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

Running head: EDUCATIONAL MODEL FOR ASSESSMENT

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

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

Correspondence concerning this paper should be addressed to Desirée Joosten, Open University of the Netherlands, Educational Technology Expertise Centre, P.O. Box 2960 6401 DL Heerlen, The Netherlands, voice: ++31-45-5762758, fax: ++31-45-5762802, e-mail: desiree.joosten-tenbrinke@ou.nl

Abstract

In the new learning approach assessment is integrated in learning and instruction and addresses complex traits (the abilities, the characteristics in a specific domain) of students. To match 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 promise to reduce those efforts by the re-use of materials. But they also raise questions: Is it a complete assessment or are there specific parts of an assessment that can be re-used? And is re-use limited to particular item formats? In order to support the re-use of both new and traditional assessment types an educational model for assessment is developed.

In this article we present this model. The model is validated against Stiggins' (1992) guidelines for the development of performance assessments, the four-process framework of Almond, Steinberg and Mislevy (2001, 2003), a specification for the exchange and interoperability of assessments and performance assessment as a new type of assessment. The educational model for assessment gives new input to the alignment of the teaching, learning and assessment.

Key words: assessment, exchange, design, 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 (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 performance assessment, portfolio assessment and peer assessment.

These new types of assessment are typically embedded in an educational context, require more stipulation of the processes of assessment and rely on higher levels of student involvement in assessment (Sluijsmans, Brand-Gruwel, van Merriënboer, & Martens, 2004). The shift from a perspective that focused on the teacher to a perspective that is centered on student learning is the greatest conceptual shift that has occurred in recent times in (higher) education (Boud, 1995). '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 are 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 claim to give tutors and students a deeper understanding of the student traits. Examples of the new types of assessment are portfolio assessment, performance assessment, self-assessment and peer assessment. In portfolio assessment students compile a portfolio to demonstrate evidence of, for example, personal

growth. In a performance assessment, a person has to show competent behavior. Self-assessment and peer assessment are ways to involve students in assessment. These types require learners to think critically about what they are learning, to identify appropriate standards of performance and to apply them to their own work (Dochy, Segers, & Sluijsmans, 1999). The new assessment types have great potential, but problems in terms of quality criteria and resources as well. The assessment 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) and the development of reliable and valid assessments is time-consuming and expensive.

The question then is how we can combine these additional demands with the limited resources. 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 alone will not solve problems of re-use or exchange of assessments. Assessment developers must also share the same conceptual framework of the assessment domain to understand what can be re-used or exchanged. 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 (i.e. programs, courses, study tasks), portfolios, learning objectives and curriculum structures.

Interoperability of assessments

Whenever we require that assessment experts can exchange assessments in an electronic form, whatever software and hardware systems they use, interoperability enters the scene. Interoperability, as used here, is the 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. Once interoperability is achieved, parts of assessments, like assessment items or assessment descriptions, can be exchanged between experts. They all can edit, store and re-use them. 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). Although the use of computers in assessment as an administration tool has a long history (Epstein, & Klinkenberg, 2001), the possibilities of computers are not fully elaborated in the context of exchange in the phase of assessment design. To support exchange a specification for interoperability and exchangeability 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 offers a shared (controlled) vocabulary in which core concepts and ideas about a specific topic area can be expressed. In this article we will present such a vocabulary for assessment. For the classical assessment types, like multiple choice assessments and open-ended question assessment, such a vocabulary already exists. The next paragraph describes this specification.

The QTI specification

The leading specification for the exchange and interoperability of assessments is the Question and Test Interoperability (QTI; IMS Question & Test Interoperability, 2006). The primary goal of this specification is to enable the exchange of questions (called 'Items') and tests (called 'Assessments') between Learning Management Systems. The Question and Test Interoperability specification describes questions and tests by (a) providing a well documented content format for storing items independent of the authoring tools that were 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 QTI interoperability information model is based on the four-process framework (Almond, Steinberg, & Mislevy, 2000). Almond et al. (2000) discuss the relationships between the functions and responsibilities of the processes and the objects in the QTI information model. The QTI specification is limited to those assessment types for which an unambiguous definition in technical terms can be specified. Examples of these assessment types are multiple choice items, open-ended questions, matching items. The simple structure of these items makes them well-suited for storage in item bank systems and delivery in digital format. The QTI specification supports the 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 will match with 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 match, 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 formats to students. The structure of the items must be comprehensive with regard to the domain to make them useful for domain specialists. The implementation of the full QTI specification has proven to be difficult. In a review of software applications that claim to support QTI, Gorissen (2003) found that in almost all cases the support was restricted to the item layer, leaving the Assessment and Section layer aside. The latest version of the QTI specification (IMS QTI, 2006) is a minor upgrade that provides additional features on the Assessment and Section level. This may enhance the storage and exchange of complete tests with a strong focus on computer-based assessment, but does not provide a solution for new types of assessment like portfolio assessment or peer assessment.

The assessment triangle

Although QTI does not support the new types of assessment, there are several other initiatives that will 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 Pellegrino, Chudowski and Glaser (2001), who defined a new framework for assessment based on the assessment triangle of ‘cognition’, ‘observation’ and ‘interpretation’. 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 student essays (linking observation to interpretation). In this framework items (tasks) provide part of the evidence that is linked to the learning objective and they must support decisions that are based on the assessment results. The

items that are 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 makes clear how the collected observations constitute evidence about the learner's competencies.

