focus on design patterns results in new questions and more use of question types other than True/False and MC Guidelines A1 and A2 are not considered because cases are supposed to direct too much attention of the student to a small part of the subject matter that has to be covered according to the definition of the course. As it was agreed that the consultant would take the lead also A3 did not get much attention B1 had already been done in the previous project Once more scanning B2 was not inspiring B3.2 (sound) and B3.4 (video) were not considered because of capacity constraints A number of new design patterns were used. These patterns will be presented in a publication on design patterns. D9 (abstract concepts) and D10 (what to remember) were not considered F (prior knowledge of student as starting point) was not considered useful by both the lecturer and the question designer G1,2,4,5 were not used because of time constraints H was not considered useful by the lecturer and the question designer I was used implicitly whenever a suggestion of the consultant had to be discussed. I was also implicit in the textbook J is not relevant for activating learning material Presenting design patterns and focussing on design patterns was much more effective in generating a variety of innovative questions than presenting guidelines or inspirational dimensions. The design patterns sometimes 'use' one guideline but often 'use' more guidelines C1, C2 and J4.1 were felt to be useful to create equivalent exams. The guidelines D were used in the sense that the learning material is scanned for inspiration. Directional requirements F (students), H (motivating) and I (validity) are used but are not considered to provide inspiration. Remark: The exam was to be digital. Technical and organisational aspects required much attention of Question Designer as well
WU4	Role : summative Team : SME assistant	<ul style="list-style-type: none"> Lecture notes Hand-outs of presentations Articles 	<ul style="list-style-type: none"> C1-3, D1i, J4.1 	<ul style="list-style-type: none"> C1, C2 and J4.1 were felt to be useful to create equivalent exams. The guidelines D were used in the sense that the learning material is scanned for inspiration. Directional requirements F (students), H (motivating) and I (validity) are used but are not considered to provide inspiration. Remark: The exam was to be digital. Technical and organisational aspects required much attention of Question Designer as well

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
WU5	Role : activating Team : assistant	<ul style="list-style-type: none"> Textbook Handouts of presentations 	<ul style="list-style-type: none"> C1 and C3, D, E 	<ul style="list-style-type: none"> The guidelines D were used in the sense that the learning material is scanned for inspiration. Guidelines concerning the interaction types (B) were used unconsciously as already a lot of experience had been gained by developing other questions. The guidelines F (students), H (motivating) and I (validity) are seen as important issues that require attention but that are not concerned to provide inspiration. ("Yes but HOW") J is not relevant for activating learning material.
WU6	Role : activating Team : assistant	<ul style="list-style-type: none"> Handouts of presentations Articles 	<ul style="list-style-type: none"> C1 and C3, E, G1 and G5 	<ul style="list-style-type: none"> Guidelines concerning the interaction types (B) were used unconsciously as already a lot of experience had been gained by developing other questions. The guidelines F (students), H (motivating) and I (validity) are seen as important issues that require attention but that are not concerned to provide inspiration. J is not relevant for activating learning material. As there was no textbook guidelines D were not really helpful, but instead guidelines G1 and G5 were.
WU7	Role : diagnostic Team : SME = ET	<ul style="list-style-type: none"> Textbook many examples of closed questions for Food Chemistry in FLASH though often not specifically for exactly the same subject matter 	<ul style="list-style-type: none"> Guidelines that were mainly used : B1, B2, B3, C3, C4, C7, D1.i except D1.7, E1, E2, E3, F.i, , G1, G3, G5, H2, H4, I1 and I3 	<ul style="list-style-type: none"> The SME/ET could clearly explain why she did not use the following guidelines: A1 (cases) was difficult to match with the test matrix A2.i (LaDuca) did not match the purpose of the diagnostic test A3.1 (tips, tricks) did not match the purpose of the diagnostic test A3.2 (surprise in profession) incidentally provided inspiration C1 and C2 did not match the purpose of the diagnostic test C1 and C2 are actually not very useful unless one wants to develop a set of exams

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
WU8	Role : diagnostic Team : SME assistant	<ul style="list-style-type: none"> Detailed list of learning objectives Lecture Notes Handouts 	<ul style="list-style-type: none"> New guideline "cluster of five", E2, I1 	<ul style="list-style-type: none"> C5 and C6 (for designing and developing text based questions) did not match very well the subject matter D1.7 (exceptions) did not help at all. In the related courses it is not usual to pay attention to exceptions E4 (target competencies) was not yet useful because the target competencies are only defined at curriculum level and articulating them at the course level is considered to be a task that does not fit within the scope of the project. Fi (students) were all used but F1 and F2 more than F3 and F4 G2 (ask content experts) and G4 (brainstorm sessions) were not used because not within the budget. H1 (gain attention) and H3 (aim for confidence) did not strongly match with the purpose of the questions I2 was not used Bottom-line A very experienced designer can use about two thirds of the guidelines and can give a clear explanation of any reasons not to use a specific guideline. Content Expert already had gained some experience in case WU1 Quickly decided to focus on MC, MA, ordering, match and fill-in-the-blank and not to use any diagrams or pictures. Subject matter does not require such diagrams Quickly decided to use new guideline ("design and develop cluster of five equivalent questions approach") Questions were designed in MS Word, later formulated by technical assistant in QTI 2.0 Most design guidelines were not used Initial confrontation with the complete initial set of guidelines resulted in very limited use On basis of that it was agreed to focus on the following subset : B2 interaction types – B 3.4 graphs – B 3.5 diagrams– B 3.6 process diagrams– C 3 completion - C 4 introduce error – D systematically scan learning material (self developed) – G2 ask food safety experts – G5 other sources – H1 capture attention E use detailed learning objectives
WU9	Role : diagnostic (self -) Team : Assistant ET	<ul style="list-style-type: none"> Scientific articles Learning Material that was designed and developed in parallel with the design of closed questions 		

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
WU10	<p>Role : summative (not open book)</p> <p>Team : SME Assistant ET</p>	<ul style="list-style-type: none"> • Lecture Notes • Articles • Handouts of Presentations • The handouts include many diagrams and graphs and other pictorial information • The handouts include many procedures and computations • Computer Practical instructions 	<ul style="list-style-type: none"> • Guidelines that were mainly used: A1 and A4, C3, D 	<ul style="list-style-type: none"> • Together with an educational technologist, new design patterns were developed • The educational technologist presentation that covered most of the subject matter and this presentation contained a wealth of diagrams and figures to be used as foundation for closed questions • New design pattern: match symbols in a given equation with data in a given problem description. Thus understanding of operational semantics of an equation can be separated from the ability to execute a calculation • Technical implementation was delegated to a FLASH programmer. • Questions developed in MS Word and MS PowerPoint • Focus by ET on design patterns (guidelines C) that imply the use of pictures • Not limited to the few design patterns that were initially available. Result: Many more design patterns were conceived. • Preliminary conclusion: • The combination of: <ul style="list-style-type: none"> ○ availability of many digitized diagrams, graphs and other pictures ○ many computations and corresponding chains of inference ○ many questions ○ high degree of involvement of the content expert/instructor • is in keeping with the hypothesis that - the more conditions are satisfied the more guidelines are useful and the better a condition is satisfied the more one tends to focus on the guidelines that match this condition • In this case study many PowerPoint slides formed an obvious basis for a question. • In particular application of guidelines D in combination with C and some new design patterns was effective. • Guidelines A1 and A4 were followed to develop cases. A2 and A3 were not useful as the question designer did not have practical experience. • I was used unconsciously whenever the questions were discussed with the content expert

Case	Role / Development team	Initially available material	Which Guidelines used And How	Summary of case report
FO1	Role : diagnostic Team : SMEs	<ul style="list-style-type: none"> Textbook(s) 	<ul style="list-style-type: none"> A1 	<ul style="list-style-type: none"> Only guideline A1: (develop cases) was used When the initial set of guidelines was presented representatives of the team indicated that they would not adopt these guidelines Fundamental critique was that the presented guidelines suggested too much focus on individual questions instead of sets of questions that the set of guidelines killed creativity It was agreed to develop 30 questions and record what alternative guidelines were actually used. <p>The team however did not succeed in formulating any alternative guideline.</p>

5 Design of Digital Closed Questions: Procedures for Deciding when to Use which Guidelines

Rob Hartog
Wageningen University

Silvester Draaijer
Vrije Universiteit Amsterdam

Abstract

Based on experience with a large set of guidelines for the design and development of closed questions in higher education, procedures for handling this large set of guidelines have been developed. The procedures should help educational technologists to select adequate guidelines and apply them at the right moment, based on an assessment of the design and development context.

5.1 Introduction

Literature has been rather implicit with respect to guidelines that focus on the initial stages of design and development of innovative closed questions. Insofar such guidelines could be found, they were found in different locations. Chapter 4 presented a comprehensive table of guidelines focussing on the initial stages of design and development that was extracted from literature and was in 15 case studies presented to design teams. The case studies made clear that design teams could not adequately handle such a complete set. Most teams wanted a carefully selected set of three to five guidelines matched to their situation. Based on the experience in the case studies, procedures have been developed that help ET's to select adequate guidelines at the right moment.

In order for an ET to secure a successful outcome for a test and question design effort, (s)he should assess the context of the project and adapt his or her collaboration strategy with SME's and assistants accordingly. Based on the characteristics of contexts of projects for the D&D of closed questions, the next sub sections describe procedures for assessing such a context and selecting adequate design guidelines. The procedures focus on the two extremes along the question role dimension: the role of activating learning material (the ALM role) versus the role of Computer-based Assessment (the CBA role). A discussion of possible roles along a scale between these extremes requires more elaboration.

The procedures assume a realistic budget. Based on experience in the ALTB project such a budget assumes an average D & D time per question of about 2 hours. Ideally, there should be adequate division of labour and consequently the actual costs per hour will have to be calculated based on the actual division of labour. A rough estimate of the necessary budget for summative assessment following the 'aim at clusters of five' guideline can be based on the following line of reasoning.

If the exams for which we aim contain 60 questions per exam then the minimum project size will be $60 * 5 * 2 = 600$ hours. The number of clusters will – in practice - depend on the amount of subject matter and – in practice - on the amount of time scheduled for the exam. For the design and development of closed questions that will stimulate active learning a much smaller number of questions can already make sense.

5.2 A procedure for Design & Development of closed questions for the CBA role

This section describes a procedure for the design and development of closed questions aiming at summative assessment in higher education.

1. Assess in detail the possibilities and limitations of the CBA system that will be used for assessment.
2. Skip those question formats for which no accepted and transparent scoring model can be found.

Step 1 and Step 2 constitute a variation of guideline B1 in Chapter 4. The other guidelines in category B are not considered applicable anymore in projects that aim at the assessment role of questions.

The next steps should result in numeric or non-numeric labels for learning objectives. Each label of a learning objective will also be a label for a cluster of five. Although non-numeric labels may support easy look-up one should take care to keep the effort in formulating non-numeric labels very low (see also Chapter 7).

1. IF there is a list of detailed learning objectives:
THEN number each learning objective. The outcome of Step 3 is a set L_A of labelled learning objectives.
2. IF there is learning material that *clearly implies* the set of detailed learning objectives
THEN attach a label to each implied learning objective. Find a way to make the relationship between these learning objectives and their labels transparent. For instance, a label could be attached to one or more slides of a presentation. Alternatively, the label could be attached to a specific exercise in the textbook or to a specific graph. In any case, keep the effort of trying to construct a formulation of the learning objective in natural language to a minimum.
The outcome of Step 4 is a set L_B of labelled learning objectives.
3. IF previous exams clearly imply the learning objectives
THEN attach a label to every implied learning objective. Find a way to make the relationship between these learning objectives and their labels transparent. For instance, a label could be attached to one or more sub questions of different exam questions. Again, do not invest much time in trying to construct a formulation of the learning objective in natural language.
The outcome is a set L_C of labelled learning objectives.
4. The union of L_A , L_B and L_C is L . L is the set of learning objectives of which there is a shared understanding in the team.
5. For the design of closed questions for assessment Step 3 to 6, replace the guidelines E of Chapter 4 about learning objectives. An assistant should be able to execute step 3 to 6 with little guidance of the ET.
6. Now for each of the learning objectives in L , mark the learning objectives for which media inclusion is actually required or at least makes sense. One result is a subset M of learning objectives marked for media inclusion. The other result is a subset T of unmarked learning objectives that will probably lead to text-based questions.
7. For each learning objective in M , find photo's, graphs, diagrams and other media that are available in the textbook, the presentations and in other learning material. Also a screenshot of a situation of a computer simulation or the state of a computer program can be such an available media object.
In other words apply guidelines D1.1 and/or D1.2 and/or D1.3 of Chapter 4.
8. In addition to application of D1.1, D1.2, and D1.3, apply design patterns for photographs, diagrams and graphs. For example an 'introduce a mistake and let the student identify the mistake' pattern. Another example is the pattern that requests students to complete a diagram or a chemical structure. Select only design patterns that point at interaction types or question types that are allowed by the filter applied in which step 1 and step 2. In the ALTB project a number of design patterns have been described in detail (see Chapter 6) .

9. For each cluster, address the main concerns of the SME and the assistants:
 - a. IF the the SME and the assistant have major problems in finding distractors THEN avoid initially any design pattern that requires the formulation of distractors. Be aware that innovative question types can reduce development time because they not always require the 'conjuring up' of distractors. These might for instance involve Numeric calculated, fill-in-the-blank, Identify the mistake, Hot Spot. These question types still allow objective and transparent scoring.
 - b. IF the SME primarily wants to reduce guessing: THEN present also design patterns in which the number of distractors is large, for example Extended Matching (Case et al., 1994) or Glossary type answering options lists.
Note that design patterns that do not require distractors also reduce guessing.
 - c. IF the SME primarily wants to avoid implicit clues: THEN present design patterns of questions that give very little or no information. The former are patterns that provide very much data.
 - d. IF the SME only allows question types for which scoring rules are well established THEN use basic examples such as presented by Haladyna (2004) to abstract design patterns for the subject matter of the course AND present suggestions for the systematic generation of distractors such as the matching pairs approach and the cause – effect approach.
10. Finally, it might be attractive to search outside the available learning materials for media to be included or even to create media to be included. This requires considerable additional budget proportional to the number of questions for which additional media will be collected or created. This would imply application of guidelines B3.x and G5 of Chapter 4 .

While the arguments for presenting other guidelines or directional requirements of Chapter 4 still hold, the procedure presented here leaves little room for those guidelines. The procedure will produce not only questions but also a question matrix that covers all learning objectives.

5.3 How to use guidelines for Design & Development of closed questions for the ALM role

This section proposes an approach for the design and development of closed questions that are intended to stimulate active learning

5.3.1 No clusters of five and no complete coverage.

If the role of the questions to be designed is to function as activating learning material, the 'aim at clusters of five rule' does not apply. In addition, not every learning objective calls for closed questions. Furthermore, even a small set of closed questions only covering one topic can already be valuable. In practice however, the designers need to build up experience and routine before they arrive at a satisfactory level of quality. Thus in general, it is not a good idea that question designers only design and develop a small set of questions for one topic and never invest any more time in question design.

5.3.2 The two major situations

Two important distinctions between the combination of project contexts and goals must be made. In the first situation there is already learning material in the form of presentations and/or lecture notes and/or a textbook and the goal is to complement parts of this presentational learning material with stimuli for active learning. In the second situation, there is no such learning material and the goal is to create learning material that is directly and in itself **the** learning material.

5.3.3 Situation 1: ready available presentations, lecture notes and/or textbooks

1. Find topics and learning objectives for which the design and development will have a low cost/benefit ratio.
 - a. Looking at the benefits side, the need for 'at least something' to stimulate active learning might be stronger for the one topic than for the other topic. Closed questions that satisfy an urgent need, will have relatively high benefits. Walk through the learning material and identify the slides, paragraphs, photo's, diagrams, derivations and so on and so forth and decide where the need for additional stimuli to activate the student is most urgent.
 - b. For each identified element, look for a design pattern or a new opportunity to realize one or more closed questions to at relatively low cost. This requires cost estimation. The estimated costs will be determined by the project context and by the learning objective. For instance, the availability of design patterns for certain learning objectives will help to keep the costs low.
2. Step 1a and 1b should be summarized in an ordered list of elements that can be used as a basis for one or more closed questions. Step 1a and 1b constitute an adjusted version of the set of guidelines D*.* of Chapter 4. What is new, is the focus on cost/benefit ratio. In particular, the need to realize a few quick wins proved in several case studies in the ALTB project to be important in order to generate respect for the competence of the ET. Step 1b furthermore focuses on the guideline to use design patterns but with a new and larger set of design patterns.
3. Start from low cost and high benefit and work down the list:
 - a. Look for an opportunity to capture the student's attention. In other words, apply direction giving requirement H_1 of Chapter 4. In the ALTB project the question of SME's and assistants with respect to H_1 was often: "How to capture the students attention?". The scope of this chapter does not allow a full account of all possibilities to capture and hold the students attention. Four sub guidelines in this respect are: (i) provide surprise, (ii) minimize the effort that must be invested by the student before (s)he is rewarded for this effort in some way, (iii) make the relevance of the question explicit, and (iv) tell the students that an almost equivalent question will be in the summative assessment. Quick reward can for instance imply that the student immediately recognizes the issue that is the foundation of the question as well as extreme values that might play a role. One can compare this with newspaper headlines. Surprise can for instance be realised by taking a statement or picture that is counterintuitive but scientifically correct as the core of a question.
 - b. Look for an opportunity to ensure student's satisfaction. It is not likely that guessing behaviour will lead to real satisfaction. Rather, the student must perceive that he has accomplished something. This suggests the use of question types or interaction types and design patterns that give the student a feeling of accomplishment. The perception of having constructed a process by positioning unit operations in the right order using 'drag-and-drop' or the perception of having set up an experiment by filling in the missing step by using 'fill-in-the-blank' is likely to be more satisfactory than the selection of one choice by pressing a radio button.
4. In general, feedback should just be a reference to one ore more entries in the textbook and the learning material.

5.3.4 Situation 2: the set of closed questions is intended to be THE learning material

Note that there is nothing wrong with the intention to start developing learning material that stimulates students to active learning. Often, learning material just presents information and lines of reasoning and stimuli to do something with the information have to come from the teacher or from the student.

However if the set of closed questions to be designed and developed is intended to be THE learning material, then the ET cannot do much more.

First, if there is no learning material at all, the major contribution from the ET will be to provide design patterns including the corresponding examples. It might however not be clear how to make a selection of existing design patterns. Presenting a selection of design patterns that turn out to be useless is likely to reduce the confidence of the SME and assistants in the support of the ET. Most likely, the SME will have to make the selection of design patterns himself.

