Modeling Assessment for Re-use of Traditional and New Types of Assessment.

Desirée Joosten – ten Brinke, Jan van Bruggen, Henry Hermans, Jan Burgers, Bas Giesbers & Rob Koper

Open University of the Netherlands
P.O. Box 2960
6401 DL Heerlen, The Netherlands
Telephone +31 (0)45 5762 758
Fax +31 (0)45 5762 802

Ignace Latour
CITO
P.O. Box 1034
6801 MG Arnhem, The Netherlands
Telephone +31 26 352 11 11
Fax +31 26 352 13 56
Abstract

In assessment the tendency is no longer focused on massive standardized testing with multiple choice questions based on knowledge acquisition. In the new learning approach assessment is integrated in learning and instruction and addresses the complex traits (the abilities, the characteristics in a specific domain) of students. To fit this new approach, new types of assessment are developed, like peer assessment or competence assessment. The development of these new assessments is an expensive and intensive activity. Exchange initiatives enable the re-use of materials that are developed by others. But several questions arise: Must assessments be completely re-used, and if not, what parts of an assessment can be re-used, and is re-use limited to particular item formats?

In this article we present a conceptual model (an 'educational model') for assessment, cast in a UML class model, that offers to support both new and traditional assessment types. The model is validated against the theory of Stiggins (1992) according to the development of performance assessments, the four processes framework of Almond, Steinberg and Mislevy (2001, 2003) and the QTI specification. The educational model for assessment gives new input to the alignment of the teaching, learning and assessment.
Assessment model

Modeling Assessment for Re-use of Traditional and New Types of Assessment.

Assessments are at the core of the educational process because they have a direct impact on the learning processes of students. Summative assessments help to establish whether our students have attained the goals set for them. Formative assessments provide prescriptive feedback to assist students in reaching their goals. Where instruction and assessment are considered as separate activities, assessment is often referred to as ‘testing’ (Birenbaum, 1996).

In this article we define assessment as all the systematic methods that can be used to gather information and evidence about student properties, based on a process, a product or the progress of a student, for the purposes of certification, placement or diagnoses in formative and summative contexts. This definition includes classical tests, examinations and questionnaires, as well as newer types of assessment, such as competence-based assessment, portfolio assessment and peer assessment.

These new types of assessment are typically embedded in an educational context, requiring more stipulation of the processes of assessment and relying on higher levels of student involvement in assessment (Sluijsmans, Brand-Gruwel, van Merriënboer & Martens, 2004). The shift to a perspective which is centered on student learning from a perspective which focused on the teacher is the greatest conceptual shift which has occurred in recent times in (higher) education (Boud, 1995). Assessments have a direct effect on students’ learning processes. The type of assessment expected by students, influences the way they learn, and thus assessment is like “the tail that wags the curriculum dog” (Hargreaves, 1989). ‘New learning’ is described by Biggs (1999) as a system in which teaching, learning and assessment interact and therefore requires that the curriculum objectives, the teaching and learning activities and the assessment tasks remain kept aligned. New types of assessments try to give an adequate answer to these new
ideas. Cizek (1997) emphasizes that the new assessment types are not replacements for traditional assessments, but that they give answers to different assessment questions than the traditional types. The new types give tutors and students with a deeper understanding of the student traits. Examples of the new types of assessment are portfolio assessment, competence assessment, performance assessment, self-assessment and peer assessment. In portfolio assessment students compile a portfolio to demonstrate evidence of, for example, personal growth. Portfolios are an effective way to bring assessment into harmony with instructional goals (Shavelson, 1992). If the learning objectives are the subject of the instruction, the assessment tasks must be a logical follow up of this instruction. Through observation during instruction and collecting some of the artifacts of instruction, assessment stems directly from the instruction. Performance assessment and competence assessment are often mentioned as synonyms, but, as Nedermeijer and Pilot (2000) emphasize, a performance assessment is that part of a competence assessment in which a person has to show a performance. Competence assessment is broader and the judgment that a student is competent in a domain can be achieved by using several assessment instruments (Birenbaum, 1996). The challenge is to select the assessment types that yield the appropriate evidence. A portfolio assessment, for example, provides a measure of individual student growth with respect to individual student goals. A multiple-choice exam provides a measure for knowledge acquisition in relation to a reference group. Both assessment measures are important providers of information of student traits and can be used in a competence assessment. Self-assessment and peer assessment are ways to involve students in assessment. Self-assessment involves learners taking responsibility for monitoring and making judgments about aspects of their own learning. It requires learners to think critically about what they are learning, to identify appropriate standards of performance and to apply them to their own
work. Peer assessment involves the responsibility to make critical judgments about the learning of peers and applying the standards to the work of these peers (Dochy, Segers, & Sluijsmans, 1999).