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

In the conceptual assessment framework of Almond, Steinberg and Mislevy (2001, 2003) assessment is viewed as a process in which an administrator, who is responsible for setting up and maintaining the assessment, and a candidate, whose traits are being assessed, are actors in a system. The framework 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 it 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. In the next section we describe the requirements that such a model needs to match.

Requirements for an educational model for assessment

An educational model for assessment that enhances reusability needs to match the same requirements (derived from Koper, 2001) that any complete conceptual model should in the long run comply to:

1. Flexibility: The assessment model can describe assessments that are based on different theories and models.
2. Formalization: The assessment model describes assessments and its processes in such a formal way that it is machine-readable and automatic processing is possible. The formalization gives the possibility to extend the model if new developments in assessment arise.
3. Reusability: The assessment model supports identification, isolation, de-contextualization and exchange of useful objects (e.g. items, assessment units, competencies, assessment plans) and their re-use in other contexts.
4. Interoperability and sustainability: The assessment model distinguishes the description standards from the interpretation techniques, thus making the model resistant to technical changes and conversion problems.
5. Completeness: The assessment model covers the whole assessment process, including all the typed objects, the relations between the objects and the workflow.
6. Explicitly typed objects: The assessment model expresses the semantic meaning of different objects within the context of an assessment.
7. Reproducibility: The assessment model describes assessments in such a way that replicated execution is possible.
8. Medium neutrality: The educational model for assessment, where possible, supports the use of different media, in different (publication) formats, such as computerized assessments on the web or paper and pencil tests.

9.Compatibility: The assessment model matches available standards and specifications.

We developed an extensible educational model for assessment to provide a broader basis for interoperability specifications for the whole assessment process from construction to evaluation. The model allows a tight embedding of assessments in educational practice and it caters for new types of observation and interpretation. During the development of the model we gave priority to the following requirements: completeness (covering the whole assessment process #5), interoperability (#4), flexibility (#1), reproducibility (#7), reusability (#3), and formalization (#2). Requirements on explicitly typed objects (#6), medium neutrality (#8) and compatibility (#9) were given less priority because they either have no direct influence on the quality of the conceptual model.. The next section describes the method of model construction, which is followed by a descriptions and illustrations of the model itself.

Method

The development of an educational model for assessment elaborates the original work by Hermans, Van den Berg, Vogten, Brouns, and Verhooren (2002). The conceptual model was developed in three consecutive steps: (1) development of a first version based on expert input, (2) validation of the first version using 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 with five assessment experts from educational and specialized testing institutes in the Netherlands and a small project team (two assessment experts, one modeling expert, one educational technologist, one scribe and one project leader).

The model was cast in the Unified Modeling Language (UML), a standard language for specifying, visualizing, constructing, and documenting concepts or artifacts. This is in line with the learning technologies specifications and facilitates compliance to requirements #3

(reusability), #7 (reproducibility), and #6 (explicitly typed objects). 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). Previous experiences with UML class diagrams (Hermans et al, 2002; Hermans, Manderveld, & Vogten, 2003) also favored the use of UML. UML class diagrams are suitable to model processing rules in a declarative way, rather than modeling the process itself in a procedural way. In a description, attached to the model, the processes are described. Business rules are defined to constrain aspects of the assessment model and how they relate to each other. After the development of the model the model was validated and adjusted.

Results

In this section, first the result of the model development, the educational model for assessment, is given; second the results of the validation studies of this model are described.

The assessment model

The model is built on several sub-models, each matching a different stage in the assessment process as depicted in Figure 1. In the assessment design the objectives for the assessment are clarified. Decisions in this stage influence the elaboration of the next stages.

[Insert Figure 1 about here]

The assessment model is constructed in UML. UML classes are represented as squares and the lines indicate the kind of relation between the UML classes. In the following discussion the concepts that are part of the model are in italics. The characteristics described are omitted from the figures to increase readability.

Assessment design

The reasons for using assessments are expressed in the stage of assessment design. The challenge in assessment design is to select the assessment types that yield the appropriate evidence of students' competence, skills or knowledge. A competence assessment, for example, can consist of a portfolio assessment, that provides a measure of individual growth with respect to individual goals, in combination with a multiple-choice exam that provides a measure of knowledge acquisition. Both assessment measures are important providers of information of student traits and both can be used in a competence assessment. The concepts and their relations in assessment design are represented in Figure 2.

[Insert Figure 2 about here]

The *assessment policy* of an educational institute is the basis for the development of an assessment (Van Zutven, Polderdijk & De Volder, 2004). This framework enumerates *assessment types* that are allowed according the policy of the institute. 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. An example of an assessment plan is an assessment to measure writing skills. The plan stipulates the *decision rules* that set down how a *decision maker* will come to a decision on a candidate. The *assessment function* in the *assessment plan* stipulates the purpose of the decision. *Assessment functions* include diagnosis of individual candidates, formation of groups, selection or certification. The *assessment plan* addresses one specific *population*. The *assessment plan* prescribes which *assessment types* can be used for units of assessment. These must be *assessment types* that match the *assessment policy* of the institute. The *assessment scenario* is part of the assessment plan. An *assessment scenario* determines the

mandatory and optional *units of assessment* for a candidate, as well as their sequence and time schedule. 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 will be taken. These *trait*s are important for educational contexts because they give the criteria for education in terms of level and direction. A *trait* is determined in advance for 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 3. The main concept in this stage is the item.