Second, in most case studies in the ALTB project, SME's and assistants were open to the direction giving requirements H* of Chapter 4 that aim to motivate the student. Thus, the ET should at least present these requirements. Again, the problem might be 'how to capture the attention of the students?' Presenting the sub guidelines given above will not help much without examples that are closely related to the subject matter of the course.

Third, the ET might comment on the initial products of the design and development team if such comments are appreciated. If so, the ET should try to assess the ARCS values of each question but also look for feedback that tends to be the beginning of a poorly designed set of lecture notes. In the ALTB project, no guidelines were given with respect to feedback. Case study WU2 (see Chapter 4) revealed however the danger of feedback beginning to start a life of its own leaving little time to the design of innovative questions. The ET might signal such a misdirected evolution and suggest developing feedback that consists of closed questions, but only if the available systems support this. This has been tried in case study FO1. In a trial, students were quite satisfied. However, the technical implementation of such questions in N@tschool involved relatively much effort. Elsewhere, also some success has been reported with this approach (Schaaf, 2007). One of the strong aspects of the QTI specification (IMS, 2005) is that it supports interactions of which the feedback essentially consists of interactions.

Finally, note that the guidelines in categories A, B, C, D and E in Chapter 4 are also still relevant in the sense that some of them will automatically be used by the SME but are not useful in the sense that they will be considered as helpful advice to the SME or assistant.

5.4 Concluding remarks and further research

It has been argued in this chapter, that developing questions for different roles requires different approaches. On the other hand, assessment modules of learning management systems and systems for computer-based testing are used both for testing as well as for providing students stimuli for active learning. Therefore, the practice in higher education will require educational technologists to be able to support projects with goals ranging from pure summative assessment to pure stimulation of activity. In addition, some of the subject matter experts who start developing closed questions for the purpose of stimulating students in active learning, will want to move on to using closed questions in exams. Further research should be aimed at the design development and evaluation of learning materials and a course for training ET's for their role in a design team. The following learning goals for educational technologists should have priority.

- Firstly, an ET should be knowledgeable both on the classic approaches to test and question design and instructional design.
- Secondly, the ET must be well informed on the opportunities and limitations of the assessment modules and functionality of LMS and CBT systems. In particular, the ET should be able to prevent discussions on the limitations of these systems so that valuable time can be invested in the design and development of questions.
- Thirdly, the ET should be able to assess the design and development situation and to select three to five guidelines for the design and development that can be presented to the SME's and assistants.

- Finally, the ET should have extended knowledge of digital closed question formats and should have access to a great number of examples of questions, question formats and design patterns. These examples and design patterns should of course match the selected guidelines.

The case studies revealed the value of design patterns: design patterns can have more impact on the creative design of digital questions than general guidelines and therefore should receive more attention in a methodology. Design patterns tend to give a more direct problem-solution relation. A designer needs many patterns. There are reasons to think of thousands of patterns (Alexander et al., 1977; Simon, 1996) even though no records of sets of thousands of patterns are known. Currently for design of closed questions, very few patterns are available. Experience in the ALTB project showed that scanning a set of about thirty patterns to find a match with a specific and detailed learning objective already turned out to be difficult when the patterns are only available on paper or even in a small database. Design patterns must be fully internalized. However, non-specialist question designers cannot be expected to have stored (or to store) a large number of suitable design patterns in memory. Therefore, ET's can prove their value by presenting the right design pattern at the right moment. In terms of required competency of the ET, this means an educational technologist must hold a large number of design patterns in memory to be able to offer an adequate design pattern in a discussion with a subject matter expert. Further research should aim at collecting and representing design patterns, making these patterns accessible and supporting ET's in constructing personal knowledge that integrates these design patterns.

5.5 References

- Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language : Towns, buildings, construction*. New York: Oxford University Press.
- Case, S. M., & Swanson, D. B. (1994). Extended matching items: a practical alternative to free-response questions. . *Teaching and Learning in Medicine*, 5, 107-115.
- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (Third ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.
- IMS. (2005). Question and Test Interoperability. Retrieved august 25 2006, from <http://www.imslobal.org/question/index.html#version2.0>
- Schaaf, H. v. d. (2007). *Design of digital learning material for bioprocess-engineering education*. PhD thesis, Wageningen University Wageningen.
- Simon, H. A. (1996). *The sciences of the artificial*. (3rd ed.). Cambridge: Mass.: MIT Press.

6 Design Patterns for Digital Item Types in Higher Education

Published as "Draaijer, S., & Hartog, R. J. M. (2007). Design patterns for digital item types in higher education. e-Journal of Instructional Science and Technology (e-JIST), 10(October)."

Silvester Draaijer
Vrije Universiteit Amsterdam

Rob Hartog
Wageningen University

Abstract

The ALTB project has produced a set of design patterns for digital item types in response to challenges identified in case studies on the design and development of digital closed questions by subject matter experts in higher education. The goal of the projects in question was to design and develop formative and summative tests, and to develop interactive learning material in the form of quizzes. The subject domains involved were mainly in the life sciences, medical sciences and engineering sciences. The use of digital item types and facilitating the process of designing items were typical examples of the challenges involved. From the viewpoint of subject matter experts, the main challenge in digital item type design was to design items that test for understanding. Furthermore, lecturers want to reduce student behaviour that is based on guesswork. With these conditions in mind, this chapter presents a set of design patterns for digital items, together with a standard format for describing these patterns.

6.1 Introduction

6.1.1 Focus on design patterns

This chapter presents one of the results of the ALTB project (Hartog, 2005). The aim of this project was to develop a methodology for the design and development of digital items. The methodology is intended to bridge the gap between currently available literature and the day-to-day work of designing digital items in higher education. A number of *design patterns* which were brought to light by this project, and which have now been incorporated into the methodology, are presented in this chapter.

Design patterns are intended to reduce the cost of designing and developing digital items. In addition they should enhance the validity of questions by reducing the chance that someone could arrive at the correct answer by means of guesswork and by enabling the intended objective to be measured more directly. In section 6.2, the concept of design pattern will be explained in more detail and applied to the design of digital items. A template for describing design patterns is presented. Its purpose is to support the design and development of digital items. A number of design patterns are also presented, together with arguments in support of their instructive value and versatility of purpose.

6.1.2 New opportunities for designing items for computer-based assessment and learning management systems

Currently available Computer-based Assessment systems (CBA) offer a great variety of digital item types (Bull et al., 2001; Mills et al., 2002; Parshall et al., 2002) such as multiple answer, drop-down lists, numeric, hot-spot, drag-and-drop. These systems also enable a variety of item types to be deployed within a single assessment. The availability of CBA systems and the Internet make it easier than ever before for Subject Matter Experts (SMEs – professors, academics, lecturers, tutors, instructors) to use such innovative item types. Also, other digital options can be used such as the inclusion of images. Several authors have referred to these item types as innovative. SMEs in many higher education courses are already using digital item types that are made available via CBA systems and Learning Management Systems (LMSs). One recurring problem, however, is how to make optimal use of these new possibilities.

6.1.3 User roles in designing digital items for higher education

Within the field of higher education, digital test items are usually developed within the context of a course taught by SMEs and their assistants. In general, it must be assumed that SMEs and their assistants have limited time for designing and developing such items, as well as limited skill and experience in this area. In practice, Educational Technologists (ETs) are increasingly being asked to advise on, and participate in, small to midsized projects to design and develop pools of digital test items. These items are generally used for summative assessment, and in quizzes aimed at stimulating active learning. ETs need a methodology for the design and development of digital items if they are to provide the best possible advice to those involved in projects of this kind.

6.1.4 ALTB project

The SURF ALTB project (Hartog, 2005) was carried out in 2005 and 2006. That project incorporated 15 small to midsized projects on the design and development of digital items. The aim of these various subprojects was to develop sets of questions for summative use, and for use in quizzes intended for formative applications. A systematic approach to the design and development of digital items was used under a range of conditions, in situations involving various forms of collaboration and types of task division. The intention was to identify the potential of digital items and to determine how they can best be used, to collate people's experiences, and to formulate the lessons learned. These experiences were used as input for the development of a methodology for digital item design.

6.1.5 Information sources on the Design and Development of digital items

In the ALTB project a methodology for the design and development of digital items has been defined as (1) a set of design requirements, (2) a set of design guidelines, (3) definitions of available components and item types (4) a library of paradigm examples (5) a library of design patterns (6) task structures and scenarios in which resources are allocated to subtasks along a time-line (Hartog, 2005).

In the ALTB project, attempts were made to collect information on these methodology ingredients. In this section the usefulness of available information that is intended to support the process of designing and developing innovative digital items will be explored.

Design guidelines

The literature contains long lists of design guidelines for multiple-choice items (T/F, alternate choice, four options) to be used in assessments. See, for example, Haladyna and Downing (2002). During the ALTB project, however, it was found that SMEs regard most of these guidelines to be unhelpful. This is due to the fact that such guidelines often actually are requirements in stead of pointers for inspiration. The projects showed that ETs should avoid focusing their advice and participation on the promotion of such guidelines.

Available item type taxonomies

Some researchers have developed frameworks within which both traditional and innovative question types can be categorized (Haladyna, 2004; Scalise et al., 2006). Such categorizations should preferably lead to the appropriate development and use of the items in question. These frameworks offer a perspective that is based on a combination of stimuli presentation and item formats. These frameworks are based on the categorization of item formats ranging from very low complexity (e.g. True/False questions) to a greater complexity (e.g. drag-and-drop items, constructed response and essay-type items). Additional dimensions involving knowledge and cognitive processes are sometimes added to this framework, as an overlay. Parshall (2002) has indicated five dimensions in which digital items could be described as "innovative". These dimensions are the item format (the response obtained), the response action (for example key presses, mouse clicks), media inclusion (images, photographs, graphs, video, animation, etc.), level of interactivity (system responses) and scoring method (how responses are converted to scores).

In the ALTB project, these frameworks were used to help SMEs and their assistants get their projects up and running. Although helpful in this way, the frameworks were not able to provide those involved with inspiration. The project participants regarded these frameworks as interesting instruments for the analysis and categorization of items, but not as a means of conceiving items for use in their own particular courses.

Examples of digital items

During the project, desk research was undertaken to identify possible sources of sample digital items for use in higher education. The number of such sources was found to be relatively limited (Bull et al., 2001; King et al., 2001; Mills et al., 2002; Parshall et al., 2002; Scalise et al., 2006). For the most part, the samples available from these sources are derived from secondary education and from subject domains other than those involved in the 15 small to midsized projects (life sciences, medical sciences and engineering sciences). The ALTB project showed that ETs and SMEs were seldom able to use these examples as paradigm examples or as a source of inspiration. One major problem was that SMEs encountered great difficulty in abstracting the examples. That imposes a barrier to subsequent transformation of those examples for applicability for their own courses.

Another issue that was often encountered in the cases dealt with by the ALTB project involved indicators for the effort needed to develop questions beyond the stage of the initial concept. "How much time will it take to flesh out that question within my own authoring environment?", "Can I author it myself or do I need a specialist for this?". Not one of the sources consulted was able to provide a satisfactory answer or approach to this problem.

The importance of the concept of design patterns as an instrument for a methodology derives from the limitations of individual examples, and the limitations of factors such as the usefulness of guidelines and the value of frameworks. In the next section, which explores the concept of design patterns, it is argued that one of their functions is to bridge the gap between abstract guidelines and isolated examples.

6.1.6 Design patterns

The term "Design Pattern", which was introduced by (Alexander, 1979) in the seventies of the last century is a concept used in architectural design. It was adopted for use in software engineering (Gamma et al., 1994) about 15 years later. Relations between components that repeatedly occur in different designs in answer to specific design challenges are called design patterns. The central idea is that it is not realistic to suppose that designers design from scratch. On the contrary: an experienced designer is supposed to have very many design patterns in his mind. "It is only because a person has a pattern language in his mind, that he can be creative when he builds" (Alexander, 1979, p. 206) .

Design patterns are generic combinations of solutions to recurring problems within problem-solving or design domains. Competent designers can instantly match a problem to the appropriate design pattern to arrive at satisfactory solutions to given problems and contexts. Design patterns are therefore an integral component of design methodology.

6.1.7 Design patterns for item design

Thinking in terms of design patterns for digital items takes the associated thought processes to another level. When applied to the design of digital items, design patterns bridge the gap between learning objectives and the item types currently available in CBA systems and LMSs. Design patterns span the divide between guidelines for item designers and examples that are already available. They also reinforce the importance of the distinction between design on the one hand and the development of digital items on the other. Lastly, by sharing design patterns, designers are able to learn from one another. In the interests of an efficient flow of information among ETs, a shared and accepted pattern language or format to describe patterns is necessary.

With regard to question design, the present authors found just a single publication that intentionally adopts a design-pattern-based approach. The design pattern concept is used in the Principled Assessment Designs for Inquiry project (PADI), which focuses on designing high-quality assessments of scientific inquiries. "The design patterns that are being developed as part of the PADI system are intended to serve as a bridge or in-between layer for translating educational goals into an operational assessment" (Mislevy et al., 2003).

To date, it is likely that most ETs have mentally internalized only a few design patterns for digital design, or that they have very limited numbers of these resources to hand. Yet ETs have the most to gain from the design pattern approach. It would enable them to provide better support for the SMEs, by supplying appropriate design patterns at just the right moment in item-development projects. The design pattern approach allows for a faster, more economical, yet more varied deployment of digital items.

6.2 A template for describing design patterns for digital items

6.2.1 Introduction

A common way to describe a design pattern is to provide a set of attributes and to describe the particular characteristics of each design pattern in terms of those attributes. To a large extent, the value of design patterns is determined by the ease with which a designer can identify a match between a pattern and a given problem. Accordingly, the set of attributes selected must provide adequate support for this process. In the case of a large set of patterns, we assume that the approach would be to use a browser to search for patterns in an online database. This might for example, involve entering specific values to search for specific attributes. Alternatively, free text searches could be conducted across all attributes.

The PADI project (Mislevy et al., 2003) describes design patterns on the basis of quite a large number of attributes: Title, Summary, Rationale, Focal KSAs (Knowledge, Skills and Abilities), Additional KSAs, Potential observations, Potential work products, Potential rubrics, Characteristic features, Variable features, I am a kind of, These are kinds of me, I am a part of, Educational standards, Templates (task/evidence shells), Exemplar tasks, Online resources, References, Miscellaneous associations. A worked out design pattern consists of tabulated text that takes up as much as two pages of A4. However, there are few specific item and task examples in a design pattern.

In most cases within the ALTB project, the implementation of the design pattern concept of Mislevy and Hamel was felt to be too abstract for digital item design. ETs in the field of higher education require design patterns that are less elaborate, to facilitate the process of searching for them. Another factor is the finding that design patterns must provide a clearer bridge to actual examples. At the same time, innovative digital items require greater emphasis on item format, in combination with the use of media. Lastly, the time required to design and develop real items are vitally important, if design teams are to allocate resources effectively. Therefore, it was decided to:

- limit the number of attributes;
- be more specific concerning the components of items (stimuli, prompts, item formats);
- add attributes relating to the design and development effort;
- add an attribute relating to the chance of arriving at the correct answer by guesswork alone;
- add an attribute relating to the possible presence or absence of extraneous cognitive load;
- provide more examples.

All of the attributes are listed and described below.

Title

The Title is intended to be a short description of the pattern’s core concept.

Context

The Context attribute describes the situation in which the design pattern in question can be used. It can contain information on the type of learning objective involved, together with details of the relevant domain of interest. It also describes the conditions in which the design pattern would be of use. The context provides references to specific sources, for further discussion of the design pattern in question.

KSA focus in a Summative Test

The focus on measuring Knowledge, Skills and Abilities (KSA) is a short description of the type of learning objectives that are to be measured. It is a combination of subject matter (i.e. domain knowledge), knowledge types, and cognitive processes. The descriptions of this attribute incorporate suggestions regarding the classification of the pattern within the taxonomy proposed by Anderson and Krathwohl (2001). As it is increasingly being used to classify objectives within education, this taxonomy is expected to remain a stable indicator for the foreseeable future. Its core concept is that educational tasks can be categorized on the basis of two factors, the knowledge dimension and the cognitive process dimension. This concept results in the following table.

Table 14: Two Dimensional Framework by Anderson & Krathwohl (2001)

The knowledge dimension:	The cognitive process dimension:					
	1: remember	2: understand	3: apply	4: analyse	5: evaluate	6: create
A: Factual knowledge	A1	A2	A3	A4	A5	A6
B: Conceptual knowledge	B1	B2	B3	B4	B5	B6
C: Procedural knowledge	C1	C2	C3	C4	C5	C6
D: Metacognitive knowledge	D1	D2	D3	D4	D5	D6

Within the context of design patterns for digital items, the range of questions turned out to be bound by dimensions A, B and C and by cognitive process dimensions 1, 2, 3 and 4. That is in line with observations (King et al., 2001).

KSA focus in a Quiz

The learning focus is a short description of the type of cognitive process or line of reasoning that can be induced by a question based on this pattern and knowledge type. With regard to the descriptions of this attribute, here too suggestions are made concerning their classification within the taxonomy table proposed by Anderson and Krathwohl (2001).

Pattern Core

The pattern core is a description of the pattern that is sufficiently generic in nature to enable an item to be generated concerning various specific situations within the context. At the same time the description is very tangible, in that it lists the individual components of the question. Furthermore, this list sometimes contains suggestions regarding the spatial arrangement of these components, which are specific elements of the question (stimulus, prompt, item format).

Design Effort

Design Effort is the amount of time needed to arrive at, or compile, the main conceptual idea of a question. On the basis of the experience gained in the 15 small projects on the design and development of closed questions, we are able to distinguish two levels of Design Effort:

- **Low:** Less than 15 minutes.
Design Effort can be minimal if – for example – use of the pattern does not require the designer to develop distractors or to develop new representations of knowledge.
- **High:** From 15 minutes to several hours. This type of effort usually involves finding and formulating distractors or new representations of knowledge.

Realization Effort

The Realization Effort is the estimated amount of time required during the ALTB project to develop and implement the conceptual idea of a question in an authoring environment. It also comprises the time that is needed to check, discuss and revise the question. We distinguish three levels of Realization Effort:

- **Low:** Less than 10 minutes. On average, this amount of development is needed for text only, standard type question formats such as True/False, alternate choice, multiple-choice, fill-in-the-blank.
- **Medium:** Between 10 minutes and 40 minutes. On average, this amount of development effort is required for more elaborate question formats such as hot spot, matching, multiple drop down lists, numeric and calculated formula. Some media resources, such as any images that are available, will often still need to be processed in order to make them suitable for display on screen.
- **High:** More than 40 minutes and up to 3 hours. This level of development effort might, for instance, be due to the fact that the questions involve the integration of video and animation. The creation of drag-and-drop questions with multiple markers also tended to require considerable effort.