In the development of assessment systems, the evolution towards new types of assessment enables the creation of assessments that facilitate learning and instruction in ways that paper measures in the initial achievement of computerizing traditional multiple-choice tests cannot (Bennett, 2002). In this assessment creation, the developers have to cope with quality criteria, like authenticity, meaningfulness, fairness and educational consequences (for a full description of assessment quality criteria for competence assessment, see Baartman, Bastiaens and Kirschner, 2004). The question then is how we can combine these additional demands with the limited resources, such as time and persons, that are usually available. After all, the development of reliable and valid assessments is time-consuming and expensive. One way out of this dilemma is to design assessments in such a way that they can be shared amongst assessment developers and re-used in other contexts (Williamson, Bauer, Mislevy & Behrens, 2003). Here technology can play a role. Mislevy, Steinberg, Breyer, Almond and Johnson (1999) state that advances in technology allow more complex performances to be captured in assessment settings. They use an evidence-centered approach that presents a design framework that incorporates integrated structures for modeling the student traits, designing tasks, and extracting and synthesizing evidence. These technological advances are the basis for the use of assessments that are developed by others.

But technological improvements are not the whole answer to the re-use of assessments. Assessment developers must also share the same conceptual framework of the assessment domain to understand what can be exchanged or re-used. In the next section the interoperability
of assessments in relation to open specifications is described, and current initiatives in the interoperability of assessments and their limitations are given. From these limitations we state several requirements that in the long run any complete conceptual model should comply to. Such a conceptual model is referred to as an 'educational model', i.e. a model of an educational subsystem (see: http://hdl.handle.net/1820/275), in this case assessment. Examples of other subsystems that can be modeled are units of learning, portfolios, learning objectives, and curriculum structures.

**Interoperability of assessments**

For assessment experts a limited set of rules may be sufficient to ensure that each expert knows what to expect and how to work with the printed assessments that they exchange. If we require that these experts can exchange the assessments in an electronic form, using whatever software and hardware systems, interoperability enters the scene. Interoperability as used here is a capability of software systems to use the same formats for storing and retrieving information and to provide the same service on different hardware and software platforms. Hence parts of assessments, like assessment items or assessment descriptions, can be exchanged between experts. They all can edit, store and re-use them. In a computer-interpretable form, these assessments might be delivered to a candidate by a computer. The key issue here is to create and manage information in such a way that opportunities for exchange and re-use, either within or between institutions, are maximized (Miller, 2000). To reach such an ambitious goal a specification for exchangeability and interoperability of assessments is required. A specification prescribes, in a complete, precise, and verifiable manner, the requirements, design, behavior, or characteristics of a system (Beshears, 2003). One of the main benefits of a specification is that it
Assessment model

The leading specification for the exchange and interoperability of assessments is QTI, developed by the IMS Global Learning Consortium (2004a). One of the core concepts of this model is the assessment structure model (ASI) that defines Assessment, Section and Item layers. The QTI specification includes a set of XML bindings to describe questions and tests. It does so by (a) providing a well documented content format for storing items independent of the authoring tool used to create them; (b) supporting the deployment of items and item banks across a wide range of learning and assessment delivery systems and (c) enabling systems to report results in a consistent manner (Joosten-ten Brinke, Gorissen & Latour, 2005). The primary goal of this specification is to enable the exchange of questions (called ‘Items’) and tests (called ‘assessments’) between Learning Management Systems. QTI supports different types of questions and it is split up in two parts, the content of the evaluation part and the results from the evaluation part. Both parts can be used separately or together. The QTI interoperability information model and bindings are based on the four-process framework (Almond, Steinberg & Mislevy, 2000). Almond et al. (2000) discuss the relationships between the functions and responsibilities of these processes and the objects in the QTI information model. The QTI specification is more or less limited to those assessment types for which an unambiguous definition in technical terms can be specified. The structure of multiple-choice items proved to be well-suited for storage in item bank systems and delivery in digital format as the structure was offers a shared (controlled) vocabulary in which core concepts and ideas about a specific topic area can be expressed. Using open specifications means that many people look critically at another's work, resulting in a more stable, and ultimately more satisfactory, result.
not complex. The QTI specification offers good opportunities for exchange of items in standardized assessments. Test developers in an educational program may use colleagues’ multiple-choice items. For example, items about ‘knowledge of the learner’ in teacher education developed at university X may fit into the educational program of teacher education in university Y. The test developer who wants to use these items has to make sure that these items fit, based on learning objectives, their wording and format. Often these multiple-choice items are stored in item bank systems. By using a specification such as QTI to code them, these items may be exchanged between different platforms and presented in various format to students. The structure of the items must be comprehensive with regard to the domain to make them useful for domain specialists. Unfortunately, the implementation of the full QTI specification has proven to be difficult. In a recent review of software applications that claim to support QTI, Gorissen (2003) found that in almost all cases the support is restricted to the item layer of the ASI model. In addition, none of these applications supported all test formats defined in QTI. So, the interoperability is limited to “classical” multiple-choice items and their variations. Given the evolution towards new types of assessment, this is a problem for the realization of more advanced learning networks.