[Insert Figure 3 about here]

In this stage the concepts *elementary trait* and *population* that were described in the previous stage are the guiding lines for the construction of *items*. Indicators measure the elementary trait. Often, however, 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 whether the teacher understands how students learn. To that end, *indicators* are specified that provide evidence on the trait. These *indicator*s 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 applying a calculation rule on the scores. For every *indicator* items can be developed that are suitable for the *population* to which the assessment plan applies. The term Item in this model has to be interpreted in a broad sense. For example, it applies to a multiple-

choice item with four answering options, as well as to a task in which a candidate has to show a performance. Candidates can provide answers in a number of formats, such as 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 usually has a *prompt*, a *case text*, *hints* and *feedback*. The *prompt* is the explicit message to the candidate that makes clear what is expected (within the item) of the candidate. In unannounced workplace observations it is possible that the prompt is not given. The *case text* is a description of a context in which the item has to be made. *Hint*s and *feedback* are both instruments to give the student supportive information, the hint beforehand, and feedback afterwards. For all relevant indicators an item must have a *rating instruction*. The *rating instruction* specifies for each item the characteristics of 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 presented in Figure 4.

[Insert Figure 4 about here]

The central concept in this stage is the *unit of assessment*. This is a composite of items that will be presented to a candidate based on a *unit of assessment definition*. In this definition the composition rules describes the structure of the assessment. Composition rules may be used in advance to generate an assessment, as well as dynamic during assessment sessions to select new items, for instance in adaptive assessment. The *assessment type* of a unit of assessment are restrained to the *types* that were defined in the assessment plan. 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, the possible roles the candidates have to fulfill in the *unit of assessment* and several rules. These rules are 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 assessed in a specific unit of assessment (*unit of assessment trait)* and which indicators are used to 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 irrelevant in the context of 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 the *assessment item indicator*.

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 5. The central concept in this stage is the *session*.

[Insert Figure 5 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 person*s or *group*s. 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. During 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 6.

[Insert Figure 6 about here]

After an *assessment take* an assessor must assess the item responses. 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 get 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 7.

[Insert Figure 7 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 kind of 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*.

This initial model was put to the test in a number of validation studies that are reported below.

Validation of the assessment model

The validation studies were conducted to test if the model satisfied the requirements of flexibility (#1), formalization (#2), reusability (#3), interoperability (#4), completeness (#5), and reproducibility (#7). Descriptions of existing assessment frameworks and assessment cases were gathered for this validation. A team of experts and UML modelers analyzed assessments and tried to express the identified assessments and concepts in the model. Whenever a problem was encountered, a problem description was compiled and any solutions were proposed using a change request.

Here, we present the results of the validation studies of the model. First, the model was validated on Stiggins' (1992) guidelines for performance assessments. Second, the model was validated on the four-process framework of Almond et al. (2001, 2003). Third, the model was validated on the Question and Test Interoperability specification. Fourth, describing a performance assessment in teacher education using the models terminology tests the model's expressiveness. Finally, the adjusted model was sent to international experts on assessment and UML modeling.

Validation 1

Stiggins (1992) provides guidelines for the design of performance assessments, i.e. evaluations of the application of knowledge and skills in authentic learning situations. The steps in this framework 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. In our model 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 knows what kind of tasks a candidate has to fulfill. A 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 is modeled in the classes *assessment plan*, *population*, *trait*, *indicator* and *rating instruction*.

The third step addresses 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 matches the objective has to be selected. For example, this may mean that multiple-choice questions are selected in a knowledge domain and that portfolio assessment is selected when students have to present evidence of their competences. In this step, the assessment designer also has to decide on the number of measurements that are necessary to make an assessment reliable. Furthermore, one has to decide whether or not students will be

made aware beforehand that they are being assessed. The classes in our 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), the persons 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*.

This first validation study shows that the educational model for assessment can describe assessments that are constructed according to the framework of Stiggins (1992) and that the model covers the whole assessment process as described by Stiggins. Therefore, this validation study is a positive indication that the model meets the requirements of flexibility (#1) and completeness (#5).

Validation 2

The second validation was done on the basis of the four-process 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-process 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 educational model for assessment, 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-process 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 that the items measure and later these measurements on the indicator level are summarized in an assessment score. In terms of Almond

et al., the indicators are estimates of participant proficiency(ies). One or more assessment scores give 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 used characteristics 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 that the unit of assessment will give about the candidate's trait. Almond et al. mention two types of feedback: 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 the 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. Almond et al. catch this in the assembly model. The presentation process describes how a particular task has to be presented. In our model this is described in the *assessment session*. Information on 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 referred to 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. In our model this is described by the *composition rules*.

The conclusion of this validation study is that the four-process framework of Almond et al. (2001, 2003) can be described using the educational model for assessment. This validation study is a further indication that the model meets the requirements of flexibility (#1) and completeness (#5).

Validation 3

The third validation was done on 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) phrase 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 QTI specification elaborates the assessment items in detail. A long list of item types are described, including simple multiple choice items, hotspot items, match items and several others. Since the QTI specification was published before the assessment model was construed, we could decide to leave the detailed specification of the selection item types to QTI. In the model the item types are therefore 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 reacts to an item in a *unit of assessment*

within an *assessment session*. This unit of assessment can consist of one item, but contains more information than the item, such as composition rules. *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.