Extraneous Cognitive Load

One of the most essential requirements for any item is validity. The options for more direct measurement of the intended construct (Parshall et al., 2002) in particular are put forward as an argument in favour of the design, development, and deployment of digital items. Extraneous cognitive load occurs when the student is required to allocate cognitive processing capacity to cognitive actions that are actually irrelevant to the correct answer. In particular this is the situation when the spatial arrangement of stimuli and response mechanisms requires a lot of eye movement or mental re-arrangements of facts and concepts. Eliminating this aspect as much as possible results in questions with no extraneous cognitive load.

Guess Change

The high probability to arrive at the correct answer by pure guesswork is often seen as a drawback for the use of multiple-choice questions. A number of design patterns have a set up that decreases this probability. For ETs it therefore is an interesting attribute. In the attributes, a **high** guess chance is given to the traditional T/F and 4-option multiple-choice questions (~ 0.5 to ~ 0.25). The value **intermediate** is given to design patterns that decrease that chance somewhat (to ~ 0.2 to 0.1). The value is set to **low** if this chance is decreased much more ($< \sim 0.1$).

Iconic Examples

The Iconic Examples section is an important attribute of design patterns. Iconic Examples clarify the semantics of pattern definition. In some examples, extra directives are mentioned as noteworthy aspects. However, we would like to emphasize the importance of abstracting from the example, rather than regarding the example as identical to the pattern. It gives details of real situations involving the use of the design pattern in question, either past or present, and of the solutions that were generated.

Scoring Rules

Scoring is of major importance for summative purposes, and must be considered carefully. Many of the 15 projects showed that various design patterns give rise to time-consuming discussions about scoring rules. It is good practice to inform students about the scoring of an item upfront. Accordingly, decisions about scoring should be made before the items in question are deployed in an actual test. Firstly, the scoring of questions should be discussed in relation to the goal of the item, and to that of the test in which it has to function. Secondly, characteristics such as answering time and the probability of guessing the correct answer should be considered.

Thirdly, the mutual interdependence of answering options must be taken into account when deciding on scoring rules. Finally, it is important to note that the specific characteristics of the CBA system in question impose limitations on the options for devising scoring rules. During the ALTB project no useful information was found in the literature that might lighten this task, nor could clear and univocal scoring rules for most patterns be devised.

In general, SMEs were comfortable with the idea of providing as much *transparency for students* as possible when it comes to scoring rules. For that reason, it is proposed that the following rules be applied (regardless of the type of design pattern involved):

- Let S_i be the maximum number of points that a student can get for question i ;
- Let p_i be a rational number between 0 and 1. Call p_i the partial credit factor for question i ;
- Now, S_i should be:
 - proportional to the weight allocated to a specific question within a test;
 - proportional to the amount of time that a student is supposed to allocate to this question within the test.
- Now, p_i should be:
 - proportional to the number of correctly chosen or constructed elements of an item.
- Given the above mentioned aspects, the attribute of Scoring Rules is left out in the design patterns. Ideally, however, SMEs, their assistants, and ETs should not have to invest any time in establishing scoring rules for questions.

6.3 Selected design patterns for digital items

About thirty design patterns were identified and described in the fifteen small to midsized projects on the design and development of digital items. In this section we present 10 archetypical design patterns. These patterns were arrived at on the basis of the instructional qualities that they bring to item design and their usefulness in a number of other contexts such as domain, task structure, knowledge and cognitive characteristics. They:

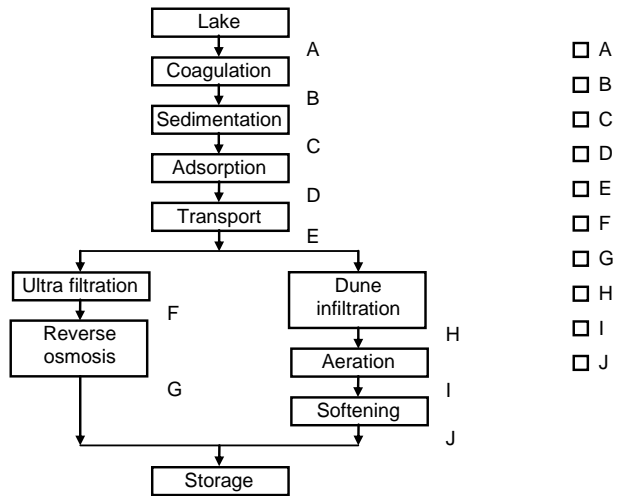
- require little design effort;
- allow much of the design and development work to be allocated to assistants and ETs.
- minimize guessing behaviour and unintended answering strategies (such as the elimination of options);
- are aimed at those knowledge categories and cognitive processes that are considered important by many SMEs in higher education (B2, B3 and C2, C3 of Bloom's taxonomy, as revised by Anderson and Krathwohl).

Each pattern takes up two pages of A4. On the first page, the values of the attributes are described. The facing page illustrates one or more examples derived from the pattern in question. This presentation format allows for easy browsing, retrieval and presentation of the design patterns.

An experienced designer will have internalized many design patterns. The list of patterns and examples should boost the experience of any novice question designer. In particular, educational technologists should study and internalize the patterns before becoming involved in projects for the design and development of questions. Experience in the ALTB project suggests that, while the examples sometimes may be considered as useless when they do not match exactly with the subject matter in the course, the design patterns may be considered useful because they reveal an abstraction of what several questions have in common in relation to a type of learning objective. Furthermore, for an educational technologist it is more difficult to control a discussion based on an example from subject matter, which he does not master than a discussion on a design pattern which is more generic. Finally, the table of patterns and examples is intended as a source of inspiration by any subject matter expert or assistant who really takes time to study the details of the table.

Indicating positions of sub processes in a process diagram.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
001	Any type of subject matter that uses process diagrams. At least some process diagrams must be available in the learning material or in the literature.	<p>Measuring the ability of a student to position a specific sub process within a given process.</p> <p>In general the student will not be able to deduce the answer without detailed knowledge of the inputs, outputs, and function of each of the sub processes. Questions based on this pattern can test understanding effectively provided that students have not previously encountered any of the specific sub processes used.</p> <p>A&K: B2, B3 C2, C3</p> <p>See also Roid & Haladyna, 1982 (1982: pp. 169-170)</p>	<p>Stimulating the student to think about the function, inputs and outputs of a specific sub process. Also the students must be aware of the inputs and outputs of each of the other sub processes. Stimulates student to scan the whole process.</p> <p>A&K: A2, A3, A4 B2, B3, B4 C2, C3, C4</p>	<p>A diagram of the whole process. An indication of possible placements of the sub process with symbols.</p> <p>A name or description of a specific sub process.</p> <p>A prompt that tells the student to indicate which of the indicated possible placements of the specific sub process makes sense, given the function of the whole process.</p> <p>Multiple response. Or Drag-and-drop.</p>	Low	Medium	No	Medium

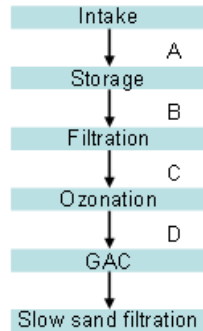
In the figure a scheme is shown of a groundwater treatment plant. Select all the positions in which filtration functionally can be placed, also if combinations of placements could be correct.



- A
- B
- C
- D
- E
- F
- G
- H
- I
- J

Course Drinking Water Treatment, L. Rietveld, Delft University of Technology.

Indicate possible locations of coagulation in the treatment train (more answers can be possible).



- A
- B
- C
- D

Course Drinking Water Treatment, L. Rietveld, Delft University of Technology.

Which of the positions indicated by a question mark is the correct position for a gel filtration unit in the given purification process?

Fermentor

Output (V: 10.00m ³ water)					
Naam	c	D (nm)	P (kg/m ³)	pI (-)	
protein1	4.30 kg/m ³	3.0	1090	6.5	
protein7	3.20 kg/m ³	20.0	1030	5.5	
Target	0.40 kg/m³	5.0	1110	7.0	
<i>E.coli 913</i>	20.10 kg/m³	1000.0	1105	7.0	
protein8	4.80 kg/m ³	11.0	1120	8.8	

new Centrifuge cost: €20 200

Output (V: 10.0m ³ water)		Waste (V: 10.0m ³ water)	
protein1	4.29 kg/m ³	protein1	0.01 kg/m ³
protein7	3.18 kg/m ³	protein7	0.02 kg/m ³
Target	0.40 kg/m³	<i>E.coli 913</i>	20.10 kg/m³
protein8	4.74 kg/m ³	protein8	0.06 kg/m ³

new Ion-exchange cost: €14 865

Output (V: 20.0*10 ⁻³ m ³ water)		Waste (V: 10.0m ³ water)	
protein1	50.00 kg/m ³	protein1	4.19 kg/m ³
protein7	19.30 kg/m ³	protein7	3.14 kg/m ³
Target	191.79 kg/m³	Target	0.01 kg/m³
protein8	35.69 kg/m ³	protein8	4.67 kg/m ³

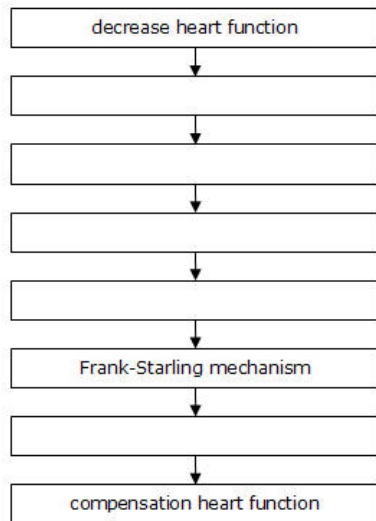
endpoint

Content (V: 20.0*10 ⁻³ m ³ water)					
Naam	c	D (nm)	P (kg/m ³)	pI (-)	
protein1	50.00 kg/m ³	3.0	1090	6.5	
protein7	19.30 kg/m ³	20.0	1030	5.5	
Target	191.79 kg/m³	5.0	1110	7.0	
protein8	35.69 kg/m ³	11.0	1120	8.8	

Course Process Technology, H.vd. Schaaf / R Hartog, Wageningen University.

Indicating relationships between qualitative changes of variables in a model.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
002	Any type of subject matter that uses quantitative or qualitative models. This pattern is useful in any type of subject matter that uses diagrams to illustrate the qualitative relationship between changes of process variables.	<p>Measuring the ability of a student to indicate qualitative relationships between process variables, between processes, or between individual phenomena within a process.</p> <p>The student is forced to demonstrate his mastery of the process as a whole.</p> <p>A&K: B2, B3 C2, C3</p> <p>See also Roid & Haladyna (1982: pp. 169-170).</p>	<p>Stimulates qualitative reasoning with respect to quantitative and qualitative models.</p> <p>Stimulates the student to think about the process as a whole.</p> <p>A&K: B2, B3 C2, C3</p>	<p>A symbol or passage of text representing a qualitative change of each process variable.</p> <p>A graphical configuration of most of these symbols or texts indicating the relationships between process variables.</p> <p>Placeholders for some of these symbols or passages of text.</p> <p>A prompt asking the student to drag the appropriate markers to the correct positions.</p> <p>Drag-and-drop.</p>	Low	High	No	Low

The diagram below depicts the hemodynamical process that occurs when a person has a beginning form of heart failure which was compensated. Drag the correct given processes (at the bottom of the diagram) in the correct positions.

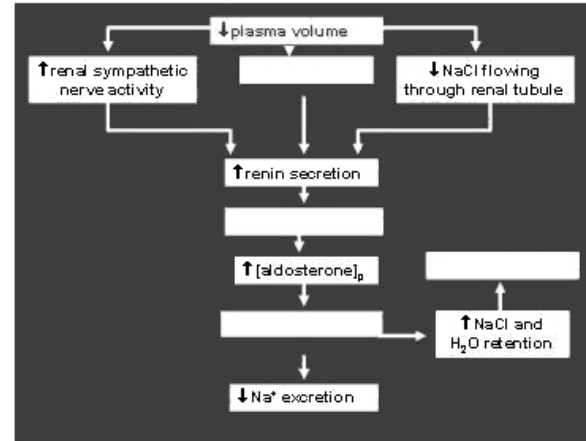


- decrease heart filling decrease effectively circulating volume
- deactivation sympatica/RAAS decrease blood volume (a.o.)
- increase effectively circulating volume increase HMV
- increase blood volume (a.o.) increase heart filling
- decrease HMV activation sympatica/RAAS

Course Physiology, S. Draaijer, Vrije Universiteit Amsterdam.

Note that all boxes are of equal size in order to prevent any cuing because of text length.
 Note that also foil text markers are present, this lowers the probability of a correct guess.

Drag the appropriate label to the appropriate box on the diagram.



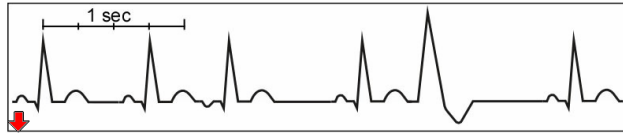
- ↓ Na+ reabsorption ↑ plasma volume ↓ plasma volume
- ↑ [angiotensinII]p ↓ [angiotensinII]p ↑ Na+ reabsorption
- ↑ arterial blood pressure ↓ arterial blood pressure

Course Phase 1, N.J. Part, University of Dundee.

Recognizing characteristics of phenomenon in a graph.								
	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
	This pattern is useful in any type of subject matter that uses graphs to visualize recordings of natural phenomenon or to depict deviations of normal situations (in economy, medicine, earth sciences, chemistry, physics).	Measuring the ability of a student to recognize the characteristics of a specific phenomenon in a graph. A&K: A1, A2 B1, B2	Stimulates the student to look carefully at the graph and to search for the characteristics of a phenomenon. Stimulates the student to attach the label of a phenomenon in his mind to a specific set of characteristics. A&K: A1, A2 B1, B2	A graph that represents a recording of the actual behaviour of a system over time or other variable. A label of a phenomenon. A prompt requesting to indicate the characteristic of the phenomenon. A marker. Drag-and-drop. OR Hot-Spot.	Low	High	No	

Some strokes of the ECG-waves depicted in the diagram below, show compensation breaks. Drag the marker in the diagram below to the position where the normal regular heart beat would have been, if NO extrasystoles had occurred.

ATTENTION: put the marker in a position (right) in line with the time-axis!



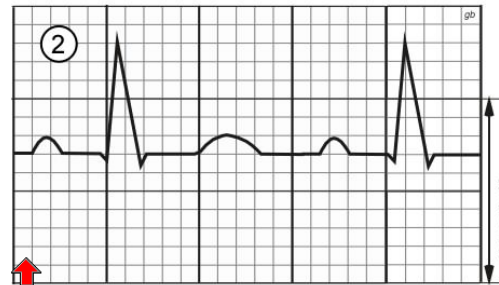
Course The Heart, R.J.M.P. Musters, Vrije Universiteit

During depolarisation a potential difference is present in parts of the heart.

Drag the pointer in the ECG-diagram below, to the position in which the potential difference is reduced to zero (indicating the end of the depolarization cycle).

ATTENTION: put the marker in the first PQRS interval and put is somewhere ON the trace!

Papiersnelheid: 25 mm/sec



Course The Heart, R.J.M.P. Musters, Vrije
Universite

The graph below depicts the registration of the Korotkow tones (as function of time) during the measurement of the blood pressure in the arteria brachialis.

Indicate - by positioning the red arrows - in the trace when the SYStolic (arrow pointing up), respectively the DIAstolic blood pressure (arrow pointing down) is reached.

Attention: drag the arrows as closely as possible to the correct Korotkow tones!

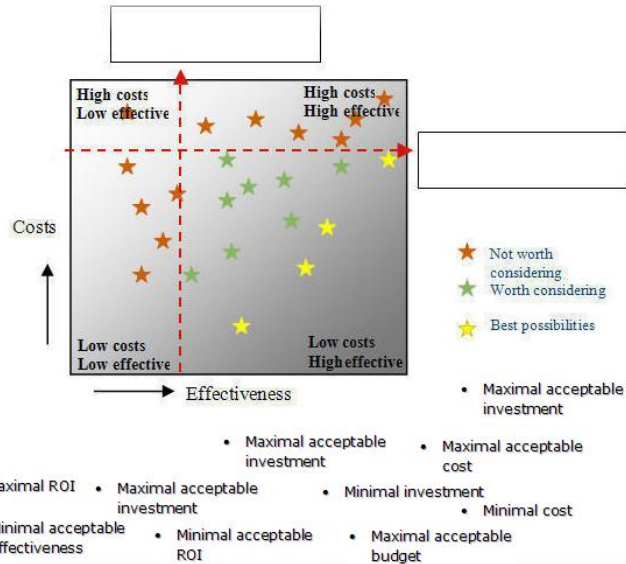


Course The Heart, R.J.M.P. Musters, Vrije

Recognize or recall the legend of a diagram, graph or table.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
008	This pattern is useful in any type of subject matter that uses diagrams, graphs and tables to denote important characteristics of concepts.	<p>Measures whether the student knows which variable belongs to which axis and/or which phenomenon belongs to which landmark point and/or which phenomenon belongs to which set of landmark points.</p> <p>A landmark point might be a maximum or a minimum or an intersection or some other "special" point in the graph</p> <p>A&K: B1, B2, B3</p>	<p>Stimulates students to focus on the meaning of a graph where the visual representation is already well known. Might make the students aware that they have not yet fully grasped the meaning of the graph.</p> <p>A&K: B1, B2, B3</p>	<p>A diagram (or graph or table).</p> <p>A prompt that asks the student to analyze the diagram and to determine what relations it depicts.</p> <p>Drag-and-drop. OR Drop down list. OR Fill-in-the-blank.</p>	Low	<p>High Drag-and-drop</p> <p>Low Drop down list and Fill-in-the-blank</p>	No	Low - Medium

Below, a decision diagram is depicted.

Drag one text label to each of the two boxes indicating the meaning of the red dotted lines.



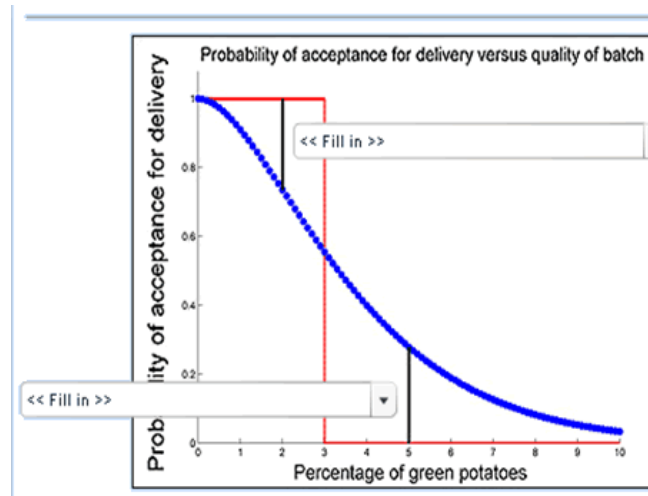
Course Food Safety Economics, A. Velthuis / R. Hartog, Wageningen University.