Other specifications

Other specifications with a close relation to assessment are IMS Learning Design (IMS LD, 2003), IMS ePortfolio (2004), IMS Learner Information Package (2001), and IMS Rubric (2004). IMS LD does not include assessment tools and strategies, but these can be included if assessments were referred to as content elements. The Educational Modeling Language (EML, the precursor of IMS LD) included educational content and test items integrated into the learning
design. In IMS LD these content and test items have been replaced by XHTML (with the namespace inclusion of 'global elements' to provide content interactions) and QTI respectively. In the new release of QTI the integration with IMS LD properties will be realized. IMS ePortfolio will support the possibilities of portfolio assessment in a technical way; it enables portfolios to be exchanged from school to work, it allows educators and institutions to better track competencies, it enhances the learning experience, and it improves employee development. This will all be related to the technical aspects of the portfolio as an artifact and not yet to the personalized information of portfolios of a candidate in an assessment. The Learner Information Profile specification is important to affirm the information of the learner in a population. The Rubric specification deals with the assessment of a portfolio; no other assessment types are addressed. An educational model for assessment can provide insight into gaps between the different specifications to support assessment exchange initiatives.

The assessment triangle and the four processes framework

There are already several initiatives to support the new types of assessment. These initiatives are based on insights into how people learn, how knowledge and knowledge structures develop and how they relate to the assessment of competences. These new insights are explained by an example of Pellegrino, Chudowski and Glaser (2001), who defined a new framework for assessment based on the assessment triangle (Figure 1) of ‘cognition’, ‘observation’ and ‘interpretation’.

[Insert Figure 1 about here ]
Here, cognition is a model of how a learner represents knowledge and develops competencies; observations are tasks or situations in which (complex) behavior can be observed, and interpretation is a means by which one can make sense of the observations. Pellegrino et al. (2001) provide several examples of new linkages, such as the use of concept mapping to assess knowledge structures (linking cognition to observation), or the use of Latent Semantic Analysis to interpret essays (linking observation to interpretation). Van Bruggen, Sloep, Van Rosmalen, Brouns, Vogten, Koper et al. (2004) describe latter use. In this model items (tasks) provide part of the evidence that is linked to the learning objective and must support decisions that are based on the assessment results. These items selected for observation should be developed with the purpose of the assessment in mind (i.e. going from cognition to observation). The evidence gathered still needs to be interpreted. This interpretation expresses how the observations derived constitute evidence about of the learner’s competencies.

An important consequence of the new foundations of assessment seems to be that any exchange of assessment has to entail all three points of the triangle rather than being limited to the exchange of the test items. Several questions arise from these deliberations: Must assessments be re-used completely, and if not, what parts of assessment can be re-used? Can all the assessment aspects (cognition, observation, interpretation) be described using specifications as mentioned above? Or have the specifications to be extended? Can we describe all sort of tasks and situations, or are we limited to particular formats such as multiple-choice items?

The conceptual assessment framework of Almond, Steinberg and Mislevy (2001, 2003) promotes reusability of both objects and processes and can thus provide a start in answering our questions. But there are some limitations to the model. It is developed with computer-based assessment in mind and is focused on the execution phase of an assessment. The limitation of the
QTI specification for assessment interoperability is that it is concentrating on the ‘observation’ point of the assessment triangle, where it offers support to rather traditional tasks. To include the other vertices a more encompassing model is needed. http://hdl.handle.net/1820/275

Requirements for an educational model for assessment

To develop an educational model for assessment there are several requirements (derived from Koper, 2001) that any complete conceptual model should in the long run comply to:

1. Flexibility: The assessment model must be able to describe assessments that are based on different theories and models.

2. Formalization: The assessment model must be able to describe assessments and its processes in a formal way, so that it is machine-readable and automatic processing is possible.

3. Reusability: The assessment model must make it possible to identify, isolate, de-contextualize and exchange useful objects (e.g. items, assessment units, competencies, assessment plans), and to re-use these in other contexts.

4. Interoperability and sustainability: Separation must be made between the description standards and interpretation techniques, thus making the assessment model resistant to technical changes and conversion problems.

5. Completeness: The assessment model must cover the whole assessment process, including all the typed objects, the relationship between the objects and workflow.