This validation study has a somewhat different character than the studies before, because in the development of the model a part of the QTI specification has been incorporated in the educational model for assessment. This validation study indicates that that the requirement of formalization (#2) can be met, because of the possibilities of this type of incorporation.

Validation 4. Performance assessment

In this section we use the assessment model to describe a performance assessment in teacher education. 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 and they are the basic assumption for the curriculum of any institute for teacher education (*assessment policy*). An example standard in teacher education is ‘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 interview twice, in their first year and in their third year (*assessment plan*). Two assessors (*assessor*) observed the interviews. 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 to 0 points, in which 4 means 'the report is easy to read and uses appropriate format. It has correct spelling, capitalization, punctuation and absence of 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.' A zero rating 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 prepares the report (*unit of assessment 2)*. For a specific student (*candidate)* the interview takes ten minutes (*assessment session; item response*) and the report must be delivered to the assessor within a week after the interview (*item response*). The assessor assigns two ratings (*assessment indicator scores)*, one for the interview and one for the report. If the mean of the ratings is above 6, the mean will be the end score (*trait score)* for this competence. The institute has stated in the assessment plan that students whose scores are all above 6 may start their final exams.

This validation study gives evidence that the educational model for assessment can describe a complete performance assessment and therefore indicates that the model can meet the requirements of flexibility (#1) and completeness (#5).

External review of the educational model for assessment

After several sessions with assessment experts and a UML modeler and the internal validations we agreed on a consolidated educational model that covered various types of assessments, like portfolio assessment, group assignments and self-assessment. The model then was presented to a number of international experts in assessment and UML modeling outside the development group for review. The comments of these experts mainly addressed modeling peer assessments and group assignments. These experts found the model useful and indicated that they could model assessment series and plans in an effective and satisfactory manner. Their comments and suggestions were incorporated in the model.. From this review we conclude that the model is understandable and useable for experts in the assessment domain with knowledge of UML.

Conclusion and discussion

In this article we have described the development of an educational model for assessment and formulated a number of criteria such a model should meet.. The model itself was cast in terms of UML object classes in order to facilitate meeting the requirements of reusability (#3), reproducibility (#7), and the use of explicitly typed objects (#6). In order to validate the model we selected assessments based on different theories and several assessment frameworks and tried to describe them using our model. Since each of the validation studies indicated that the model met the requirements of flexibility (#1) and completeness (#5), we conclude that the model can describe assessments that are based on different theories and models (requirement #1) and that

the model covers the whole assessment process, including all the typed objects, the relation between the objects and the workflow (completeness ##5). The review by experts on assessment and UML modeling also hinted to this conclusion. By the possibility of incorporating a part of the QTI specification, the model makes a start in support of the requirement of formalization (#2). This does not mean that the model is directly machine-readable, but a first step is made. Moreover, to 'prove' whether the model meets the demand of formalization (#2) tooling will be built. The model will be subject to further investigation on the formalization requirement. 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 (IMS LD, 2003) is focused on 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.

The semantic meaning of different objects within the context of an assessment (requirement #6) might differ for different assessment background. In the model new terms for assessment objects are defined to overcome this problem. Future use of the model in different backgrounds has to show whether this requirement holds. The educational model for assessment will comply with the requirement of medium neutrality (#8), but depends on the media defined in the type of assessment. The requirement of compatibility (#9) will be reached if the model matches available standards and specifications.

Although the experts on assessment and UML were positive about the model, the model can be improved by a short UML explanation for assessment experts without UML knowledge.

Experts without UML expertise experienced difficulties in reading the model. Another shortcoming in the current assessment model is the lack of statistical and psychometric information. This information plays an important role in the four-process model. In our model this information is often formulated in several rules. As already mentioned in the method section of this article, the model can be extended here. For the structure rules between assessment scenario and unit of assessment, an example of more detailed modeling is depicted in Figure 8.

[Insert Figure 8 about here]

In conclusion, an educational model has been constructed that matches the new approach of assessment, and that can be used to describe new assessment types. The model is not competitive to other frameworks of assessment, but supplementary in the development of assessment frameworks and their interoperability. The model can provide input to the further development of the 'assessments' and 'sections' parts of the QTI specifications and to other specifications with a close relation to assessment. These are specifications on learning design (IMS LD, 2003), portfolio's (IMS ePortfolio, 2005), information about learners (IMS Learner Information Package, 2005), and scoring rubrics (IMS Rubric, 2005). An educational model for assessment can provide insight into gaps between these different specifications to support assessment exchange initiatives.

Acknowledgement: We like to thank the following experts for their valuable contribution to the development of the educational model for assessment: Dr. G. Maris and Drs. D. W. Schönau and (Cito, the Netherlands), I. Heijmen-Versteegen MSc and M. B. C. Jaspers (Fontys Polytechnic University, the Netherlands), Dr G. Moerkerke (OUNL, the Netherlands), Dr V. van

Deventer (University of South Africa, SA), Dr A. B. Steyn (University of Pretoria, SA), Dr R.Young (CETIS, UK), Dr S. Lay (UCLES, UK). We also like to thank B. Wilkinson MSc (University of Maastricht, the Netherlands) for comments on an earlier draft.