Note that all boxes are of equal size, in order to prevent cuing based on the length of the passage of text.

Note that the single combination of this design pattern and the same graph may give rise to several digital items

In the last slide we use a sampling plan of 50 sampling units ($n=50$) and accepting the batch by 1 of fewer green potatoes. The critical limit of acceptance is called c , in this case $c=1$

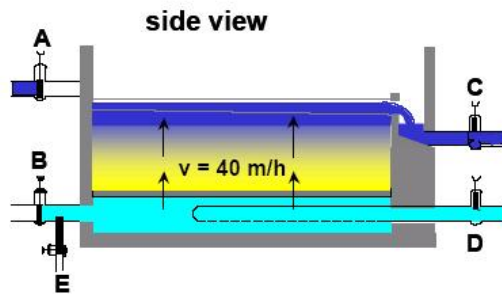
Select the right term for the two black lines in the graph to the below:



Course Sampling and Monitoring, E. Boer / R. Hartog, Wageningen University.

Ordering steps in a process or procedure.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
016	This pattern is useful for any type of subject matter that deals with specific linear or cyclical processes or with the sequencing of events.	<p>Measuring the ability of a student to remember or deduce the specific ordering of a specific process.</p> <p>Many instructors feel that a student who can provide an ordering that makes sense "understands" the related subject matter.</p> <p>A&K: B1, B2, B3 C1, C2, C3</p> <p>See also Roid & Haladyna (1982: p. 170).</p>	<p>Stimulates the student to scan each process step, possible orderings based on matching inputs and outputs of process steps, and on the intended function of the whole process.</p> <p>May also stimulate the student to learn about specific process steps, and about specific inputs and outputs.</p> <p>Is perceived as "creative" by some students. Finding the correct answer is believed to be more satisfactory than answering a traditional multiple-choice question</p> <p>A&K: B1, B2, B3 C1, C2, C3</p>	<p>A set of process or procedural steps in terms of a verbal or diagrammatic description.</p> <p>A definition of the function or intended output of the process or procedure.</p> <p>A prompt that asks the student to present an ordering of the steps such that the sequence of steps constitutes a complete process that realizes the given function or procedure.</p> <p>Ordering. OR Drag-and-drop.</p>	Low	<p>Medium For ordering.</p> <p>High For drag and drop.</p>	<p>Yes For Ordering.</p> <p>No For drag and drop.</p>	Low

What is the order of opening and closing valves for backwashing. Use the figure.



- Close valve A
- Open valve A
- Open valve C
- Close valve E
- Close valve B
- Open valve B
- Close valve D
- Open valve D
- Close valve C
- Open valve E
- Close valve A

Course Drinking Water Treatment, L. Rietveld, Delft University of Technology .

Design an experiment to test if a mouse that over expresses the NMDA receptor is more intelligent than a mouse that has a normal expression level of the NMDA receptor.

Place the following steps in the right order:

- Determine the genotype of the borne mice by Southern analysis
- Microinject the construct into a fertilized oocyte and transplant the fertilized oocyte into a mother mouse
- Make a construct
- Determine the genotype of the offspring by Southern analysis
- Test memory of the mutant mice
- Cross mutant mouse with wild type mouse
- Test transgene expression level in mutant mice by Western blot analysis

Genetics course, T. Aegerter-Wilmsen / T. Bisseling, Wageningen University.

Note that, in this example, use is made of the ordering question format. A drag-and-drop format is depicted for example in design pattern 002.

The chain of sampling and food microbiological analysis can be described in several steps.

Put in the right order:

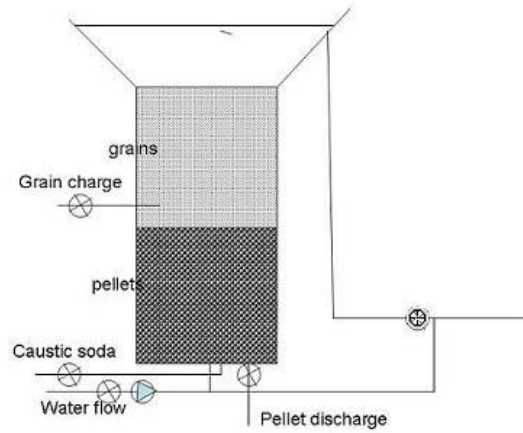
- Report to quality manager
- Interpretation by microbiologist
- Labelling and transport of sample
- Microbiological analysis
- Storage of sample in laboratory
- Decision quality acceptable / unacceptable
- Product or production proces
- Collecting samples

Course Sampling and Monitoring, E. Boer / R. Hartog, Wageningen University.

Note that, in this example, use is made of the ordering question format. A drag-and-drop format is depicted, for example, in design pattern 002.

Identify the error in process design.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
018_2	This pattern is useful with any type of subject matter that uses diagrams to describe processes.	<p>Measures the ability of the student to detect errors in a process design.</p> <p>For large models or designs etc., the effort required of the student might be out of proportion to the information generated by measurements using this question.</p> <p>A&K: B2, C2</p>	<p>Stimulates the student to study a design, model or process in total and to write a critique of it.</p> <p>A&K: B2, C2</p>	<p>A model OR A design</p> <p>An error introduced into the model or design</p> <p>A representation in the form of a diagram or a picture.</p> <p>A prompt requesting the student to identify and indicate any errors.</p> <p>Hot Spot. OR Drag-and-drop.</p>	Low	Low	No	Low

Indicate the error in the flow scheme of the pellet softening reactor.



Selected Coordinates 355, 266

Clear

Course Drinking Water Treatment, L. Rietveld, Delft University of Technology.

Which parameter setting in which operation is **not** correct when the function of the complete process should be to isolate the target protein?

Fermentor

Title: Fermentor

new Centrifuge
cost: €20 200

Title: new Centrifuge

Use: pellet

S (size): 200 (m²)

Resuspend in: 10 (m³)

new Ion-exchange
cost: €39 722

Title: new Ion-exchang

Type: Kation

Eluent: 1 (m³)

pH On: 6,5

pH Off: 8

endpoint

Title: endpoint

total costs: €59 922

purity: 0.00%

costs recovered: 0.00%

Output (V: 10.00m ³ water)				
Naam	c	D (nm)	p (kg/m ³)	pl (°C)
protein2	6.00	5.0	1050	8.0
protein7	3.20	20.0	1030	5.5
Target	0.40	5.0	1110	7.0
<i>E.coli 913</i>	20.10	1000.0	1105	7.0
protein8	4.80	11.0	1120	8.8

Output (V: 10.0m ³ water)		Waste (V: 10.0m ³ water)	
protein2	0.02	protein2	5.98
protein7	0.02	protein7	3.18
<i>E.coli 913</i>	20.10	Target	0.40
protein8	0.06	protein8	4.74

Content (V: 1.0m ³ water)				
Naam	c (nm)	D (kg/m ³)	p (°C)	pl
protein2	0.02			
protein7	0.02			
<i>E.coli 913</i>	20.10			
protein8	0.06			

Course Process Technology, H.vd. Schaaf / R. Hartog, Wageningen University.

Identify a detail error in a model-based calculation.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
018_3	<p>This pattern is useful with any type of subject matter in which model-based calculations are performed.</p> <p>See also pattern ID 018_4</p>	<p>Measures the ability of the student to detect errors in a calculation.</p> <p>For elaborate calculations, the effort required of the student might be out of proportion to the information generated by measurements using this question.</p> <p>A&K: B2, C2</p>	<p>Stimulates the student to study a computation in total and to become aware of forms of accuracy.</p> <p>A&K: B2, C2</p>	<p>A given problem.</p> <p>A computation for solving the problem.</p> <p>A detail error introduced into the computation.</p> <p>A prompt requesting the student to identify any errors.</p> <p>Hot Spot. OR Drag-and-drop.</p>	Low	Low	No	Low

The temperature of the water is 10 °C, the flow velocity through a pellet softening reactor is 70 m/h, the density of the pellets is 2700 kg/m³ and the pressure drop over the bottom 50 cm of the filter is 40 cm. Then the pellet diameter can be calculated. Indicate the spot of the mistake in the calculation.

$$H = (1-p)L \frac{\rho_p - \rho_w}{\rho_w} = 0.4 \Rightarrow p = 1 - \frac{0.4}{0.5} \cdot \frac{1}{1.7} = 0.53$$

$$H = 130 \cdot \frac{v^{0.8}}{g} \cdot \frac{(1-p)^{1.8}}{p^3} \cdot \frac{v^{1.2}}{d^{1.8}} \cdot L =$$

$$0.4 = 130 \cdot \frac{(1.3e-6)^{0.8}}{9.81} \cdot \frac{(1-0.53)^{1.8}}{0.53^3} \cdot \frac{70^{1.2}}{d^{1.8}} \cdot 0.5 \Rightarrow$$

$$d = 0.26m$$



Selected Coordinates

Clear

Course Drinking Water Treatment, L. Rietveld. Delft University of Technology.

Note that the calculation contains a detail error regarding the use of units within it.

Selecting the primary problem-solving strategy for a calculation problem								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
032	<p>This design pattern is useful for any type of subject matter that requires a specific problem-solving strategy. The subject matter categorizes problems and solutions. Examples can be found in statistics, mechanics, mathematics etc.</p> <p>Successful problem solving is conditional on the ability to select a strategy that is appropriate to the problem in question.</p> <p>See also the literature on factors for successful problem solving (Gick & Holyoak, 1983; Sweller, 1989).</p>	<p>Measuring the ability of a student to select the primary problem solving strategy.</p> <p>A&K: B2, B3 C2, C3</p>	<p>Stimulating the student to acquire factual knowledge about the functions and goals of processes.</p> <p>A&K: B2, B3 C2, C3</p>	<p>A prompt asking the student to select the correct options</p> <p>An option list that gives the standard set of tools and/or operations and/or processes that is available in the subject matter domain</p> <p>Multiple Response.</p>	Low	Low	No	Medium

Suppose we would like to test a lot of powdered milk on *Salmonella*:

- A lot of 20.000 kg powdered milk is produced
- 15 sample units of 25 g are taken randomly
- A lot is only accepted if all samples are negative
- Suppose you know that 100.000 nests of *Salmonella* are present in the lot and are homogeneously distributed over the lot.

Which of the following statistical tools and methods should you use to calculate the probability of accepting the lot.

- A. Binomial distribution
- B. Normal distribution
- C. Lognormal distribution
- D. Poisson distribution
- E. Uniform distribution
- F. Standard deviation

Course Sampling and Monitoring, E. Boer / R. Hartog, Wageningen University.

Which of the following statistical tools and methods can you use to calculate the probability that a sample of 1 ml contains no micro-organisms? The density d is equal to 2.5 organisms per ml.

- A Binomial distribution
- B Normal distribution
- C Lognormal distribution
- D Poisson distribution
- E Uniform distribution
- F Standard deviation

Course Sampling and Monitoring, E. Boer / R. Hartog, Wageningen University.

Distinguishing relevant laws, values, formulas etc. from irrelevant ones, to solve a calculation problem.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
029	<p>This design pattern is useful in situations where the subject matter calls for the application and execution of subject-matter relevant mathematical operations.</p> <p>This design pattern can be used in situations where it is necessary to perform calculations, but where additional information needs to be retrieved from the answer given.</p> <p>Compare this pattern with pattern ID 019.</p>	<p>Measuring the ability of students to potentially arrive at a correct answer to questions requiring the use of calculations.</p> <p>Understand the role of specific variables in calculations, without having to apply them.</p> <p>Selecting what is necessary for a computation.</p> <p>A&K: A2, A3 B2, B3</p>	<p>Stimulate the student to study and apply subject-matter specific, mathematical and solving algorithms.</p> <p>A&K: A2, A3 B2, B3</p>	<p>A prompt presenting a question about what is needed for a given calculation.</p> <p>A list with possible constants, variables or operations. Note that many textbooks include such a list as an appendix.</p> <p>Multiple Response.</p>	Low	Low	<p>No</p> <p>Note that the student needs to work on paper to be able to determine the correct choices.</p> <p>The student may be allowed to use a sheet containing formulas that are relevant to the subject matter.</p>	Medium

Indicate which formulae are necessary to calculate the backwash velocity in a clogged sand filter bed for drinking water treatment



$$H = 180 \cdot \frac{v}{g} \cdot \frac{(1-p)^2}{p^3} \cdot \frac{v}{d^2} \cdot L$$

$$H_{\max} = 1 - p \cdot L \cdot \frac{\rho_f - \rho_w}{\rho_w}$$

$$H = 130 \cdot \frac{v^{0.8}}{g} \cdot \frac{1 - p_e^{1.8}}{p_e^3} \cdot \frac{v^{1.2}}{d^{1.8}} \cdot L_e$$

$$R = \frac{B \cdot H}{B + 2 \cdot H}$$

$$Re = \frac{v_0 \cdot R}{v}$$

Course Drinking Water Treatment, L. Rietveld, Delft University of Technology.

Indicate which formulae are necessary to calculate the clean bed resistance of a sand filter for drinking water treatment



$$H = 180 \cdot \frac{v}{g} \cdot \frac{(1-p)^2}{p^3} \cdot \frac{v}{d^2} \cdot L$$

$$H_{\max} = 1 - p \cdot L \cdot \frac{\rho_f - \rho_w}{\rho_w}$$

$$H = 130 \cdot \frac{v^{0.8}}{g} \cdot \frac{1 - p_e^{1.8}}{p_e^3} \cdot \frac{v^{1.2}}{d^{1.8}} \cdot L_e$$

$$R = \frac{B \cdot H}{B + 2 \cdot H}$$

$$Re = \frac{v_0 \cdot R}{v}$$

Course Drinking Water Treatment, L. Rietveld, Delft University of Technology.

Note that in this example, the same formulae are used as in example to the left.

Distinguishing relevant classes of information for problem solving from irrelevant ones.								
ID	Context	KSA summative	KSA Quiz	Pattern Core	Design Effort	Development Effort	Extraneous Cognitive load	Guess chance
030	Any subject matter that relates problem solving to classes of information.	<p>Measuring whether a student knows what information is relevant to finding or creating solutions to a given problem.</p> <p>A&K: B2, B3 C2, C3</p> <p>Also: direct measurement focussing on highest level in SOLO taxonomy (Biggs, 1999)</p>	<p>Stimulating students to be aware of the distinction between information that is either relevant or irrelevant to a given problem, and encouraging them to apply this awareness.</p> <p>A&K: B2, B3 C2, C3</p>	<p>A list of information classes.</p> <p>a problem.</p> <p>a prompt asking which classes of the list of information classes is relevant to attempts to deal with this problem.</p> <p>Multiple Response.</p>	Low	Low	No	Medium

What information should typically be taken into account in an effectiveness analysis that is part of Risk Management?

Information about:

- Risks
- Interventions
- Effectiveness
- Benefits
- Alternatives
- Costs
- Societal Acceptance
- Environmental Consequences

Course Food Safety Economics, A. Velthuis / R. Hartog, Wageningen University.

What information is necessary to calculate the concentration of oxygen in water that is in open connection with the outside air?

- Temperature of the water
- Temperature of the air
- Volume percentage of oxygen in the air
- Valence number of electrons in the oxygen
- Molecular mass of oxygen
- Partial pressure of oxygen
- Ion strength
- Partitioning's Verdelingscoefficient H
- molair fraction oxygen
- Atmospheric pressure
- Molecular mass of H₂O
- The gas constant

Course on Drinking Water Treatment, L. Rietveld, Delft University of Technology.

6.4 Conclusions

About thirty design patterns were identified in fifteen small to midsized projects on the design and development of digital items. Ten design patterns are presented in full. It is thought that many more design patterns can be devised. A format has been developed and used to describe the set of design patterns. The format helps ETs to quickly scan through the patterns and to make matches between a given learning material, a given learning objective, and a given pattern.

A scan of the selected set of design patterns show that some patterns use the drag-and-drop item format. This supports statements by other researchers (King et al., 2001; Scalise et al., 2006) that item types involving drag-and-drop operations hold great potential for use in digital environments. The design patterns described also demonstrate how the drag-and-drop format allows for a more direct measurement of the construct intended, through the alignment of conceptual, spatial and textual information. In this way, for example, the effects of construct-irrelevant variance on the basis of students' reading level ability (Downing et al., 2004) and extraneous cognitive load are avoided. At the same time, developing drag-and-drop items induces more development effort.

A number of the selected design patterns are related to performing calculations. Calculation problems represent a challenging problem for question design. To date, most calculation problems are worked out in multiple-choice questions in which students have to select the correct numerical or algebraic answer to the given problem or in fill in the blank questions. Some design patterns described in this chapter show options that go beyond that approach by presenting problems in which students have to identify the mistake in a calculation or in which they have to select the appropriate laws and formulas needed to arrive at the correct answer for a given calculation problem.

One aspect of the concept of design patterns is that there are a great number of possible patterns. Scanning patterns to find one that matches a specific and detailed learning objective is time consuming, as they are only available on paper. This problem has already been encountered with the thirty patterns developed during the ALTB project. It is also unreasonable to expect SMEs to learn and internalize every single pattern. This is one area in particular in which ETs in Higher Education can prove their worth, by internalizing as many design patterns as possible. In interviews with SMEs, they will then be able to offer an appropriate design pattern on a "just-in-time" basis. This will undoubtedly boost the level and efficiency of item design and development. The next step in the concept of design patterns for item design is to familiarize a group of ETs with the concept of design patterns, and to increase the number of available patterns. The ETs will then have to invest effort in memorizing a large set of design patterns and in working with them. This will enable them to effectively internalize these patterns. In addition to this chapter on the subject, a tutorial has been developed to instruct participants in the use of design patterns for digital item design. The first workshop on the basis of this tutorial, which attracted fifteen participants, has already been evaluated. Average overall satisfaction was rated at just above 8, on a scale of 1 to 10.

The problem of determining scoring rules for some of the design patterns, has had an impact on the extent to which design patterns are perceived to useful. Furthermore, the lack of generally accepted scoring rules for the most promising design patterns has given rise to considerable debate on the validity of some of the design patterns in question. Further progress in the use design patterns and digital item types will require considerable input from the field of psychometrics.