6. Explicitly typed objects: The assessment model must be able to express the semantic meaning of different objects within the context of an assessment.
7. Reproducibility: The assessment model must describe assessments so that repeated execution is possible.

8. Medium neutrality: The description of an assessment, where possible, must be medium neutral, so that it can be used in different (publication) formats, like the web, or paper and pencil tests.


We developed a first version of an educational model for assessment that can be extended to provide a broader basis for interoperability specifications for the whole assessment process of construction till evaluation. The model allows stronger embedding of assessments in educational practice and it caters for new types of observation and interpretation. At this moment the requirements used are restricted to completeness (the whole assessment process #5), interoperability (#4), flexibility (#1), reproducibility (#7), reusability (#3), and as an initial step towards formalization (#2), the model is cast in terms of UML class diagrams. The next section describes the method of the model construction, followed by a section describing and illustrating the model itself.

Method

The development of an educational model for assessment elaborates the previous work of Hermans, Van den Berg, Vogten, Brouns, and Verhooren (2002). Hermans et al. discuss the design of a framework for testing interactions, which has resulted in a preliminary UML domain model. The method used to develop the conceptual model consists of three main activities: (1) the development of a first version of the conceptual model, (2) validation of the first version of
the model by cases and literature, (3) adjustment of the model on the basis of the validation results.

The first version of the model was constructed in a series of sessions (brown paper sessions) with a small project team (two assessment experts, one UML expert, one educational technologist, one scribe and one project leader) and six assessment experts from educational and specialized testing institutes in the Netherlands. A chairman facilitated these expert sessions, as well as the majority of activities during the sessions. The UML modeler was available to translate expert knowledge into UML and to have these models directly confirmed by the experts. A running record was kept of the developing model, the data dictionary and any background material. Finally, a scribe took notes of the sessions and did regular reporting.

As input to the modeling process, expert knowledge was elicited using techniques such as brainstorming concepts, concept sorting and clustering. These techniques are suited to eliciting declarative domain knowledge only. Subsequent discussions took place with assessment experts about all aspects and characteristics of assessments to derive a first version of an educational model for assessment. During the construction of the model it was clear that as many experts on assessment were involved as there were ideas about the model. The model developed through discussions about the interpretation level (institutional, program or course) of the model, the selection of the UML classes and their characteristics, and the extent of details in the model.

In line with the learning technologies specifications and requirements #3 (reusability) and #7 (reproducibility), this model has been cast in UML class diagrams (dealing with #6). Previous experiences with UML class diagrams (Hermans et al, 2002; Hermans, Manderveld, & Vogten, 2003) also indicated the use of UML. UML class diagrams give a clear and unambiguous description of the elements and structures of the domain. It has become more or less the standard
modelling language in the field of object-oriented system design (Warmer & Kleppe, 2001). UML class diagrams are suitable to model declarations on the existence of processing rules instead of modeling the process part itself in a procedural way. In a description, attached to the model, the processes are described. In addition, business rules are set up to define or constrain aspects of the assessment model in relation to other aspects in the model.

After the expert sessions, the model was fixed for a first validation round in which the model’s fit was tested using assessment cases and existing assessment frameworks. A team of experts and UML modelers analyzed documents about assessment on how to express the identified assessments and concepts in the model. Whenever a problem was encountered, a standardized problem description was compiled and any solutions were proposed using a change request. To validate in particular the requirements of flexibility (#1) and completeness (#5), existing assessment frameworks and assessment cases have been gathered externally to test the model’s fit.

**The assessment model**

The model that was constructed in UML is depicted in Figure 2. The UML classes are the squares and the lines indicate the kind of relation between the UML classes.

[Insert Figure 2 about here ]

The model is built on several sub-models, each fitting to a different stage in the assessment process as depicted in Figure 3.
In the following discussion the concepts that are part of the model are written in italics. The characteristics described are omitted from the figures for reasons of readability.

Assessment design

The reasons for using assessments are expressed in the stage of assessment design. The concepts and their relations in assessment design are represented in Figure 4.

To draw up the policy of an educational institute on the subject of assessment a framework has to be developed in which the educational vision must be made clear (Van Brakel & Heijmen-Versteegen, 2003). An example of such an institutional policy might be a prescription on assessment types, or that only standardized mass exams may be used, or that students must be involved in the development of assessments. Within the scope of this assessment policy one or more assessment plans can be designed. An assessment plan includes the basic assumptions for an assessment. It defines the decision rules which set down how a decision can be taken for a candidate by a decision maker. Also, part of the assessment plan is the blue-print. The assessment function in the assessment plan stipulates the reason that a decision has to be taken. Possible assessment functions are diagnosis for individual candidates, the formation of groups, selection or certification. The assessment plan addresses one specific
population. The assessment plan also prescribes which assessment types can be used for units of assessment. These must be assessment types that fit with the assessment policy of the institute.