References

Almond, R. G., Steinberg, L., & Mislevy, R. J., (2001). A sample assessment using the four-process framework. CSE Report 543. Center for study of evaluation. Retrieved November 15, 2005 from http://www.cse.ucla.edu/CRESST/Reports/TECH543.pdf. Los Angeles: University of California.

Almond, R. G., Steinberg, L., & Mislevy, R. J. (2003). A four-process architecture for assessment delivery, with connections to assessment design. CSE Report 616. Center for study of evaluation. Los Angeles: University of California.

Baartman, L. K. J., Bastiaens, T. J., & Kirschner, P. A. (2004). *Requirements for Competency Assessment Programmes.* Paper presented at the Onderwijs Research Dagen. The Netherlands: Utrecht.

Bennett, R.E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, *1*, *1*. Retrieved November 15, 2005 from http://www.jtla.org.

Beshears, F.M. (2003). *Open Standards and Open Source Development Strategies for e-Learning.* Presentation for IS224 Strategic Computing and Communications Technology 10/2/2003. Retrieved October 11, 2004 from http://ist-socrates.berkeley.edu/~fmb/events/sakai-2004-01-12/IS224-2003-10-02.ppt. UC Berkeley: Educational Technology Services.

Biggs, J. B. (1999). *Teaching for Quality Learning at University.* Buckingham: Society for Research in Higher Education & Open University Press.

Birenbaum, M. (1996). Assessment 2000: towards a pluralistic approach to assessment. In M. Birenbaum, & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp.3-29). Kluwer Academic Publications.

Boud, D. (1995). *Enhancing learning through self-assessment.* London: Kogan Page.

Cizek, G. J. (1997). Learning, achievement, and assessment: constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment: learning, achievement, and adjustment* (pp. 1-32). San Diego: Academic Press.

Dochy, F., Segers, M., & Sluijsmans, D.M.A. (1999). The use of self-, peer-, and co-assessment in higher education: a review. *Studies in Higher Education*, *24*, 331-350.

Epstein, J., & Klinkenberg, W. D. (2001). From Eliza to Internet: A brief history of computerized assessment. *Computers in Human Behavior*, *17*, 295-314.

Gorissen, P. (2003). *Quickscan QTI.* Retrieved November 15, 2005 from http://www.digiuni.nl/digiuni//download/35303.DEL.306.pdf. Utrecht: De Digitale Universiteit.

Hermans, H., Berg, van den, B., Vogten, H., Brouns, F., & Verhooren, M. (2002). *Modelling test-interactions.* Educational Technology Expertise Centre (OTEC). Open University of the Netherlands.

Hermans, H., Manderveld, J.M., & Vogten, H. (2003). Educational modelling language. In W. Jochems, J. van Merriënboer, & R. Koper (Eds). *Integrated E-learning*. London: Kogan Page.

IMS ePortfolio (2005). *IMS ePortfolio Specification.* Version 1.0. Final Specification. IMS Global Learning Consortium, Inc. Retrieved November 15, 2005 from http://www.imsglobal.org/ep/index.html

IMS LD (2003). *IMS Learning Design Specification.* Version 1.0. Final Specification. IMS Global Learning Consortium, Inc. Retrieved March 22, 2005 from http://www.imsglobal.org/content/learningdesign/

IMS LIP (2005). *IMS Learner Information Package.* Version 1.01. Final Specification IMS Global Learning Consortium Inc. Retrieved November 15, 2005 from http://www.imsglobal.org/profiles/

IMS Rubric (2005). *IMS Rubric Specification.* Version 1.0. Final Specification. IMS Global Learning Consortium, Inc. Retrieved November15, 2005 from http://www.imsglobal.org/ep/epv1p0pd/imsrubric_specv1p0pd.html

IMS Question & Test Interoperability (2006). *IMS Question & Test Interoperability.* Version 2.1. Public Draft Specification. IMS Global Learning Consortium, Inc. Retrieved May 2, 2006 from http://www.imsglobal.org/question/index.cfm.

Joosten – ten Brinke, D., Gorissen, P., & Latour, I. (2005). Integrating assessment into e-learning courses. In E. J. R. Koper & C. Tattersall (Eds.), *Learning Design: a handbook on modelling and delivering networked education and training* (pp. 185-202). Springer Verlag.

Koper, E. J. R. (2001). Modelling Units of Study from a Pedagogical Perspective: the pedagogical meta model behind EML (http://eml.ou.nl/introduction/docs/ped-metamodel.pdf).

Koper, E. J. R., Pannekeet, K., Hendriks, M. & Hummel H. (2004). Building communities for the exchange of learning objects: theoretical foundations and requirements. *ALT-J, Research in Learning Technology*, *12*, *1*, 21-35.

Koper, E. J. R. & van Es, R. (2004). Modeling units of learning from a pedagogical perspective. In McGreal (Eds.), *Online education using learning objects (open and flexible learning).* Canada: RoutledgeFalmer.

Miller, P. (2000, June). Interoperability. What is it and Why should I want it? *Ariadne*, *24*. Retrieved March 10, 2004 from http://www.ariadne.ac.uk/issue24/interoperability/

Mislevy, R., Steinberg, L. S., Breyer, F. J., Almond R. G., & Johnson, L. (1999). *Making sense of data from complex assessments.* Retrieved November 15, 2005 from http://www.education.umd.edu/EDMS/mislevy/papers/MakingSense.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: the science and design of educational assessment.* Washington, DC: National Academy Press.