6.5 References

- Alexander, C. (1979). *The timeless way of building*. New York: Oxford University Press.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.
- Bull, J., & McKenna, C. (2001). *Blueprint for Computer-assisted Assessment*: RoutledgeFalmer.
- Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327-333.
- Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Amsterdam: Addison and Wesley Professional Computing Series.
- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (Third ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education* 1, 15 (3), 309-334.
- Hartog, R. (2005). ALTB website. Retrieved march 20 2007, from <http://fbt.wur.nl/altb/>
- King, T., & Duke-Williams, E. (2001). *Using Computer Aided Assessment to test higher level learning outcomes*. Paper presented at the 5th CAA Conference, Loughborough.
- Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-Based Testing, Building the Foundation for Future Assessments*. London: Lawrence Erlbaum Associates.
- Mislevy, R. J., & et al. (2003). *Design Patterns for Assessing Science Inquiry*. Menlo Park: Center for Technology in Learning (SRI International).
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York,: Springer-Verlag.
- Scalise, K., & Gifford, B. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment* 4(6).

7 Practical Aspects of Task Allocation in Design and Development of Digital Closed Questions in Higher Education

Published as "Hartog, R., Draaijer, S. & Rietveld, L.(2008).Practical Aspects of Task Allocation in Design and Development of Digital Closed Questions in Higher Education. Practical Assessment, Research and Evaluation 13(2)"

Rob Hartog
Wageningen University

Silvester Draaijer.
Vrije Universiteit Amsterdam

Luuk Rietveld
Delft University of Technology

Abstract

For projects on design and development of digital closed questions in higher education the task structure is analyzed. Based on fifteen small to mid sized projects in four universities, a practical set of tasks is defined and practical aspects of task allocation are described and discussed. Ten design and development scenarios are presented. Based on time registrations in the projects and on a few experiments, estimates are given for the most important cost categories in the budgets for the scenarios.

7.1 Introduction

The increased availability of Learning Management Systems and facilities for computer-based assessment (CBA), induce more and more teachers in higher education to invest in the design and development of pools of digital closed questions. A closed question is any fixed response item that can be administered by a computer. Digital closed questions are being developed for computer-based assessment but also for use as activating learning material (ALM). In practice, several hybrid roles for closed questions can be distinguished.

To take full advantage of innovative closed questions, considerable knowledge is required, regarding question design, educational measurement and multimedia development. In addition, a set of practical skills is needed with respect to question editing and entry, image processing and management of questions and pictures. Therefore, design and development of question pools in higher education is often a matter of teamwork in projects. The number of students that will use the questions resulting from such a project will generally be much lower than the number of participants in a nationwide or large scale test or exam. Because the costs per student tend to determine what budget is acceptable, smaller numbers of students in practice correspond to smaller project budgets. Thus, realistic budgets for the design and development of a set of questions in higher education are much lower than budgets for large scale tests. If there would be hard quality criteria for digital closed questions to be used for CBA or in an ALM role and if these quality criteria would be widely accepted, reality might be different. Currently, quality is de facto an implicit derivative of the quality of the design and development team and their working procedures.

This chapter focuses on small to mid sized projects for the design and development of closed questions in higher education with no explicit quality criteria. These projects were projects in a larger program (Hartog, 2007) aimed to develop a methodology for the design and development of digital closed questions. One of the aims of this program was to identify what aspects design and development of digital closed questions for different roles (ranging from pure ALM role to pure CBA role) can have in common. The first author was supervisor of this program and took part as educational technologist in seven of the projects. The second author took part as educational technologist in two of the projects. The third author took part as subject matter expert in two of the projects. The program also involved a number of projects on educational measurement issues related to innovative closed questions and interoperability. The results of these latter projects fall outside the scope of this paper (see Hartog, 2007).

The chapter describes the most common classes of human resources, defines and discusses the tasks and matches these tasks to possible functions that might be defined within the university. Suggestions are given to prevent waste of efforts. Furthermore, the chapter presents a set of scenarios and corresponding budget templates. For a number of entries in these templates, cost estimates are given.

7.2 Method

Data for this chapter were collected from fifteen small to mid sized projects in higher education in which closed questions for learning goals and objectives in natural sciences, engineering sciences and social sciences were designed and developed. On average in the projects, about one third of the developed questions were the common type multiple choice questions. About two thirds of the developed questions make use of other questions types such as multiple correct, matching, ranking, hot-spot, drag-and-drop. Examples of innovative use of question types are presented in (Hartog, 2007).

The aim of the projects was twofold: first to design and develop pools of digital closed questions and second to develop design requirements, design guidelines and design patterns for new design and development projects in higher education. As such, the projects can be classified as developmental research projects (Richey et al., 2004).

Table 15 presents an overview of these projects. In the table, the case, course level, course subject, number of target population of students, role of the questions, the authoring software, the set-up of the development team and the calculated average design and development time per question are listed.

Progress and experience was reported at regular time intervals. Each project was evaluated and attempts were made to use experience in the form of requirements, guidelines and patterns in the next project. The most tangible results of the projects were more than 2000 questions and about 30 design patterns.

At regular time intervals, initially every three months, progress in terms of newly developed questions was reported. For reasons of accountability, the time invested by every person in the project was registered. Furthermore, observations were reported as to inefficiencies, problems and issues that were recognized as important. From now on, the term 'case study' will be used to refer to the body of qualitative and quantitative data and the corresponding analysis of a project.

Analysis of the collected quantitative data (numbers of questions, designed and developed and corresponding time registrations) and qualitative data (observations, descriptions of working procedures) revealed a common task structure, and was a basis on which ten scenarios were developed for small to mid sized projects for design and development of closed questions in higher education. The next sections describe resources and roles of team members, tasks and options for allocating this task and issues related to the costs of this task. In particular, one section presents a budget template for each of a number of scenarios.

7.3 Description of the Design and development contexts

In this section, resources that are needed for question design and role descriptions of team members within a design and development project are described.

7.3.1 Question authoring and delivery environments

First of all, without any hard- and software system, there would not be any project for digital question design. Probably every institute for higher education now has a Learning Management System and sometimes a dedicated computer-based assessment system. These systems offer support for authoring questions and for managing questions, pools of questions and exams. In eight of the fifteen case studies, Blackboard version 6.0 (2006) was used as Learning Management System. In one of the projects N@tschool (ThreeShips, 2007) was used as Learning Management System. Most of the Learning Management Systems offer support for 'quizzes' and 'tests' that primarily contain closed questions. In four of the fifteen projects Questionmark Perception version 3.x (Questionmark, 2002) was used. Finally, in two projects, a Questions and Test Interoperability v2.x (QTI 2.x) delivery system was used. For these two projects, the questions were edited directly in QTI 2.x XML templates.

Instructors use Learning Management Systems and computer-based assessment systems to present 'quizzes' to students and for summative assessment in regular courses. These systems provide a number of new question types, or seemingly new question types, which are often referred to as 'innovative'. A number of these types involve the use of multimedia.

7.3.2 Different functions and different competencies

The case studies revealed four specific functions within a design and development project. These functions are:

- Subject Matter Expert;
- Assistant SME;
- Educational Technologist;
- Rendering Specialist.

The Subject Matter Expert usually is a professor or associate professor. The professor is also the principally responsible person for the content of a course, the learning goals and for the development of questions. Alternatively, the subject matter expert may be an invited speaker for instance from industry.

Educational Technologists are the designated persons to provide knowledge and skills with respect to the design and development of closed questions, the possibilities and limitations of the available authoring and question delivery environments. The educational technologist is assumed to have broad knowledge from the field of instructional design and educational measurement. Typical sources used in the projects were (Scalise et al., 2006), (Haladyna, 1997; Haladyna, 2004) and (Bull et al., 2004). In addition, the educational technologist has to play an important role in the project definition and project set-up. On that basis, educational technologists are to provide design guidelines and present design patterns.

Table 15: Overview of 15 small/midsized D&D projects

Case	Course Level	Course Subject	Number of students per year (about)	Role of the questions	Software	Development team	Average D&D time/question (in minutes)
WU1	Master	Food Safety (Toxicology/Food Microbiology)	30	summative	QM	SME and assistant	160
WU2	Master	Food Safety Management	30	activating	Bb	SME and ET	150
VU1	2nd year	Heart and Blood flow (physiology, ECG measurement and clinical ECG interpretation)	300	diagnostic and summative	QM	SME and ET	220*
VU2	3rd year	Special Senses (vision, smell, hearing, taste, equilibrium)	300	summative	QM	SME and ET	80
TUD1	3rd year	Drinking water treatment	30	activating	Bb	SME and assistant	85
WU3	Master	Epidemiology	100	summative (open book)		SME and assistant	130
TUD2	3rd year	Sanitary Engineering	50	activating	Bb	SME and assistant and ET	95
WU4	Master	Food Toxicology	100	summative	QM	SME and assistant	130
WU5	Master	Food Micro Biology	40	activating	Bb	assistant	80
WU6	Master	Advanced Food Micro Biology	30	activating	Bb	assistant	130
WU7	Entry Master	Food Chemistry (general introduction module for candidate students)	open self test WWW	diagnostic	QTI delivery	SME = ET	120**
WU8	Entry Master	Food Toxicology	open self test WWW	diagnostic	QM	SME and assistant	120
WU9	Master	Sampling and Monitoring	30	diagnostic (self-)	Flash	SME and Assistant and ET and Flash programmer/design patterns used	80**
WU10	Master	Food Safety Economics	30	summative (not open book)	Bb and on paper	SME and assistant and ET/design patterns used	***
FO1	1st year	Curriculum: General Sciences	30	diagnostic-'plus'	N@t-school	SMEs and Rendering specialist	160**

Note. WU = Wageningen University, VU = Vrije Universiteit, TUD = Delft University of Technology, FO = Fontys University of Applied Science, QM = Questionmark Perception, Bb = Blackboard LMS, QTI = Question and Test Interoperability 2.0 format, N@tschool = N@tschool LMS, SME = Subject Matter Expert, ET = Educational technologist, ASME = recently graduated student or student-assistant with subject matter expertise but not at SME level

* time included extensive training sessions of SME with ET, aiming at using other than MC questions.

** For a number of questions only time for design in Word was registered. For those questions an average of 20 minutes/question for question rendering was added.

*** Time registration included too many other activities for which correction was not possible

A separate parallel project was defined for investigation of issues with respect to educational measurement. Insofar the design and development projects encountered questions with a strong educational measurement component these questions were passed on to this parallel project. For the actual design and development none of the teams incorporated an educational measurement specialist. For the design and development of questions for the ALM role this was not deemed relevant because the primary function of those questions was not measurement. For the design and development of questions for the CBA role, the combination of the educational technologists, access to literature was supposed to be sufficient and the link to the parallel project was considered sufficient.

An important aspect of design and development projects in higher education is that many of the relevant learning objectives cannot be understood or grasped by the team members who are not subject matter experts. This puts a tension on the position and possibilities of the educational technologist within such projects. The educational technologist costs less per hour than a subject matter expert.

The assistant of the subject matter expert has some subject matter knowledge but cannot be considered an expert. Often, the assistant of the subject matter expert is an almost or just graduated student within the relevant discipline. The subject matter knowledge of the assistant is greater in comparison to the subject matter knowledge of the educational technologist. The assistant however, will usually not have any specific question design and development competence. In most universities, the typical assistant subject matter expert will be hired just for the project. The assistant is always considerably cheaper than the educational technologist. For the majority of the small to mid sized projects, an assistant was appointed to contribute to the design and development of the questions.

Above, the rendering specialist refers to the question entry and picture processing specialist or service. This specialist (or pool of specialists, for example within an institution's audio-visual services department) is someone who is proficient with desktop computers and has a lot of routine with question entry and elementary picture processing tasks. Thus, the productivity of the rendering specialist can be very high as long as his tasks are well defined. In practice the latter implies that the questions to be entered are available in a very clear format (for example MS-Word documents with sufficient annotation) and that the entry task can be completely outsourced to such a team member. In one of the case studies a rendering specialist as defined above, did most of the question entry and picture processing work. The rendering specialist is not necessarily cheaper than the assistant of the subject matter expert.

Table 16: Roles, competencies and costs for question design and development

	SME	ET	ASME	RS
Cost/hour	High	Medium	Low	Very low
Subject Matter Knowledge	High	Low	Medium	None
Question design and development Knowledge	Low	High	Low	Low
Educational Measurement	None	Medium	None	None
Knowledge of Authoring environment	Low	High	Low	Medium
Routine with the Authoring environments and other computer tools	None	Medium	Medium	High

Note SME = Subject Matter Expert; ET = Educational Technologist; ASME = assistant of the Subject Matter Expert; RS = Rendering Specialist

The four functions do not imply that every design and development project necessarily involves each function. For example, it is imaginable that a subject matter expert decides to fill a question bank all by himself with a readily available authoring environment. Furthermore, a subject matter expert can perfectly well realize a question bank by appointing a subject matter expert or a rendering specialist and delegate work to them (for example a help-desk employee). The scenarios that are presented below, take these set-ups into account. Because the authoring environments are usually considered as overhead costs, accounted for within institution wide budgets, human resources are the most dominant factors for the costs of a project. In Table 17 the roles, competencies and relative costs of the team members of mid sized question design and development project are listed.

7.4 Practical TASK ANALYSIS

Because of similarity in used question types and software tools, the task of design and development of closed format questions for both Computer Based Assessment and Active Learning Material always includes a number of common tasks. In this section, the tasks in mid sized projects on the design and development of closed questions are described. The tasks cannot be mapped one to one to phases in a project because tasks can overlap considerably. The practical task analysis has been carried out from the perspective of actual design and development of innovative questions. Furthermore, we have tried to highlight what design and development of digital closed questions for different roles have in common and what the differences are. A task analysis primarily focused on the delivery of a complete assessment would have resulted in a different set of primary tasks.

7.4.1 Defining the Project

Every project requires that some effort is invested in assessing the context of the project and the context of the project results. On that basis, a realistic project plan and a corresponding budget can be defined and financial means for the project can be acquired. The project plan should specifically describe the intended output of the project, the role of the questions, available resources and a deadline. In the case studies, these variables have shown to be important determinants for the quality and success of the project in terms of effectiveness and efficiency. Determining the available resources at the start of the project implies:

- which Learning Management System or Computer Bases Assessment system will be used (or is available);
- whether there are well defined learning objectives or previous questions or exams available;
- what learning materials are available;
- whether there are design patterns available;
- what the number and type of questions to be designed and developed should be.

For the CBA role, the case studies showed that in practical contexts a tangible and sensible goal is a pool of questions that is sufficient for four exams and a trial exam. The reason for this is that subject matter experts generally will need several equivalent exams for a few successive cohorts of students. Furthermore, subject matter experts need a trial exam which shows the students what to expect for an upcoming digital exam. When an exam contains about sixty closed questions this implies that about sixty clusters of five equivalent questions need to be designed and developed. In fact, the first question that is designed for such a cluster should be a good operational definition of the detailed learning objective in this cluster. For exams with sixty questions, about three hundred questions will have to be designed and managed. This requires that the clusters are labeled. Part of defining the project should involve a conscious decision with respect to the composition of the development team.

7.4.2 Setting Up the Project

Given an approved project, the project plan can be worked out in more detail. A development team can consist of one or more subject matter experts, an educational technologist, an assistant of the subject matter expert and a rendering specialist. For reasons of cost efficiency and because the time of most subject matter experts is scarce, it should be the intention in a project to delegate as much as possible of the subsequent tasks to educational technologists, assistants and rendering specialists. For example, for entering questions in a CBA system, a subject matter expert or educational technologist is actually too expensive. Such work is more appropriate for an an assistant or a rendering specialist. Also, the assistant or rendering specialist will often have more routine in question entering and picture processing and therefore can execute the task more quickly.

Analysis of the case studies highlighted the fact that, the diary of many subject matter experts seldom display empty time slots. Given their crucial role in setting learning objectives, providing inspiration and validation of questions, everyone involved should be prepared to dynamically adapt the project agenda to availability of the subject matter expert. In order to avoid frustration and delays in project progression it is advisable to set as soon as possible due dates for delivery of specific batches of questions (e.g. for specific topics, learning objectives or question types) in different stages of completion. These stages are typically characterized by a draft version, a revised draft in some intermediate form of representation, and a final version in the authoring environment.

The project set-up will almost always imply some training. More specific, for computer-based assessment, a subject matter expert and an assistant must be made familiar with elementary knowledge on educational measurement and question design. Typical resources for such training are (Bull et al., 2004; Frary, 1995; Haladyna, 1997; Haladyna, 2004; Kehoe, 1995a, 1995b) Furthermore, subject matter experts and assistants must be made aware of the possibilities and limitations of digital questions and the specific software application. In the projects a number of fundamental problems with response processing for innovative question types were identified. These problems were passed on to an educational measurement project within the program and are beyond the scope of this paper (Hartog, 2007).

In the fifteen projects, no training material was found that is adequate for subject matter experts or assistants as defined in this chapter. Most of the knowledge that an educational technologist had readily at hand is based on handbooks listed above. However, examples in these handbooks stem from secondary or vocational education, or from disciplines that had not enough in common with the disciplines in the projects. Also, presenting requirements as to correct grammar and clear formulations in the form of guidelines was not appreciated. In the case studies, some training material was developed for the subject matter experts in the form of about sixty design guidelines and a set of design patterns (Hartog, 2007). Experience in the case studies suggested that design patterns were more helpful than design guidelines.

For Activating Learning Material, elementary knowledge on learning and instruction is necessary. In a number of projects use was made of (Keller, 1983; Smith et al., 1993) and (Merriënboer, 1997). In a later stage also (Fenrich, 2005) came in view.

7.4.3 Collecting and defining learning objectives

When designing and developing questions for the CBA role, it turned out to be an effective approach to define a label for each cluster of five equivalent questions in an early stage of the project. Such a label can be denominated a 'bucket' for which questions need to be designed. If there is a list of detailed learning objectives available, this will reduce the effort needed for this task. However, in the case studies, there was seldom an adequate list and when there was such a list it allowed for too many interpretations. Often, an even more specific subject matter denomination was necessary up to the level of specific micro-subjects within a course. Previously developed sets of (mostly open) questions, assignments or learning materials, such as presentations or lecture notes implicitly containing the learning objectives could partly be used to define the learning objectives up to micro level.

The task of 'labelling clusters' is irrelevant if the project aims at questions for the ALM role. In such a case, the team should make an ordered list of detailed learning objectives for which learning can be supported by closed questions. The ordering should be based on a quick cost/benefit analysis. This cost/benefit analysis should identify for which learning objectives it will require relatively little effort to develop motivating closed questions with a high expected impact.