Part of the assessment plan is the assessment scenario. An assessment scenario determines which units of assessment a candidate has to do, which units of assessment the candidate can choose, in what sequence, and in what time. The units of assessment are described in the unit of assessment definition. The last, but very important part of the assessment plan is the trait. This is the abstract concept of the characteristics of the candidate on which decisions have to be taken. These traits are important for educational contexts because they give the criteria for education in terms of level and direction. A trait belongs to the population for which the assessment plan is set up and it can be decomposed into complex traits and elementary traits.

Item construction

The model of the concepts and their relations in item construction are represented in Figure 5.

[insert Figure 5 about here ]

In this stage the concepts elementary trait and population described in the previous stage are the guiding lines for the construction of items. The elementary trait will be measured by indicators. Often direct observation of a characteristic of a student (trait) is not possible. For example, by observing a teacher in the classroom we cannot directly measure if the teacher understands how students learn. To that end indicators are specified that give evidence for this trait. These indicators are measurable descriptions of the trait. A score on an assessment has a meaning for a trait, but it is directly based on scores on the underlying indicators by a calculation
rule. For every *indicator* items can be developed that are suitable for the *population* to which the assessment plan belongs. Items in this model have to be interpreted broadly. For example, it may concern a multiple-choice item with four answering options, or a task in which a candidate has to show some a performance. Candidates can provide answers in a number of formats; it might be a construction, a selection out of response possibilities, or the demonstration of a skill. These item types are named *construction item*, *selection item* and *demonstration item*. An item has a *prompt*, the explicit message to the candidate that makes clear what is expected (within the item) of the candidate, a *case text*, a description of a context in which item has to be answered, *hints* and *feedback*. For all the relevant indicators an item should have a *rating instruction*. The *rating instruction* specifies for each item what the characteristics should be for a correct answer in relation to the indicator.

*Assessment construction*

The third stage is that of assessment construction. The model of the concepts and their relations in this stage are represented in Figure 6.

[insert Figure 6 about here ]

The central concept in this stage is the *unit of assessment*. A *unit of assessment* is a measurement instrument based on a *unit of assessment definition*. In this definition it is the composition rules that describe the structure of the assessment. It might be possible to use the composition rules in advance, but it is also possible to use the rules during the assessment session, in case of adaptive assessment for instance. The *assessment type* of a unit of assessment
must meet the required types defined in the assessment plan. The unit of assessment is a composite of items that will be presented to a candidate and is based on a unit of assessment definition. The characteristics of a unit of assessment definition are the session time, the number of candidates that may participate, the way the unit of assessment is presented to the candidate, what roles the candidate have to fulfill in the unit of assessment and rules about the composition of the assessment, rules prescribing what items may be used and in what order and rules that specify how the final score on a unit of assessment will be calculated. The definition defines which trait will be addressed in a specific unit of assessment (unit of assessment trait) and which indicators are used for this purpose (unit of assessment indicator). The items used in a unit of assessment are selected because they measure a specific indicator. They might measure other indicators as well, but that is not of interest for this unit of assessment. Therefore the assessment item is defined for a specific item in a specific unit of assessment. The assessment item indicator gives the specific indicator that is meant to be measured with this item. The scale prescribes which values can be given to that assessment item indicator based on the rating instructions.

Assessment run

As soon as the unit of assessment is composed, the assessment can be delivered to the candidates. The model of the concepts and their relations in this assessment run stage are represented in Figure 7.

[insert Figure 7 about here ]
Depending on the kind of assessment, a candidate must provide responses, or demonstrate or present something to an assessor. *Units of assessment* are presented to *candidates*, who can be *individual persons* or *groups*. The actual presentation of one or more units of assessment to the candidates is done during *assessment sessions*. Each session has a date, a starting time and a stop time. In this *session* each candidate has an *assessment take* which specifies the medium in which the unit of assessment is presented, as well as the available candidate roles. The output of a session are *item responses*. An *item response* can be an answer to a question, a performance or a report.

*Response rating*

The next stage is that of response rating. The model of the concepts and their relations in this so-called assessment run stage are presented in Figure 8.