Shavelson, R. J., Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2003). Evaluating new approaches to assessing learning. Retrieved October 21, 2003 from http://www.cresst.org/reports/R604.pdf

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. *Innovations in Education & Teaching International*, (*41*), *1*, 60 – 79.

Stiggins, R.J. (1992). Het ontwerpen en ontwikkelen van performance-assessment toetsen. [Design and development of performance assessments]. In J.W.M. Kessels & C.A. Smit (Eds.) Opleiders in organisaties/Capita Selecta (afl. 10, pp. 75-91). Deventer: Kluwer

Van Zutven, G., Polderdijk, M., & De Volder, M. (2004). *Handboek Toetsplanontwikkeling in competentiegericht onderwijs.* [Handbook Assessment Development in Competency-based education]. Utrecht: Digitale Universiteit.

Warmer, J., & Kleppe, A. (2001). *Praktisch UML.* [Practical UML]. (2nd ed.) Amsterdam: Addison Wesley Longman Nederland BV. (Original work published 1999).

Williamson, D. M., Bauer, M., Mislevy, R. J., Behrens, J. T. (2003). *An ECD Approach to Designing for Reusability in Innovative Assessment.* (Unpublished work) Retrieved September 29, 2003, from http://www.ets.org/research/dload/aera03-williamson.pdf.

Figure Caption

Figure 1. The stages in the assessment process.