When the project focuses on CBA, the assistant was usually able to extract a part of the list of learning objectives from overall learning goals in combination with learning materials such as slide presentations, textbooks and from previous exams. To some extent, the educational technologist was able to coach the assistant. However, the subtask of defining a set of labels ('buckets') could never be completed without involving the subject matter expert.

In case of a project for the design and development of questions for the ALM role, the assistant was usually able to indicate some pieces of learning material that are – at the start of the project – insufficiently complemented by activating learning material. Exam results of previous cohorts also pointed the assistant towards learning objectives that call for additional activating learning material. When the subject matter expert becomes familiar with the real possibilities of innovative closed questions, it will be the subject matter expert who can best identify the learning objectives which offer a good chance of a low cost/benefit ratio. The case studies showed that many subject matter experts need some training in order to become familiar with the real possibilities of innovative closed questions.

7.4.4 Design and intermediate representation of questions

Ultimately, design and development of a closed question implies that a micro learning objective is represented in the form of a closed question. Assuming that the designer(s) has/have such a learning objective in mind, a first idea of a question (or cluster of questions) must be generated. The remainder of the design and development of a closed question will then involve:

- deciding on the exact interaction type
- including a case
- deciding on including of media
- authoring the text-based components of the question.

How the first draft of questions comes about is dependent on the knowledge and skills of the subject matter expert assistant or the skill of the educational technologist to inspire them. The initial training may help in this process.

In three case studies, design patterns proved to support both the generation of ideas for questions as well as decision making for the used interaction types (Hartog, 2007). Design patterns can form a powerful tool to let subject matter experts and assistants see the possibilities of digital closed questions and also reduce costs through a more effective generation of first draft questions. The case studies revealed that the first drafts of questions are usually laid down in MS-Word documents with annotations on specific detail: the intermediate representation of questions. The relatively easy method of creating, editing, revising and sharing MS-Word documents is the principal reason for that approach. Another reason is that subject matter experts are familiar with a standard text editor but would have to invest considerable time in learning to use a question authoring environment. Often email-communication was used to share information. The case studies revealed that such communication is very sensitive to problems with versioning. Including media in questions may involve designing or finding a picture or designing or finding an audiovisual object.

The design task requires deep subject matter knowledge and understanding. This implies that the subject matter expert and the assistant must do most of the work. The educational technologist can provide inspiration in terms of design patterns and by suggesting guidelines. The extent to which the subject matter expert can delegate the design task to the assistant depends very much on the subject matter knowledge of the assistant, on the availability of learning materials and on the question design competence of the assistant. Within the fifteen case studies, the output of assistants in terms of quantity and quality differed widely.

The aggregation level of the case study data is not adequate for determining the costs that are involved in this part of the design process. However, the fifteen case studies highlighted many sources of inefficiency. This resulted in the following lists of don'ts in order to keep the costs within limits.

- Don't search for a specific picture, only use readily available materials.
- Don't make drawings or pictures, but if you do, use them for more than one question
- Don't develop case-based questions, but if you do, make sure it is a fertile basis for a number of questions.
- Don't start by default making traditional MC questions; do invest some time in starting with different types that do not require developing distracters.
- Don't design and develop instances of innovative question types for assessments unless scoring is adequately supported by the available CBA system and the rationale for the scoring rules used by the available system is transparent to faculty and students.
- Don't write extensive feedback.
- Don't let the assistant develop questions for which no design patterns or examples exist.

It is important that the subject matter expert has contact within short time intervals with the assistant. This will prevent that the assistant invests much time in designing questions that later turn out not to conform to the learning goals and have to be discarded. The case studies confirmed that it is difficult to represent detailed learning objectives in some form other than the question itself. In some of the projects the assistant would, based on an initial formulation of a learning objective in natural language, design questions which were completely of the mark. Furthermore, it also occurred that at the end of the course period detailed learning objectives for which questions already were developed, had to be removed from the list of detailed learning objectives. One of the reasons for this was for instance that guest lecturers tended to change ad hoc the content of their lectures.

All in all, development efforts that lead to questions that are useless, increase the average development effort per useful question. It is believed that this is one of the factors leading to gross underestimation of design and development efforts.

7.4.5 Validating questions

When a first draft of a question has been made, the question will have to be validated, checked and revised. Validating the first draft involves more than just answering the questions and checking if the answer is 'correct'. It also involves checking for errors and ambiguities in the question formulation. Most of all, the validator has to check if the question really measures (i.e. operationally defines) a learning objective (in case of CBA role) or stimulates the intended action and line of reasoning (in case of ALM role).

The case studies made clear that it is not enough to point out problematic issues within a question. In the type of small to mid sized design and development projects which are the subject of this chapter, validators cannot restrict themselves to indicating which questions are not good enough. In practice, the validator is actually co-designer. Thus (s)he has to provide a handle for improvement of the question or for a completely different approach with respect to the learning objective. Consequently, in most case studies the validation task overlapped with the task of intermediate design. This obscures good quality control. However, a more strict separation of formal validation and actual design and development would require a larger investment and a different type of projects.

The lecturer or professor who is responsible for the course and for the corresponding assessment will have to validate questions drafted by the assistant. Alternatively, when the subject matter expert has drafted the questions, an assistant and in some cases the educational technologist can check many aspects of the question such as consistency, phrasing, choice of terminology, et cetera.

Validation can often be supported by data if the questions have been used by students in previous exams or by previous cohorts. Analysis of data often points toward 'suspect' questions. However, such analysis falls outside the scope of this chapter. In the budget templates below we therefore refer to 'ex ante' validation.

7.4.6 Revising questions

In practice, many first draft questions were revised or discarded on the basis of the validation results. Often, second drafts were made and needed to be validated again and discussed again. This process results in several versions of questions and pools of questions. The case studies showed that the teams had difficulties in managing versions of questions and keeping track of which question had what qualities.

The revision task is primarily a task for the subject matter expert and assistant. From the case studies, it became apparent that the delegation of the design task and the revision task to the assistant will always induce some waste of efforts.

7.4.7 Image processing

In the case studies, a considerable number of images have been used. Even though the images were already available in digital format, they still had to be processed. This involved operations like: changing the format of the image, resizing, clipping, deleting part of the image, replacing part of the image, inserting text in the image, indicating hotspots. These operations require routine with an image processing application. Some of these operations also require routine with the question editor of the Learning Management System or CBA system. Most of this work does not require subject matter knowledge and at first sight a rendering specialist would seem the most appropriate person to execute this task. However, the case studies do not provide sufficient information to arrive at a decision rule about to what extent image processing should be delegated to a rendering specialist.

7.4.8 Realization in CBA system

In this task, the finalized draft questions are entered into an authoring environment. This includes at least: calling up the system, initiating a new question, copying text and images into the stimulus, choices, distracters, and also formatting, layout and setting scoring rules. Furthermore, this task implies question pool management. This task requires routine with the authoring environment, file management and with picture sorting and selection tools and often still requires picture resizing operations as well. In general, this task should be delegated to an assistant or to a rendering specialist.

In the practice of the fifteen projects, it was not standard procedure to check the final version of the question in the system, lay out quality et cetera. In case it is really necessary for the subject matter expert or assistant to validate the questions on screen and to send the comments back to the rendering specialist cost savings might be negligible. Therefore, for the type of these small and mid sized projects, it is deemed better to train the rendering specialist and make this person fully responsible for the final version.

The costs for entering a validated question into an authoring environment are based on the type of question that is entered and whether media is to be included or complex scoring rules need to be entered. In order to estimate how much time this would require by someone who is very proficient with authoring tools, a benchmark set of questions was entered by three proficient persons in Blackboard, Questionmark Perception and by means of editing QTII2 conformant questions in XML. Table 17 lists the results.

In practice, the task of entering questions in an authoring environment took always much longer than the figures in Table 17 suggest. In the case study in which this task was performed by a dedicated rendering specialist, the average question entry and picture processing time of almost 25 minutes was recorded. The order of magnitude was confirmed by data from two other projects apart from the case studies.

While the time registrations in other case studies are not detailed enough in order to provide more quantitative data, many time-consuming actions related to the task of question entry were mentioned. Examples are: looking up missing details, rearranging materials, rearranging desktop settings, interpreting meta information scribbled by the subject matter expert, adjusting picture sizes, moving files around, making mistakes and repairing mistakes, previewing the question, system failures and so on and so forth.

Table 17: Benchmark test for entering set of standardized questions with different authoring environments

Question-type	time in minutes		
	Bb-expert	QMP-expert	XML-editor
multiple-choice	2	3	9
multiple-choice with image	5	7	14
multiple answer	3	5	11
multiple answer with image	7	8	19
fill in blanks	4	12 ¹	17
fill in blanks (numeric) with image	3	7	22
matching	3	4	5
matching with image	6	6	13
pull down	5	20 ²	19
pull down with image	5	8 ³	16 ⁴
ranking	2 ⁵	7	8 ⁴
ranking with image	4 ⁵	8	30 ⁶
drag and drop	5	10	24
hotspot	3	5	9
select a blank	3 ⁷		15
select a blank with image	4		17

Note. Bb – Blackboard, QMP – Questionmark Perception v 3.x, XML-editor – a person familiar with XML editing who edited two sets of 80 questions in QT12 (QTI = Question and Test Interoperability).

1 – Time to enter without modifying the outcome definitions to give a score for partial correct answers: about 6 minutes 2 – On the basis of an existing question, used as template 3 – Table inserted as 1 image 4 – Implemented as select a blank 5 – Implemented as matching 6 – Implemented as drag and drop 7 – Implemented as fill in blanks

7.4.9 Additional CBA-related tasks

This chapter is based on the assumption that design and development of closed questions can be discussed as a distinct cluster of tasks. The complete process of computer-based assessment involves several other tasks. These tasks are not directly related to the actual design and development of questions. Strictly speaking, they do not fit the scope of this chapter. However many subject matter experts in higher education are interested in some indication of the point where computer-based exams become more cost efficient than 'traditional' exams. For this reason, also *organization of exams* (including configuring the exams and organizing exam sessions) and *processing of exam results* (psychometric test analyses and score interpretation and grade curving) have been included in the budget templates below.

7.4.10 Additional management and communication within the team

For projects in general, management rather than communication is usually defined as a separate task. In this chapter, the communication within the team is defined as a separate task because the cost for communication grows when more people are working in a project. The main factors that currently contribute to communication costs are threefold. Firstly, the fact that subject matter experts have in general few timeslots available for face to face communication. Secondly, a lack of subject matter expert-friendly support for workflow, collaborative design and version control. Finally, the challenge to optimize the workload of the rendering specialist whose capacity will be shared among different projects.

7.5 Ten scenarios

In this section, we present ten possible scenarios for design and development projects of closed questions. The authors believe that these scenarios cover the various set-ups of small or mid sized projects for design and development of sets of closed questions within higher education. The scenarios are intended to support initial planning and setting a budget for the project.

Table 18 describes scenarios for projects that focus on questions for a Computer Based Assessment role. Table 19 describes the scenarios for projects that focus on questions for an Activating Learning Material role. Table 18 also supports a structure for comparing the costs of a written exam, based on open questions with the costs of a digital exam, based on closed questions. Both sets of scenarios are ordered from maximum support for the subject matter expert to minimal support for the subject matter expert. The tables provide a comprehensive overview of relevant tasks within a project for the design and development of closed questions, the allocation of these tasks and the amount of time required. Such tables have not been found in literature yet and are believed to form an important tool for anyone involved in mid sized question design and development projects.

Both templates assume that a project is set up to design and develop a pool of about 300 questions. For the CBA role, this can reflect the design of 60 clusters of 5 equivalent questions. The tables highlight the cost structure and the structural consequences of reallocation of tasks. The time values in the table are estimates based on the time registrations in the fifteen projects. However, the reader can easily insert other values for certain parameters. Some of the scenarios imply independent choices, for instance, the percentage of questions that will include a picture or the amount of training to be provided for the assistant. Parts of the data are contextual data depending on the institution and often also on the country where the institution resides. The costs/hour of a subject matter expert vary widely across different countries in the world and so do the costs of the other specialists. Another example of an estimate that may vary widely for different projects is the ratio of the time for question entry needed by a subject matter expert and the time needed for this task by a rendering specialist. In the tables, this parameter is set to 1.5.

Apart from these project specific parameters, the last column in Table 15 contains the average calculated Design and Development time per developed question. This value is based on an analysis of the time sheets of every employee in each of the projects that were used for the case studies. The overall conclusion was that average design and development times were up to 2 hours. Based on experience in the case studies it is believed that in a budget for a design and development project, this time should not be set lower than two hours per question for projects. Based on experience in the case studies this average overall time is divided over different subtasks. Notice that the difference in the time between questions for the ALM role and questions for the CBA role is mainly due to the necessity to provide feedback in the former.

The budget examples presented in Table 18 and Table 19 make clear how cost efficiency gains might be realized by reallocation of tasks. For instance, with the current settings of parameters and values the budget templates suggest that the average design and development time without support will be relatively low. However, for many institutions it is likely that the costs will be higher. The actual efficiency gains for any institution can only be determined by inserting the actual data in the cells.

7.6 Conclusions and discussion

From fifteen small to midsized projects on design and development of innovative digital closed questions for natural and engineering sciences in higher education quantitative and qualitative data were collected. Analysis of these data from the shared perspective computer-based assessment and activating learning materials led to a practical task structure for such projects. For a number of these tasks this analysis has led to practical advice, which has been described in the respective paragraphs.

Based on the case studies the options to delegate tasks to an assistant of the subject matter expert, to an educational technologist and to a rendering specialist have been described. For defining, planning and budgeting such projects good estimates for an average design and development effort of closed questions, typical for a university context, are important.

However, such estimates could not be found in literature. Communication with colleagues in higher education as well as some initial experiments always seemed to point to 'about half an hour per question' as a good estimate. Time registrations within the projects have resulted in more empirical cost estimates for some of the tasks and the average total design and development time per question. On average, the latter was close to two hours per question.

Based on reports produced within the projects, sources of inefficiency were identified and a number of 'does and don'ts' are formulated. It is concluded that efficiency improvements, which are mainly based on division of labor, tend to increase the need for communication between the subject matter expert and the other members of the team. Realizing efficiency gains requires adequate control of this communication process. It is suggested that an educational technologist takes the specific responsibility to support and manage this process. In addition, the need for subject matter expert-friendly computer-based support of workflow management, version control and collaborative design was identified.

In order to support planning and budgeting of future projects, two sets of reference scenarios and budget templates for mid sized design and development projects have been developed. The reference scenarios and corresponding budget templates cover the most likely practical contexts for such projects and highlight for which tasks efficiency gains might be realized and what consequences of labor division are possible.

The scenarios presented in this paper highlight that design and development of digital closed questions for different roles ranging from the role of activating learning material to the role of questions for computer-based assessment have a number of aspects and tasks in common. Clearly, the design and development of complete assessments using innovative digital closed questions involves a need for deep knowledge and understanding of educational assessment theory. However, experiences in the projects showed that when detailed educational measurement knowledge needs to be acquired during project, it can lead to a frustration and waste of effort. In the program on which this chapter is based, expertise on educational assessment was clustered in a special project within the program. This project falls outside the scope of this chapter.

Experience in the fifteen projects suggests that educational assessment expertise that goes beyond the expertise that can be expected of an educational technologist concerns primarily two forms of experience. The first form implies understanding the possibilities and limitations for assessment of innovative question types that are available in the learning management system or computer-based assessment system at hand. This implies knowledge of theory of educational assessment combined with detailed knowledge of the system used for educational measurement. The second form implies all knowledge that is directly related to complete assessments. The educational technologist often lacks these two forms of knowledge. This will make it necessary to involve an educational assessment expert.

Subject matter experts and assistants with subject matter knowledge need training with respect to design of digital closed questions for both roles of questions, the role to function as activating learning material and the role to function within computer-based assessment. Therefore, the next step is to develop a workshop for subject matter experts and assistants with subject matter knowledge. Initial experience with the design patterns developed in the case studies suggests that these design patterns might form the core of the training material for assistants.

Table 19: Five scenarios and corresponding budgets for the development of 300 questions for the ALM role

	Max Support					Max Support No RS					Max Support only ET			Max Support only ASME			No Support														
	SME hr	ET €	ASME hr	RS €	Total €	SME hr	ET €	ASME hr	Total €	SME hr	ET €	Total €	SME hr	ASME hr	Total €	SME hr	Total €														
Design&Development of Pool of 300 questions for ALM role																															
1. making project plan, defining budget	1	90	14	980	1070	1	90	14	980	1070	1	90	14	980	1070	4	360	14	700	1060	4	360	360								
2. setup team/allocate people to tasks	1	90	4	280	1	50	1	30	450	1	90	4	280	1	50	420	1	90	4	280	370	2	180	4	200	380	1	90	90		
2a.setting up communication the team	2	180	2	140	2	100	2	60	480	2	180	2	140	2	100	420	2	180	2	140	320	2	180	8	400	580	2	180	180		
2b. training	2	180	4	280	8	400			860	2	180	4	280	8	400	860	2	180	4	280	460	2	180	8	400	580	2	180	180		
3. matching of objectives and questions	4	360			16	800			1160	4	360			16	800	1160	8	720	4	280	1000	8	720	4	200	920	12	1080	1080		
4. design/intermediate representation*					150	7500			7500					150	7500	7500	120	10800	30	2100	12900	50	4500	100	5000	9500	125	11250	11250		
4a. authoring presentational feedback					75	3750			3750					75	3750	3750	40	3600			3600		75	3750	3750	40	3600	3600			
4b. authoring interactive feedback**									PM											PM										PM	
5. ex ante validation***	50	4500							4500	50	4500					4500	50	4500			4500	50	4500			4500	50	4500	4500		
6. improving and/or replacing questions			25	1750	25	1250	25	750	3750			25	1750	50	2500	4250			75	5250	5250			75	3750	3750	25	2250	2250		
7. image processing					75	2250	2250					75	3750	3750					125	8750	8750			125	6250	6250	150	13500	13500		
entering in CBA system					100	3000	3000					120	6000	6000					130	9100	9100			120	6000	6000	150	13500	13500		
9. providing access to students					16	800			800			16	800	800					16	1120	1120			16	800	800	20	1800	1800		
12. additional communication within team	8	720	4	280	8	400	8	240	1640	8	720	4	280	8	400	1400	8	720	4	280	1000	8	720	4	200	920					
total budget costs					31210				35880					49440					38990					52290							
D&D time per question (hr)					2.1				2.1					2.1					2.2					1.9							
D&D costs per question (€)					101				117					161					127					168							

Note * e.g. in natural language in MS Word, ** in the ALTB project no data about authoring interactive feedback have been collected, *** of the intermediate representations

7.7 References

- Blackboard. (2006). Blackboard. Retrieved march 04 2006, 2006, from <http://www.blackboard.com/>
- Bull, J., & McKenna, C. (2004). *Blueprint for computer-assisted assessment*. London: RoutledgeFalmer.
- Fenrich, P. (2005). *Creating instructional multimedia solutions: Practical guidelines for the real world*. Santa Rosa, California: Informing Science Press.
- Frary, R. B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research and Evaluation*(4), 11.
- Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights: Allyn & Bacon.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (Third Edition ed.). London: Lawrence Erlbaum Associates.
- Hartog, R. J. M. (Ed.). (2007). *Design and development of digital closed questions: A methodology for midsized projects in higher education*. Utrecht: SURF Foundation.
- Kehoe, J. (1995a). Basic item analysis for multiple choice tests. *Practical Assessment, Research and Evaluation*, 4(10).
- Kehoe, J. (1995b). Writing multiple-choice test items. *Practical Assessment, Research and Evaluation*, 4(9).
- Keller, J. M. (1983). *Development and use of the arcs model of motivational design*. Enschede: Technische Hogeschool Twente.
- Merriënboer, J. J. G. v. (1997). *Training complex cognitive skills: A four-component instructional design model for technical training*. Englewood Cliffs, NJ: Educational Technology Publications.
- Questionmark. (2002). Questionmark. Retrieved 19 jan 2007, 2007, from <http://www.questionmark.com/>
- Richey, R. C., Klein, J. D., & Nelson, W. A. (2004). Developmental research: Studies of instructional design and development. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (2e ed., pp. 1099 - 1130). Mahwah NJ: L. Erlbaum Associates.
- Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "Intermediate constraint" Questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment.*, 4(6).
- Smith, P. L., & Ragan, T. J. (1993). *Instructional design*. USA: Macmillan Publishing Company.
- ThreeShips. (2007). N@tschool. Retrieved 8 februari 2007, 2007, from www.threships.nl

8 Computer Support for Design and Development of Innovative Closed Questions

Pierre Gorissen
Fontys University of Applied Science

Abstract

The IMS Question and Test Interoperability specification can be used as a way to support the design and development of innovative closed questions. The chapter describes the five dimensions of innovation that can be distinguished in closed question assessment items and links them to the functionalities supported by the IMS QTI specification. The chapter shows that the QTI specification offers enough flexibility and supports enough functionality to be used as the basis for innovative closed question items and very interactive structures of multiple individual questions.