[insert Figure 8 about here ]

After an *assessment take* the *item responses* must be assessed by an assessor. The assessor can be a computer, a teacher, peer candidates or even the candidate. The *assessor* provides a *rubric score* that addresses the *assessment item indicators*. To do so the assessor uses transformation rules to come from a rubric value to a rubric score, to an *assessment indicator score* and to a *trait score*. The *assessment indicator score* addresses the *unit of assessment indicator*, while the *trait score* addresses the *unit of assessment trait*, the *scoring prescription* and a scoring instruction.
**Decision making**

The last stage is that of decision making. The model of the concepts and their relations in this assessment run stage are represented in Figure 9.

\[\text{[insert Figure 9 about here ]}\]

At the end of the process a *decision* must be made that is based on the *score* of a *candidate* on a certain *assessment take*. The decisions that can be made are described in the *assessment plan* (see assessment design stage). Often, the person who makes the decision is a teacher, but in general, this is the institute where the candidate is enrolled. The *decision* is based on *decision rules*.

Now that the model is described, it is interesting to validate the model on the basis of earlier work in the field.

**Validation of the assessment model**

Here, we present results of the first validation round of the model. In this validation round we have validated the assessment model to the requirements of flexibility (#1) and completeness (#5). Therefore existing assessment frameworks and assessment cases have been gathered. First, the model was validated on the guidelines of Stiggins (1992) for performance assessments. Second, the model was validated on the four processes framework of Almond et al. (2001, 2003). Third, the model was validated on the QTI. After describing these three validations, the model’s fit is described by using a performance assessment in teacher education.

**Validation 1**
Stiggins (1992) provides a framework with guidelines using four steps, for the design of performance assessments, i.e. evaluations of the application of knowledge and skills in authentic learning situations. The steps are 1. the specification of a performance to be evaluated, 2. the definition of what needs to be evaluated, 3. the development of tasks used to elicit that performance, and 4. the design of a scoring and recording scheme for results.

The information that must be specified in step one contains the kind of decisions that can be made, who the decision makers are, how the results of the assessment are being used and for what population the assessment is meant. These aspects are caught in the *assessment policy*, the *assessment plan*, the *population*, the *function of assessment* and *decision*. The function of an assessment depends on the decisions that an institute wants to make.

Stiggins’ second step is focused on the subject of assessment, the characteristic of the student that must be evaluated. In this step, Stiggins demands that the assessment designer should know what kind of tasks a candidate has to fulfill. The task may be the delivery of a product, but it might also be that a process of constructing the product is more informative. Therefore it is important to define the *rating instruction* at an early stage. Stiggins emphasizes that this *rating instruction* must be derived from the authentic situation. This step of Stiggins is found in the classes *assessment plan*, *population*, *trait*, *indicator* and *rating instruction*.

The third step of Stiggins points to the development of the assessment tasks. The performance and the standards for good practice in real life are the basis for the tasks in a performance assessment. The assessment type that best fits the objective has to be selected. For example, this may mean that the use of multiple-choice questions suits best in a knowledge domain and that portfolio assessment suits best if students have to present evidence of several processes. In this step, the assessment designer also has to decide on the number of
measurements that are necessary to make an assessment reliable. Furthermore, a decision must be taken whether or not students will be made aware beforehand that they are being assessed. The classes in the model that correspond to the information in the third step are assessment function, trait, indicator, item, assessment scenario and unit of assessment.

Finally, in the fourth step a plan for judging the assessment is specified. Here, the assessment designer describes the type of scoring (holistic or analytic), who will or may act as assessors, and the exact method of scoring. The judgment is constrained by the types of decisions that are specified in the first step. Classes from the model that correspond to this step are item response, assessor, scoring prescription, indicator score, assessment indicator score, trait score, assessment plan, decision and decision rules.

Validation 2

The second validation has been done on the basis of the four processes framework of Almond et al. (2001, 2003). Whereas the Stiggins approach was derived from an assessment design perspective, the Almond et al. view is derived from viewing assessment as a process in which an administrator, responsible for setting up and maintaining the assessment, and a candidate, the person whose traits are being assessed, are actors in a system. Although the model of Almond et al. is phrased in terms of computer-based testing, it can be used in a broader sense.

The four processes framework, comparable with the basic assumptions of our assessment model, is defined from the perspective of the re-use of functional objects of assessments in different contexts. The difference is that the framework of Almond et al. starts from the processes and the assessment model, as the name says, from the concepts in assessment. The framework consists of six different types of models that specify the materials, capabilities, and other information needed by the processes necessary to deliver a particular assessment:
(1) the student model: what complex of knowledge, skills or other traits of the student is assessed?

(2) the task model: what tasks or situations should elicit those behaviors?

(3) the evidence model: a set of instructions for interpreting the result of the task,

(4) the assembly model: a set of instructions for assembling the assessment,

(5) the presentation model: how to present a particular task in a particular delivery environment and

(6) the delivery model: a container for things that affect the entire assessment.