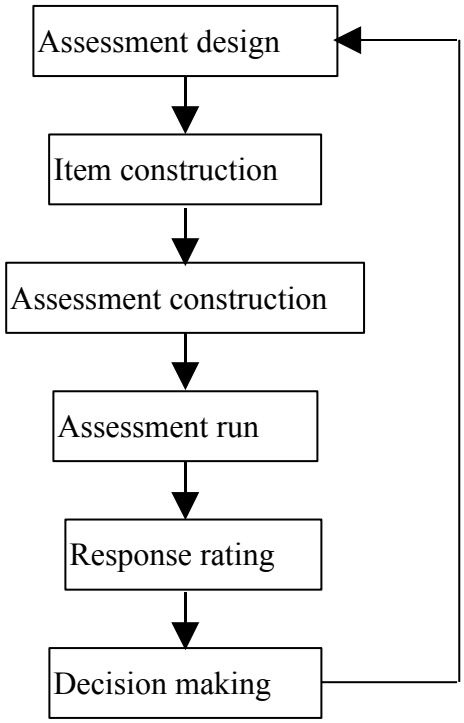

Figure 2. Assessment design

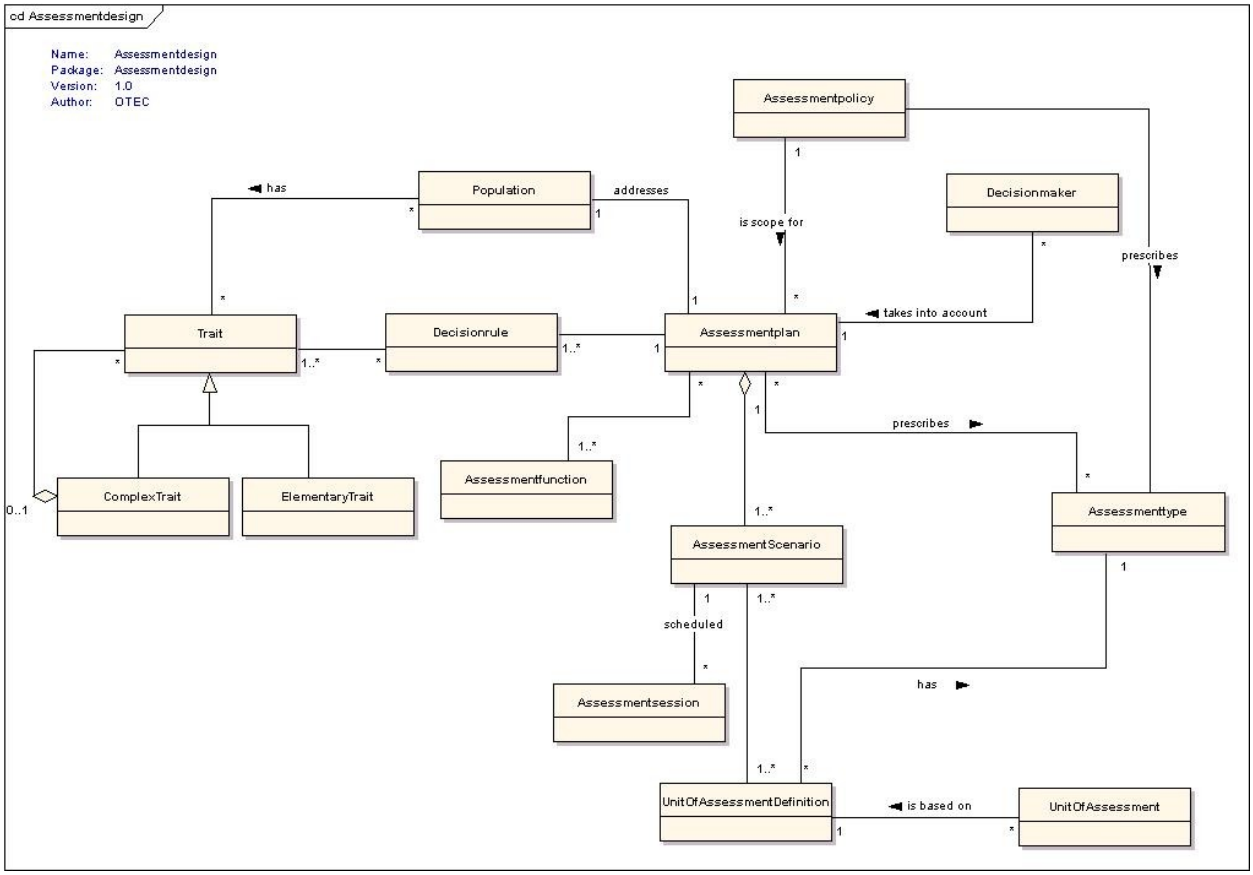

Figure 3. Item construction.