The Integrated Item Design and Development Environment or IIDDE, a fictional system, is used to describe how the required functionalities for a flexible authoring environment for assessment items based on web services in a service oriented architecture can be realized.

8.1 Introduction

This chapter discusses the use of the Question and Test Interoperability (IMS, 2006f) specification developed by the IMS Global Learning Consortium, as a way to support the design and development of innovative closed questions.

It describes the five dimensions of innovation that can be distinguished in closed question assessment items and will link them to the functionalities supported by the IMS QTI specification. Next the requirements for a flexible authoring environment for these items and the extent to which currently available systems meet these requirements.

The chapter aims at system developers, vendors, expert users of virtual learning environments and assessment systems and on decision makers currently in the process of drafting requests for proposal (RFP) for systems aimed at the design and development of innovative closed questions.

Because the ALTB project was a joint effort of four Universities (Fontys University of Applied Science, Delft University of Technology, Vrije Universiteit Amsterdam, Wageningen University) and one testing and assessment company (Cito) using a number of different systems to create and manage assessments had interoperability requirements built into the project plan. Even though technical interoperability leading to actual re-use and exchange of assessment items between the partners within the project wasn't required because of the division of the project results into topic areas, the project needed a way to develop a common vocabulary related to the design guidelines for closed questions.

An interoperability specification like the IMS Question and Test Interoperability specification, not only provides a technical method of exchanging assessment items between systems, a common vocabulary facilitates communication between content experts and item developers during the development stage of items.

8.2 IMS Question and Test Interoperability Specification

8.2.1 The IMS Global Learning Consortium, Inc.

The IMS Global Learning Consortium, Inc. (IMS) develops and promotes open specifications for facilitating online distributed learning activities such as locating and using educational content, tracking learner progress, reporting learner performance, and exchanging student records between administrative systems. IMS has two key goals: Defining the technical specifications for interoperability of applications and services in distributed learning, and supporting the incorporation of the IMS specifications into products and services worldwide. IMS endeavors to promote the widespread adoption of specifications that will allow distributed learning environments and content from multiple authors to work together (in technical parlance, "interoperate") (IMS, 2007). The first IMS specifications were released in 1999; at the moment there are a dozen specifications available. Some of the specifications are relatively new, while others have gone through a number of revisions based on input from the field. Each specification consists of an information model (describes what is in the specification), a XML binding (describes the technical implementation) and a best-practice and implementation guide (advise about how to use and implement the specification). The documents can be downloaded or read online, free of charge, at the IMS website: <http://www.imsglobal.org/>

8.2.2 Short history of QTI

The IMS Question and Test Interoperability Specification or QTI version 1.0 was released in 2000. The specification was updated in 2001 (v1.1) and 2002 (v1.2) to improve and extend the model. By 2003 it became clear that a major revision was necessary, a quick fix (1.2.1) was followed by a complete rewrite.

One reason for the need for a rewrite was the further development of in particular the IMS Content Packaging, Simple Sequencing, and Learning Design specifications since the first release of QTI and the need for a cross-specification review. Also, as the implementations of QTI matured, particularly during the phase of development between versions 1.1 and 1.2 of the specification, a number of issues have been raised that could not be addressed without making substantial changes to the specification. There was considerable pressure from the QTI community to address these issues with a revised version of the specification. Some of the issues related to the addition of functionality, in particular support for some new items types in common use. Many of them related to improving the data model generally to provide better conformance testing and better integration with modern approaches to rendering QTI content in assessment delivery engines. In September 2003 a project charter was agreed to address both the collected issues from 1.x and the harmonization issues and to draft QTI 2.0.

In order to make the work manageable and ensure that results were returned to the community at the earliest opportunity some restrictions were placed on the scope of the recommended work. Therefore, the QTI 2.0 version of the specification was released in January 2005. The scope of QTI 2.0 version of the specification was restricted to the individual assessment items. It did not update those parts of the specification that dealt with the aggregation of items into sections and tests or the reporting of results. The QTI 2.1 release, in June 2006, completes the update from version 1.x to version 2.x by replacing those remaining parts of the QTI specification. The June 2006 release was labeled Public Draft (revision 2) and not yet Final Version. The reason for that was the increase of importance within the IMS consortium of the availability of at least one reference implementation of specifications released by IMS. Despite the draft status, the specification is currently considered ready for implementation. At the moment (February 2007) the specification is going through the Final Release process. As part of that process there needs to be an internal interoperability demonstration between two or more systems. This process is expected to take about a year to complete.

8.2.3 What is new in the QTI 2.1 specification

This chapter uses the current latest publicly available release of the QTI specification, version 2.1 Public Draft (revision 2). At the moment most systems that have implemented QTI, support the 1.2 version of the specification. This section gives a brief overview of the great number of changes that have been made in the 2.x version of the QTI specification compared to the 1.2 version. This section has been previously published in the Quickscan QTI 2006 (Gorissen, 2006).

Re-alignment with other IMS specifications

Since the release of version 1.2 of QTI, a number of other IMS specifications were released or updated in a way that was relevant for the QTI specification.

- *IMS Content Packaging Specification*

Prior to QTI 2.0 there was no recommended or predefined way of *packaging* resources when transferring items, tests or processing templates between systems. This often caused problems when for example items used images or animations as part of the item. Starting version 2.0 of the specification, the use of IMS Content Packaging for this is required.

The QTI specification requires no modifications or extensions to the existing Content Packaging data model, features of that specification are used in the way originally intended. The goal was to enable the use of content packages containing assessment objects with the existing base of tools (package editors, repositories etc.) that support IMS Content Packaging (IMS, 2006d).

- *IMS Learning Resource Meta-data Specification*

Previous versions of the IMS QTI specification had a QTI specific meta-data set contained within the data structures of the items and assessments themselves. That set of meta-data elements had names which all started with the characters 'qmd_'.

In QTI 2.0, this QTI-specific meta-data has been brought into line with the IEEE Learning Object Metadata (LOM) standard in accordance with the IMS Meta-data Best Practice and Implementation Guide for LOM which is part of the IMS Learning Resource Meta-data Specification version 1.3(IMS, 2006a). The IEEE LOM standard defines a set of meta-data elements that can be used to describe learning resources, but does not describe assessment resources in sufficient detail. The application profile provided in the QTI 2.x specification extends the IEEE LOM to meet the specific needs of assessment item developers wishing to associate meta-data with items as defined by the accompanying Item Information Model. These elements for example list the interaction types used in the item, whether the item is composite, time dependent or whether the solution is available. A complete list of QTI-specific meta-data elements and a description of how to use the IEEE LOM profile can be found in the IMS Question and Test Interoperability Meta-data and Usage Data document (IMS, 2006e)

The alignment with the IMS Content Packaging specification also meant that the meta-data elements were removed from the individual QTI items and fitted into the `ims_manifest.xml` file that is part of the Content Package. The Content Package is the place to describe the resources in the package. This enables existing repositories and tools to read/write at least the generic part of the meta-data for items and assessments.

- *IMS Learning Design Specification*

The IMS Learning Design 1.0 specification offered placeholders for use of an external assessment model like IMS QTI as part of the Units of Learning that are defined within IMS Learning Design. The IMS QTI 2.x specification describes the use of IMS Learning Design properties and IMS QTI variables as a way to integrate both specifications. This integration enables the use of QTI items and assessments within an IMS Learning Design unit of learning.

Changes in the Item Content model

Version 2.x of QTI introduced a number of changes to the item content model.

The content model is the part of the specification that relates to the body of the item. The item body contains the text, graphics, media objects, and interactions that describe the item's content and information about how it is structured. The body is presented by combining it with stylesheet information, either explicitly or implicitly using the default style rules of the delivery or authoring system.

- *XHTML*

One very noticeable change compared to QTI version 1.2 is that the content model for the items now is restricted to a well defined subset of XHTML. Use of plain text or RTF is no longer allowed. Though this might seem as a more restrictive, it is much clearer defined, can be validated against the QTI schema and is easier to implement. Through support for the object-element and MathML-support the model is still flexible enough to cater for most needs.

Content that needs to be available in multiple items can be shared using Xinclude allowing for another way to optimize the content development and maintenance.

- *Interactions*

The combination of response types and rendering types that was used to determine how an item should be rendered in QTI 1.2 has been replaced by a system of sixteen interactions (IMS, 2006c): endAttemptInteraction, inlineChoiceInteraction, textEntryInteraction, associateInteraction, choiceInteraction, drawingInteraction, extendedTextInteraction, gapMatchInteraction, graphicInteraction, hottextInteraction, matchInteraction, mediaInteraction, orderInteraction, sliderInteraction, uploadInteraction

These interactions can be used in any combination within an item allowing for very sophisticated items. A final interaction, the customInteractions allows for the use of interactions not covered by the current QTI information model.

- *Adaptive items*

Especially in formative assessments finding the correct answer to a question often isn't something that needs to be limited to just one single attempt. In those cases the process of searching a correct answer is as important as actually finding it. An adaptive item allows for multiple attempts and can change the feedback, displayed information according to the number of attempts, the selection options, depending on the actions of the candidate. It even is possible to display additional interactions for example to help the candidate solving parts of the question.

- *Item templates*

A lot of items in formative and summative assessment are variations on common structures. For example if an item is designed to test if a candidate can multiply two numbers between 1 and 10, manual creation of items for all possible combinations of a multiplication of two number between 1 and 10 is not efficient.

In QTI 2.x the item designer can create one single item template that describes this multiplication question. That template can then be cloned, either during run-time or at any given time into a required set of items to be used in an assessment.

- *Inline feedback*

The introduction of inline feedback, where the feedback is displayed as part of the original item allows for much more flexible design of stimuli for the candidate. Especially if the item is adaptive, inline feedback can be valuable because the item designer can choose whether feedback given for the previous attempts stays visible for the candidate as part of the item body or not.

Because inline feedback can also contain new interactions it can also be used to have a candidate solve parts of the problem if his previous attempts have shown that he didn't fully understand the complete question.

- *Number formatting*

In many types of items the formatting of the numbers used in both, the item, the feedback and the response by the candidate can be very relevant. QTI 2.0 en 2.1 offer extensive number formatting capabilities.

Changes at Assessment level

The QTI 2.1 version introduced a number of changes and enhancements at assessment and section level

- *Item reference*

A very visible change is the fact that the XML of an assessment item is no longer included in the XML of the assessment. Instead the assessment contains just the references to the items. The advantage is that if one single item is used in three assessments, the XML file of the item is now simply referenced from within the assessment. Updates to that item only need to be done once in the external item file where in version 1.2 any change made to an item also needed to be made in each and every assessment file where that item already was in use.

- *Pre-conditions and Branching*

QTI 2.1 offers the assessment developers the use of pre-conditions to determine whether an item should be displayed. The branching option can be used to determine which item should be displayed next based on either the score or selected answer option of the previous item.

This allows for the creation of assessments that adapt the selection of the next item to be presented based on the performance of the candidate so far in the assessment.

It also enables the use of items as 'selector-items' for example as an assessment where the candidate has to demonstrate to know the best combination of steps to take to solve a problem.

8.2.4 Implementation related changes

A final category of changes relates to support for implementers and system designers.

Conformance model

QTI version 2.0 introduced a conformance model that enables a system vendor to provide an overview of the conformance level of the system.

Implementing and supporting QTI version 2.x doesn't involve simply implementing all or none of the features provided by the specification. A system developer can decide for example to start by implementing the most requested interaction types, or to add all interaction types, but not support adaptive items, or to support only a restricted set of response processing templates and not to support MathML etc. The level of implementation can be described in an XML model provided by the specification. Also, by using an XSLT it is possible to create a profile of a QTI version 2.x item and determine if it requires features not provided by the system.

Full validation of items and assessments possible

Because the XHTML used in the item body is part of the schema provided by the specification, it is possible to use XML Schema validation on the complete item and or a complete assessment.

Response processing templates

The implementation of response processing functionality that covers the full richness of the response processing that can be used within a QTI 2.x item can be too big a challenge at first for a system vendor.

The specification defines three basic yet powerful response processing templates. If a system supports at least those three, it can choose to limit the implementation efforts of the building process while still enabling basic response processing functionality.

The first response processing template (match correct) compares the response of the candidate to the correct response set in the item. If they match, the score is set to 1 otherwise the score is set to 0. The second response processing template (map response) extends that by comparing the response of the candidate to a list of responses and scores.

The third response processing template defined by the specification is the map response point template, maps point related responses to scores if they are within described areas.

Use of external response processing

In some cases the response processing can be too complex to describe in QTI and/or might require the use of external systems. This could also be the case if an item needs to be scored by a human instead of a computer.

The QTI specification allows for the use of references to these external response processing engines.

QTI Lite

As with QTI 1.2, the QTI 2.x specification also defines a Lite version, which is basically a profile that limits the number of available options.

8.3 Innovative items and QTI

The ALTB project aims at the development of a methodology for the design of closed questions. Such a methodology is envisaged to consist of design requirements, design guidelines, components, design patterns and task structures including directives for task allocation. Design and development of these kind of innovative items is expensive and still rather labor intensive.

(Parshall et al., 2002) identified five dimensions in which items may be innovative:

1. item format: the sort of response collected from the examinee e.g. selected response or constructed response.
2. response action: the means by which the examinee provides his response e.g. key presses, mouse clicks, file upload.
3. media inclusion: the addition of non-text elements in the item.
4. level of interactivity: the extent to which an item type reacts or responds to the examinees input.
5. scoring method: how examinee responses are converted into quantitative scores.

This section will discuss the five dimensions of innovation within items and link the relevant functionality supported by the IMS Question and Test Interoperability specification to those dimensions. The definitions and exact syntax descriptions for the vocabulary used for interactions, response types etc. can be found in the document describing the information model of the IMS Question and Test Interoperability specification (IMS, 2006b).

8.3.1 Item format

The item format is the sort of response collected from the candidate. The QTI specification defines both selected responses and constructed responses. A response can have either, single, multiple or ordered cardinality. If a response is said to have single cardinality it means that it can contain only one value of the specified type, multiple means that the response is a container with a list of values of the specified type. Containers may contain multiple occurrences of the same value, the order of these occurrences is taken into account when the cardinality is of type ordered.

The specification defines nine different possible data types for responses: identifiers, pair, directed pair, point, boolean, integer, float, string, file.

identifier

An identifier is simply a logical reference to another object in the item, such as an item variable or a choice. Its most common use is to return the QTI internal identifier of an answer option selected in a multiple-choice question, the identifier of a hotspot selected, or a list of identifiers in case of an ordering question.

The following interactions can return an identifier as result of the interaction:

- choiceInteraction (single or multiple cardinality)
- graphicOrderInteraction (ordered cardinality)
- hotspotInteraction (single or multiple cardinality)
- hottextInteraction (single or multiple cardinality)
- orderInteraction (ordered cardinality)

pair

A pair value represents a pair of identifiers corresponding to an association between two objects. The association is undirected so (A,B) and (B,A) are equivalent. The response type is used in the matching interactions provided by QTI 2.x . In questions based on these interactions two lists of labels (indicated by their identifier) need to be matched. The associateInteraction uses text or graphics to represent the choices. The graphicAssociateInteraction uses hotspots on a graphic image to represent the choices.

The following interactions can return a pair as result of the interaction:

- associateInteraction (single or multiple cardinality)
- graphicAssociateInteraction (single or multiple cardinality)

directed pair

A directed pair value represents a pair of identifiers corresponding to a directed association between two objects. The two identifiers correspond to the source and destination objects.

The following interactions can return a directedPair as result of the interaction:

- gapMatchInteraction (single or multiple cardinality)
- graphicGapMatchInteraction (multiple cardinality)
- matchInteraction (single or multiple cardinality)

point

A point value represents an integer tuple corresponding to a graphic point. The two integers correspond to the horizontal (x-axis) and vertical (y-axis) positions respectively. The response type is typically used when the candidate has either to select one or more specific location(s) within a graphic (selectPointInteraction) or when the candidate has to position an image once or multiple times on top of a background image (positionObjectInteraction).

The following interactions can return a point as result of the interaction:

- selectPointInteraction (single or multiple cardinality)
- positionObjectInteraction (single or multiple cardinality)

boolean

A boolean value is either true or false. The response type is not used for True/False items (which use choiceInteractions), those return an identifier. It is used solely to indicate the use of a particular interaction to end an attempt.