These models are the basis for the processes that, according to Almond et al., take place in assessment. The processes in the four processes framework are the ‘activity selection process’ (responsible for selecting and sequencing tasks, including items, set of items, or other activities), the presentation process (responsible for presenting the task to the candidate and capturing responses), the evidence identification process (responsible for identifying the essential characteristics of the response (the ‘work product’) that provide evidence about the candidate’s traits) and the evidence accumulation process (responsible for the update of the belief about the candidate’s trait). The last two phases are called ‘response processing’ and ‘summary scoring process’ in later work of Almond, Steinberg and Mislevy (2003).

The elaboration of the scoring process in the model is based on the same principles. First the scores on items are related to the indicators the items measure and later these measurements on the indicator level are summarized in an assessment score. In the terms of Almond et al., the
indicators are estimates of participant proficiency(ies). One or more assessment scores gives us information about the trait of the candidate.

The information stored in an assessment plan, like the blue-print, is the guiding input for the activity selection process in the model of Almond et al. The candidate’s current knowledge, skills and abilities (in our model named traits) is caught in the student model of Almond et al. The observable variables mentioned are our indicators. The concepts trait, item, prompt, hint, item formats, instructions are part of the task mentioned by Almond et al. in the task model. The rating instruction is comparable with the evidence rules of Almond et al. These rules (rubrics for example) describe how to identify and evaluate essential characteristics of the item response (in the terms of Almond et al., work product). The scoring record and the weight of evidence are of use in the scoring prescription. The scoring prescription gives input to an assessor to evaluate an item used in an assessment. It indicates the contribution of this item to the total amount of information the unit of assessment will give of the candidate’s trait. Two types of feedback are mentioned by Almond et al.: task-level feedback (an immediate response to the candidate’s action in a particular task, independent of evidence from other tasks) and summary feedback (a report about our accumulated belief based on evidence from multiple tasks). In the model the first type of feedback is coupled to item and the second type is a characteristic in the trait score.

Components of the assessment construction stage of the model are closely related to the activity selection process and the presentation process. The assessment scenario uses a set of instructions for assembling the assessment. This is caught by Almond et al. in the assembly model. In the presentation process it is described how a particular task has to be presented. In the model this is described in the assessment session. Information of the history of the candidate (collection of completed tasks, state of the scoring, and so on) is not described separately in the
assessment model. In Almond et al.’s model, this is mentioned as the ‘examinee record’. The decision is positioned in the ‘activity selection process’. This process makes a decision about what to do next, based on the current beliefs about the participant or other criteria.

**Validation 3**

The third validation has been done with the Question and Test Interoperability specification (2004). The four processes in the framework of Almond et al. are described as complementary processes that are meant to work with the data structures defined in QTI. As Almond, Steinberg and Mislevy (2001) phrased it, “the IMS standard for interoperability among assessment deliver and authoring systems must support both the standard multiple-choice and essay-type items, which form the core subset of current practice, and provide sufficient flexibility to grow into the advanced constructed-response items and interactive tasks we envisage as the future of assessment” (p.1). The assessment model we developed has a broader objective in the sense that it will support not only computer-based assessments but also all other presentation media and the newer types of assessment must be covered in the model. The QTI specification elaborates the assessment items in detail. A long list of item types are described, for example simple multiple choice items, hotspot items, or match items. While the QTI specification was known before the assessment model was construed, we decided to leave the detailed specification of the selection item types to QTI. In the model the item types are put in three conceptual containers: *selection items*, *construction items* and *demonstration items*. The last container is not available in QTI. The smallest exchangeable assessment object within this specification is the item. This is defined in the same way in the assessment model. A *candidate* can, however, only react to an item in a *unit of assessment* in an *assessment session*. This unit of assessment can consist of one item, but has more information than the item has, like composition rules. In the QTI specification it is possible
for a candidate to react in a item session. The unit of assessment is not within the scope of the QTI specification. The feedback component of QTI also consists of two types, modal and integrated. Modal feedback is shown to the candidate after response processing has taken place and before any subsequent attempt or review of the item. Integrated feedback is only shown during subsequent attempts or review. These two types refer to the feedback on item level in the model. As a consequence of the scope of QTI, the feedback on assessment level is not available in the QTI specification.

**Performance assessment**

In this section we will describe a performance assessment in teacher education in terms of the assessment model. In the description of the assessment the corresponding concepts of the assessment model are placed in italics between brackets.

Standards for teacher education are nationally established. Standards are the basic assumption for the curriculum of any institute for teacher education (*assessment policy*). One of the standards might be ‘The teacher works effectively in cooperation with other professionals and adults in order to promote learning’ (*trait*). This standard is translated in lower level standards. One is ‘by the end of the program, students will demonstrate that they are able to work cooperatively in the classroom with other professionals and adults, such as parents and classroom assistants (*complex trait*). On the lowest level this means that students have to demonstrate that they can manage a parent-teacher interview with the desired outcomes (*simple trait*). For this trait a performance task (*item; demonstration item*) is developed which requires students (*candidate*) to interview the parents and develop a written report of that interview. The policy of the institute (*assessment policy*) prescribes that students have to do the parent-teacher interviews twice, in the
first year and in the third year (assessment plan). The interviews are observed by two assessors (assessor). The total organization is described in an assessment plan of the institute. If a student fails the interview, the report may still be written, but the interview has to be done again. Other assessments are not dependent on the result of the interview (assessment scenario).

Accompanying the performance task is a list of performance criteria (rating instruction) for the report and the interview. The scoring rubric (scoring prescription) for the report ranges from 4 points till 0 points, in which 4 means ‘the report is easy to read and uses appropriate format. It has correct spelling, capitalization, punctuation and usage errors. It is written in complete sentences and uses paragraphs correctly. The advice given to the parents is present or it is clear that no advice was necessary.’ And zero means ‘The student failed to attempt the report.’.

A comparable scoring rubric is available for the evaluation of the interviews. After the interview (unit of assessment 1) the student makes the report (unit of assessment 2). For a specific student (candidate) the interview takes 10 minutes (assessment session; item response) and the report must be delivered to the assessor within a week after the interview (item response). The assessor gives two figures (assessment indicator scores), one for the interview and one for the report. If the mean of those figures is above 6, the mean will be the end score (trait score) for this competence. The institute has decided in the assessment plan that students whose scores are all above 6 may start the final exam.

Conclusion

After several sessions with assessment experts and a UML modeler we agreed upon an educational model that covers diverse assessments, like portfolio assessment, group assignments and self-assessment. External validation of the model, by international assessment experts, are
now taking place, to conclude if the model covers the full range (requirements #1 and #5). The model offers an understandable and descriptive insight for experts in the assessment domain with knowledge of UML (requirement #6). For those assessment experts who are not familiar with UML, this is not the case. In the description of the model a short UML explanation should be useful.

In the development the focus was on the requirements of flexibility (#1), formalization (#2), reusability (#3), interoperability (#4), completeness (#5), and reproducibility (#7). The first validation round concludes that other frameworks can be included in the assessment model. The QTI specification is much more developed in depth on the item concept. With this in mind beforehand, we did not model class diagrams in detail on item level. The assessment model can play a role in the discussion with respect to further development of QTI for assessments. After the external validation the relationship between further developments in QTI and the assessment model can be used. Another shortcoming in the current assessment model is the lack of statistical and psychometric information. This information plays an important role in the four processes model. In the model this information is often put in several rules. This solution relates to the discussion mentioned in the method section in that there might be more depth in the model. For the structure rules between assessment scenario and unit of assessment, an example of more detailed modeling is depicted in Figure 10.

[Insert Figure 10 about here]

In conclusion, an educational model has been constructed that matches the new approach of assessment, in which new assessment types fit. The model is not competitive to other
frameworks of assessment, but hopes to be supplementary in the development of assessment frameworks and the possibilities of interoperability. The model can give input into the discussions on ‘assessments’ and ‘sections’ in the further development of the QTI specifications. In the next step, in which a demonstrator will be developed to visualize what the possibilities of the model are in the framework of exchangeability, the IMS LIP specification can be used for the classes population and the candidate. Moreover, to ‘prove’ whether the model meets the demand of formalization (#2) tooling has to be built, which is part of the next step. The construction of the assessment model is one of the models in Educational Modeling (Koper & Van Es, 2004). By following the next steps ‘tooling’ and ‘use, evaluation and dissemination’, a step towards an open specification is made for the use of new assessment types in line with teaching and learning. The IMS LD specification is focused on the modeling the teaching and learning processes in a unit of learning. The QTI specification focuses on assessment, and the assessment model can give direction to the use of the IMS LD specification and the QTI specification to align teaching, learning and assessment.
References


In McGreal (Eds.), *Online education using learning objects (open and flexible learning)*. Canada: RoutledgeFalmer.

Retrieved March 10, 2004 from http://www.ariadne.ac.uk/issue24/interoperability/


Figure 1. The assessment triangle of Pellegrino, Chudowski and Glaser (2001, p. 44).

Observation  Interpretation

Cognition
Figure 2. The assessment model.
Figure 3. The stages in the assessment process.
Figure 4. Assessment design
Figure 5. Item construction.
Figure 6. Assessment construction
Figure 7. Assessment run.
Figure 8. Response processing
Figure 9. Decision making
Figure 10. Assessment plan, assessment scenario and unit of assessment modeled in more depth.