The following interaction can return a Boolean as result of the interaction:

- endAttemptInteraction (single cardinality)

integer

An integer value is a whole number in the range from -2147483648 to 2147483647. This is the range of a two's-complement 32-bit integer. It is used to either indicate the number of times a media file in a mediaInteraction has been played, the exact value selected in a sliderInteraction or a value entered in a text box for a textEntryInteraction or extendedTextInteraction.

The following interactions can return an integer as result of the interaction:

- textEntryInteraction (single cardinality)
- extendedTextInteraction (single or multiple cardinality)
- mediaInteraction (single cardinality)
- sliderInteraction (single cardinality)

float

A float value is defined as a IEEE double-precision 64-bit floating point value. It is used both for free text entry (either textEntryInteraction or extendedTextInteraction) as well as for floating point slider responses.

The following interactions can return a float as result of the interaction:

- textEntryInteraction (single cardinality)
- extendedTextInteraction (single or multiple cardinality)
- sliderInteraction (single cardinality)

string

A string value is any sequence of characters. As a response type it is only used for free text entry fields in the textEntryInteraction or extendedTextInteraction.

The following interactions can return a string as result of the interaction:

- textEntryInteraction (single cardinality)
- extendedTextInteraction (single or multiple cardinality)

file

A file value is any sequence of bytes qualified by a content-type and an optional filename given to the file either uploaded by the candidate or as a result of the interaction of the candidate with a drawing. The content type of the file is one of the MIME types defined by RFC 2045-2048 Multipurpose Internet Mail Extensions (MIME). The file response type allows for the creation of items that request the candidate to create a response using an external application, for example a spreadsheet application and upload that file as response.

The following interactions can return a file as result of the interaction:

- drawingInteraction (single cardinality)
- uploadInteraction (single cardinality)

8.3.2 Response action

Depending on the combination of interaction types used within an item, the candidate needs one or more of the following means to provide a response: key presses, mouse clicks or mouse movement (drag/drop/drawing), file upload or other interactions like for example recording of a piece of video or audio in case of a custom interaction that requires such a response.

From an accessibility point of view, an assessment player should provide alternatives for either mouse actions or keyboard actions, for example a virtual keyboard that can be controlled using a mouse or key combinations that can replace mouse movements.

The QTI specification doesn't mandate the use of a specific response action. The interaction types themselves can be rendered in many different ways depending on the settings in the player or the preference of the candidate.

For example, an assessment item containing simple choice elements to create a multiple-choice question could be rendered as an item with radio buttons as defined in HTML. But the system rendering the items could also choose to define key combinations (A, B, C, etc) as a way to select or deselect answer options. The choice is not made in the description of the QTI assessment item. This is in line with the good development practice guideline that content and layout should be kept separated.

8.3.3 Media inclusion

Through the use of the `mediaInteraction` interaction or the `image` (``) tag as defined by XHTML an assessment item can contain numerous non text elements. The XHTML object element in QTI 2 is designed to support the graceful degradation (W3C, 2003) of media objects.

Say for example that an assessment item contains a visual presentation in the form of a Scalable Vector Graphic (SVG) file.

```
<object data="example.svg" type="image/svg+xml" width="400" height="100">  
  <object data="example.jpg" type="image/jpeg" width="400" height="100">  
    Alternative text which gets displayed if the SVG and JPG versions fail.  
  </object>  
</object>
```

The above example demonstrates the use of the XHTML object element for graceful degradation. If the browser is capable of handling the SVG, it will use that file. If it can't it moves on to the JPG image. If in this case it can't handle neither the SVG or the JPG, the alternative text is being displayed.

Fragments of assessment items can be included by reference allowing for flexible combination of static and dynamic elements within an item.

The stylesheet class can be used to assign a CSS stylesheet to an item, allowing for a flexible development of items where the exact positioning of elements within the item can be determined during run-time if desired. This makes it possible to adapt the look and feel of an item to the context in which it is being displayed or the needs of the candidate for example by providing bigger font settings.

8.3.4 Level of interactivity

The extent to which an item reacts or responds to the candidates input can be divided into two types of responses:

First of all there is the client side response to actions of the candidate. This relates to for example updating the displayed slider value when the candidate moves the slider, starting or stopping media files, updating the screen during drag and drop or selection actions by the candidate. These interactions are not specifically defined or described in the item. It is left up to the assessment player to handle those kind of system interactions. Usually the browser in which the items are displayed takes care of most of this based on the XHTML it receives from the rendering engine.

The second type of response by the assessment engine is triggered by the ending of an attempt by the candidate. That can be done either by submitting the answer, or by using the end attempt interaction (endAttemptInteraction). Submission of the answer triggers the response processing for the item as defined in the item. The end attempt interaction also triggers the response processing, but the assessment engine can detect the fact that the candidate didn't want to submit an answer and can determine what kind of other response is needed.

Response processing can trigger the display of modal feedback, i.e. feedback displayed in box or window which the candidate has to close before proceeding to either the next item or before attempting the same item again in case of adaptive item. It can also trigger the display of inline feedback which can be displayed anywhere within the body of the item and can remain visible for as long as needed. An item designer can hide previous feedback after a next attempt, or could decide to add more feedback with each attempt of the candidate.

Feedback can contain new interactions, which for example can lead the candidate in small steps to the correct response for the bigger problem set forward in the main item.

Besides interactivity for items, the use of preconditions and branch rules enable an item or assessment designer to create elaborate structures of items where the next displayed item is determined by the response of the candidate within the current item. A precondition is an optional set of conditions to be evaluated during the test, that determine if parts of a test are to be skipped and branch rules (branchRule) are an optional set of rules, evaluated during the test, for setting an alternative target as the next part of the test.

(P. Gorissen 2007a and b) describe two worked out examples of test scenarios that use branch rules and preconditions to build very interactive tests. The QTI descripti

8.3.5 Scoring method

The QTI specification provides support for very simple and straight forward scoring and for very sophisticated scoring structures. The specification defines a number of basic scoring templates as part of three provided response processing templates. Response Processing is the process by which the values of Response Variables are judged (scored) and the values of item Outcomes are assigned. The match correct response processing template sets the score to either 0 or 1 depending on the outcome of the test. The map response processing template maps the value of a response variable onto a value for the score, based on a provided mapping. The map response point processing template does the same thing but now based on a response of type point. Additional templates can be developed leaving the item designer with the simple choice of just selecting the correct template for an item.

QTI 2 provides an extensive set of response conditions and expressions for evaluation of the response of the candidate. Scoring can take into account things like number of attempts so far, previous attempts, time used to submit the item and many more.

8.4 Item development in existing systems

During the ALTB project the project members have gained both considerable experience with respect to the required features that development tools for closed questions should offer. The project team also gained insight in what would be important functional and user requirements for a computer-based development environment.

The development of virtual learning environments (VLEs) and assessment tools is a process that takes into account many factors. The functionalities offered by systems are the result of carefully balancing available resources and time on the side of the vendors or developers, requirements and requests by customers and competitive considerations. Users request an easy to use editing environment which has to be implemented by the vendors within the technical and financial constraints of that moment. This section will discuss the limitations found in many of the existing systems.

8.4.1 Workflow control

One of the problems the ALTB project encountered was that, in general systems don't offer the ability to distinguish between different phases in the design and development process of items. Also, systems usually only take into account one role during the design and development process.

In a typical situation within the ALTB project, the subject matter expert (SME) won't be designing or developing the closed questions himself. The SME provides an assistant who actually designs the items with the necessary background materials to design and develop items. The item designer makes a first draft design of what the item(s) might look like, the used interactions, feedback, scoring. The designs are sent to the SME who adds comments and feedback on that draft version and sends it back to the item designer.

Depending on the complexity of the item, the item is either created by data entry users or by programmers.

An item design and development environment should be able to support different roles and different phases during the design and development process.

8.4.2 Data Entry Templates

To date, the ALTB team hasn't found existing VLEs or even more specialized assessment management systems that offer the item design and development capabilities required by midsized projects on design and development of closed questions in higher education. All systems require the item developer to first choose an *item type* which then limits the options the item developer has during the development of that item. Usually a resulting restriction on an item is that it is not possible to mix different interaction types within an item.

Another major restriction of most tools currently available is the lack of control of how the item is presented to the candidate. The positioning of the different elements in the item body and the location and format of the display of feedback are usually restricted to one built-in design.

Designing attractive and interesting questions within these constraints is not possible.

A number of items created in a programming environment at Wageningen University (Aegerter-Wilmsen, 2005; Hartog et al., 2003; Schaaf, 2007; Sessink et al., 2007) show that from a candidate point of view these kinds of items look much more attractive than the ones regularly found in VLEs and this can positively stimulate the candidate while taking the formative test. These examples also show that the interaction with the candidate in itself aren't more complex than offered by the other tools, most of them are 'simple' matching questions of multiple-choice questions.

Response processing as defined within the IMS QTI specification is almost completely absent in most assessment centers available in VLEs. Usually, an item designer can assign scores or fractions of the total score to different answer combinations, but other factors like previous attempted answers, time taken to answer or more complex combinations of answers, cannot be implemented. Specialized assessment tools score better in that respect.

In essence, what current systems do is apply data entry templates with predefined combinations of the possible interactions and response processing structures and limit the options available to the item designers to a limited set chosen by the developer of the system.

8.4.3 Import / Export

Most systems offer export and import functionality to their own file format and possibly a number of other existing formats. There is no universal shared format across the board yet. Even IMS QTI, though it is the only vendor neutral format available, doesn't play that role yet. Vendors in general implement import filters for specific item bank formats if their customers have a sufficient amount of items in a specific format and they want to enable them to use those items in their system. Since there isn't a major publisher yet that provides a significantly big amount of items in IMS QTI format that need isn't that urgent yet for most vendors.

A problem related to building an import filter is that the vendor needs to map the set of functionalities provided by the format they are importing into their system.

In the case of IMS QTI, the problem for most vendors is that they have to map the rich functionality set supported by QTI to the templates they use internally. Because of that, building an export filter that exports items to QTI format is easier.

8.5 An Integrated Item Design and Development Environment

This section describes a not yet existing integrated environment that supports this process of design and development, the *Integrated Item Design and Development Environment* (IIDDE)

8.5.1 Service oriented architecture

The IIDDE is not intended to be a single monolithic system. Instead it is a set of coupled services provided by multiple software components linked together in a service oriented architecture. This approach allows for the re-use of existing components by the IIDDE. The R2Q2 project (R2Q2, 2006) for example offer webservices for the rendering and response processing of QTI items. A number of content repositories offer webservices eliminating the need to custom build that functionality for IIDDE.

8.5.2 Generic functionality

Like any other development environment the IIDDE needs support for things like version control, user management, metadata, search, import and export of assessment items, tests and other resources. Ideally this functionality should be provided by existing systems or web services.

8.5.3 Design and Development of items

The design and development environment has to take into account that there are different roles and stages within the design and development process of an item.

The three roles are:

- The subject matter expert (SME), the person who has the knowledge about the subject area for which items have to be created, but usually doesn't know how to create them;
- The item designer, the person who takes care of the design of the item and in most cases the creation of the item;
- A programmer, helps the item designer in cases where the item requires additional programming.

The design and development environment has an editing environment similar to what modern WYSIWYG-editors for web pages look like. The item designer can quickly *sketch* the item by dragging interactions unto the canvas of the item body.

The item designer can create and save multiple designs for one item and keep them together as one object.

After the designs are finished, the item designer sends a unique link to the SME. The SME uses that link to logon to the environment. The SME can then annotate the item designs to indicate what changes need to be made and which design he likes best. The item designer will automatically be notified (e.g. by mail) when that has been done.

Once the first version of the designs has been approved, the item designer creates a graphical representation of the response processing of the item. This allows the SME to approve the response processing without having to be exposed to the actual XML-code of the item.

Next the graphical representation is automatically converted by the system into the rules needed to handle the response processing.

In rare cases the WYSIWYG-editor isn't enough. In particular this will be the case when an SME wants very sophisticated items to be developed with graphical requirements that go beyond the capabilities of standard XHTML. In that case the item designer consults with the programmer during the design phase to make sure that the required design can be constructed afterwards by the programmer.

Sometimes the item designer will save items that have been designed as (QTI-)templates. The response processing structure of an item could be saved into a QTI response processing template, allowing him to re-use (parts of) the response processing of an item. The designer could also decide to save the item into a QTI item template used for cloning. He would then be able to re-use the item as a basis for new items.

8.5.4 Combining items

The two worked out examples in (P. Gorissen 2007a and b) show the use of multiple, interrelated items. Unlike a lot of summative tests, the example in (P. Gorissen 2007a) shows a test not created by (randomly) selecting a number of items from one or more big pools of items.

Here the items build a non-linear structure with branching options for the candidate eventually leading to the candidate demonstrating the ability to answer one complex question.

The example in (P. Gorissen 2007b) uses items picked from a number of different item pools. The structure of the flow in the test is not one easily designed or developed in current assessment tools.

The IIDDE has a test design module or web service that offers the functionality needed for this kind of test. Using a canvas similar to that of the Learning Activity Management System (LAMS, 2007) it enables the item designer to simply drag and drop items in position and link them by drawing lines between them.

8.6 Conclusions

This chapter explained how the IMS Question and Test Interoperability (QTI) specification can contribute to the development of innovative closed question items. When measured against the five dimensions of innovative items the QTI specification offers enough flexibility and supports enough functionality to be used as the basis for innovative closed question items. The worked out examples show that the IMS QTI not only enables the creation of individual items, but also can be used to describe very interactive structures of multiple individual questions.

The problem with the development of innovative closed question items at the moment is the mismatch between the functionality offered by existing VLEs and assessment systems and the needs as far as the design and development of these items is involved. They lack the needed support for the workflow and support for different roles during the design and development process and don't have the needed flexibility for the more sophisticated items.

The *Integrated Item Design and Development Environment* or IIDDE can realize the required functionalities based on web services in a service oriented architecture.

The system should be developed in association with existing SOA frameworks and initiatives like the e-Framework for education and research (e-Framework, 2007). Partners in the e-Framework are the UK's Joint Information Systems Committee (JISC, 2007), Australia's Department of Education, Science and Training (DEST, 2007), New Zealand's Ministry of Education (MoE, 2007) and the Dutch SURF Foundation (SURF, 2007).

The system should leverage the existing functionalities offered by for example the R2Q2 web services and the Learning Activity Management System (LAMS, 2007)

8.7 References

- Aegerter-Wilmsen, T. (2005). Digital Learning Material for Experimental Design and Model Building in Molecular Biology. PhD thesis, Wageningen University, Wageningen.
- DEST. (2007). Homepage of Department of Education, Science and Training. Retrieved feb 22 2007, from <http://www.dest.gov.au/>
- e-Framework. (2007). The e-Framework for Education and Research. Retrieved feb 22 2007, from <http://www.e-framework.org/>
- Gorissen, P. (2006). Quickscan QTI 2006. Retrieved feb 22 2007, from <http://www.gorissen.info/Pierre/QTI/>
- Gorissen, P. (2007a). The laboratory test Retrieved feb 14 2008, from <http://www.fbt.wur.nl/altb/QTIlabtest.htm>
- Gorissen, P. (2007b). The Pre-Bachelor test Retrieved feb 14 2008, from http://www.fbt.wur.nl/altb/QTIpre_bachelor.htm
- Hartog, R., Schaaf, H. v. d., & Verver, J. (2003). eLearning. *Agro Informatica*, 16(4), 9-11.
- IMS. (2006a). IMS Learning Resource Meta-data Specification Version 1.3 - Final Specification. . Retrieved feb 22 2007, from <http://www.imsglobal.org/metadata/>
- IMS. (2006b). QTI : IMS Question and Test Interoperability Assessment Test, Section, and Item Information Model Version 2.1 - Public Draft Specification Version 2. Retrieved feb 22 2007, from http://www.imsglobal.org/question/qti2p1pd2/imsqti_infov2p1pd2.html
- IMS. (2006c). QTI : IMS Question and Test Interoperability Assessment Test, Section, and Item Information Model Version 2.1 - Public Draft Specification Version 2 - section 7. Interactions. Retrieved feb 22 2007, from http://www.imsglobal.org/question/qti2p1pd2/imsqti_infov2p1pd2.html#section10076
- IMS. (2006d). QTI : IMS Question and Test Interoperability Integration Guide Version 2.1 - Public Draft Specification Version 2 - section 3. Content Packaging. Retrieved feb 22 2007, from http://www.imsglobal.org/question/qti2p1pd2/imsqti_intgv2p1pd2.html#section10003
- IMS. (2006e). QTI : IMS Question and Test Interoperability Meta-data and Usage Data Version 2.1 - Public Draft Specification Version 2. Retrieved feb 22 2007, from http://www.imsglobal.org/question/qti2p1pd2/imsqti_mdudv2p1pd2.html
- IMS. (2006f). QTI : IMS Question and Test Interoperability Specification Version 2.1 - Public Draft Specification Version 2. Retrieved feb 22 2007, from <http://www.imsglobal.org/question/>
- IMS. (2007). About the IMS Global Learning Consortium. Retrieved feb 22 2007, from <http://www.imsglobal.org/aboutims.html>
- JISC. (2007). the Joint Information Systems Committee. Retrieved feb 22 2007, from <http://www.jisc.ac.uk/>
- LAMS. (2007). Learning Activity Management System. Retrieved feb 22 2007, from <http://www.lamsinternational.com/>
- MoE. (2007). New Zealand's Ministry of Education. Retrieved feb 22 2007, from <http://www.minedu.govt.nz/>
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Practical considerations in computer-based testing. New York,: Springer-Verlag.
- R2Q2. (2006). Rendering and Response processing services for QTIv2 questions. Retrieved feb 22 2007, from <http://www.r2q2.ecs.soton.ac.uk/>
- Schaaf, H. v. d. (2007). Design of digital learning material for bioprocess-engineering education. PhD thesis, Wageningen University Wageningen.
- Sessink, O. D. T., Schaaf, H. v. d., Beefink, H. H., Hartog, R. J. M., & Tramper, J. (2007). Web-based Education in Bioprocess Engineering. *Trends in Biotechnology*, 25(1), 16 - 23.
- SURF. (2007). SURF Foundation. Retrieved feb 22 2007, from <http://www.surf.nl/smartsite.dws?id=5289&ch=ENG>

W3C. (2003). Web Accessibility Initiative - Glossary - Graceful Degradation. Retrieved feb 22 2007, from <http://www.w3.org/WAI/GL/Glossary/printable.html#def-transform-gracefully-1>