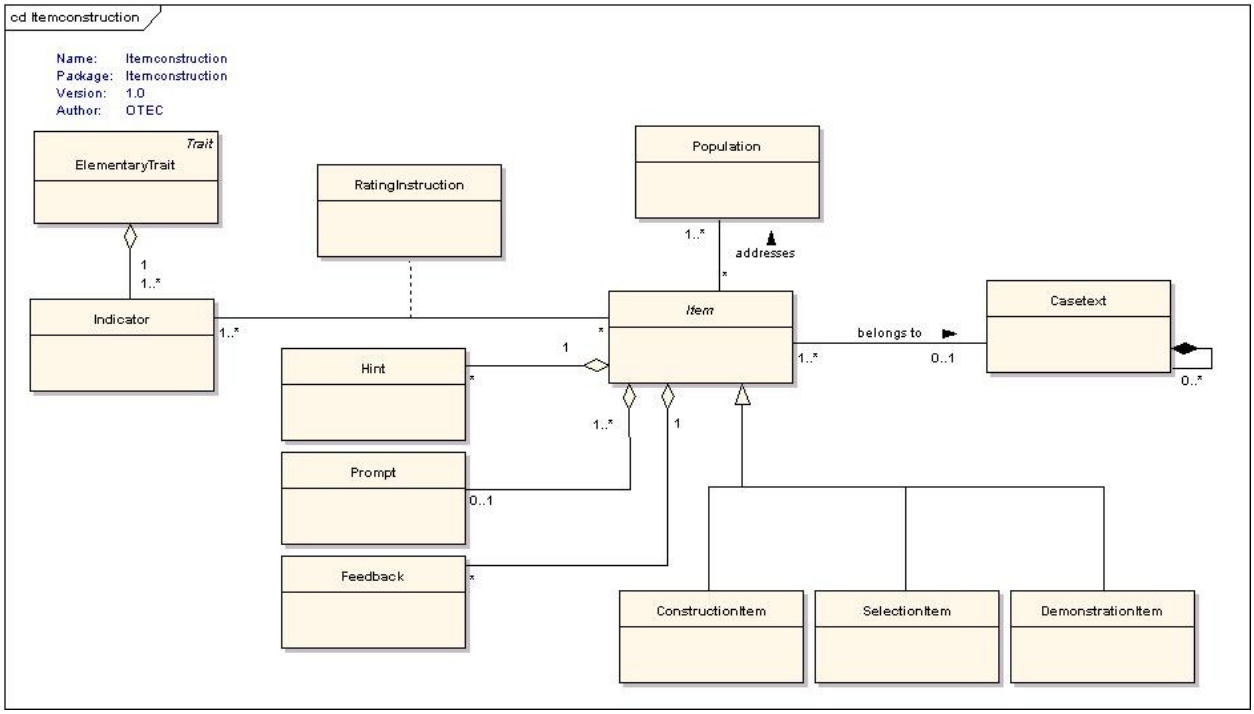

Figure 4. Assessment construction

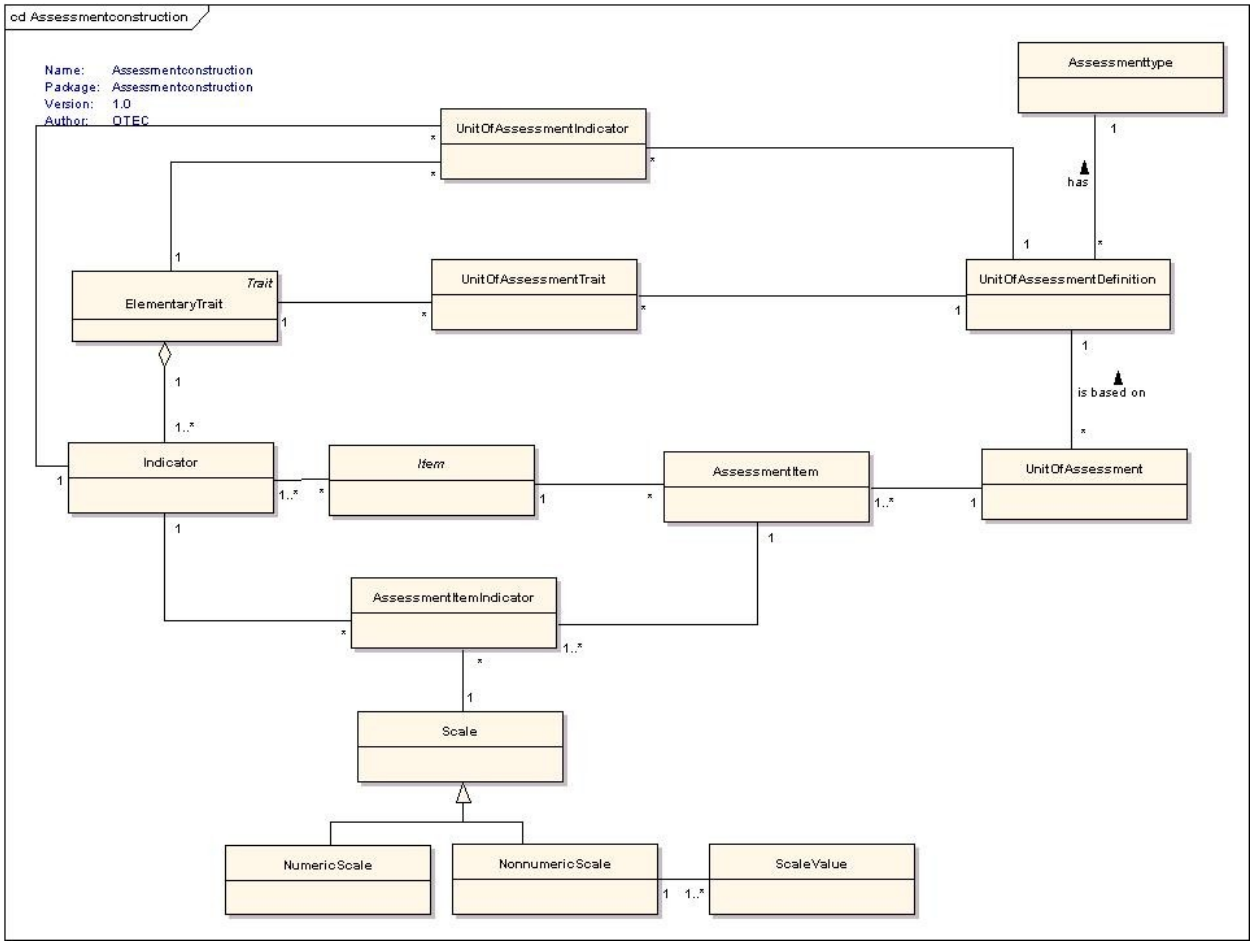

Figure 5. Assessment run.

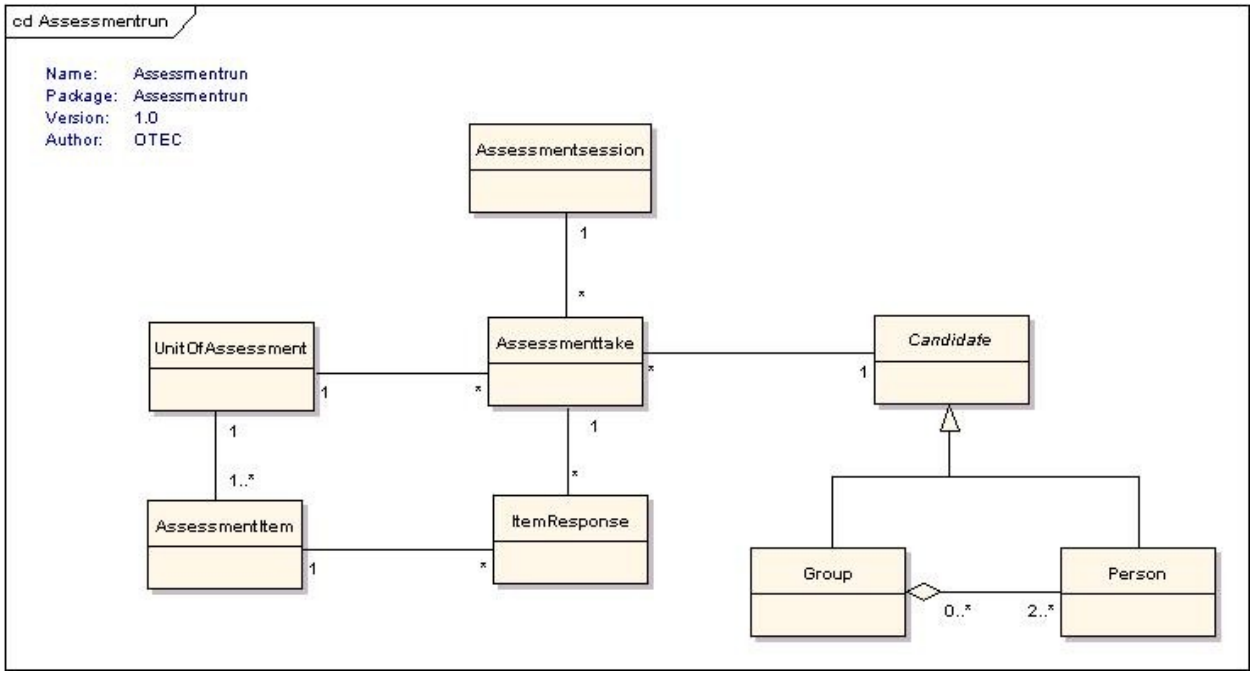

Figure 6. Response processing

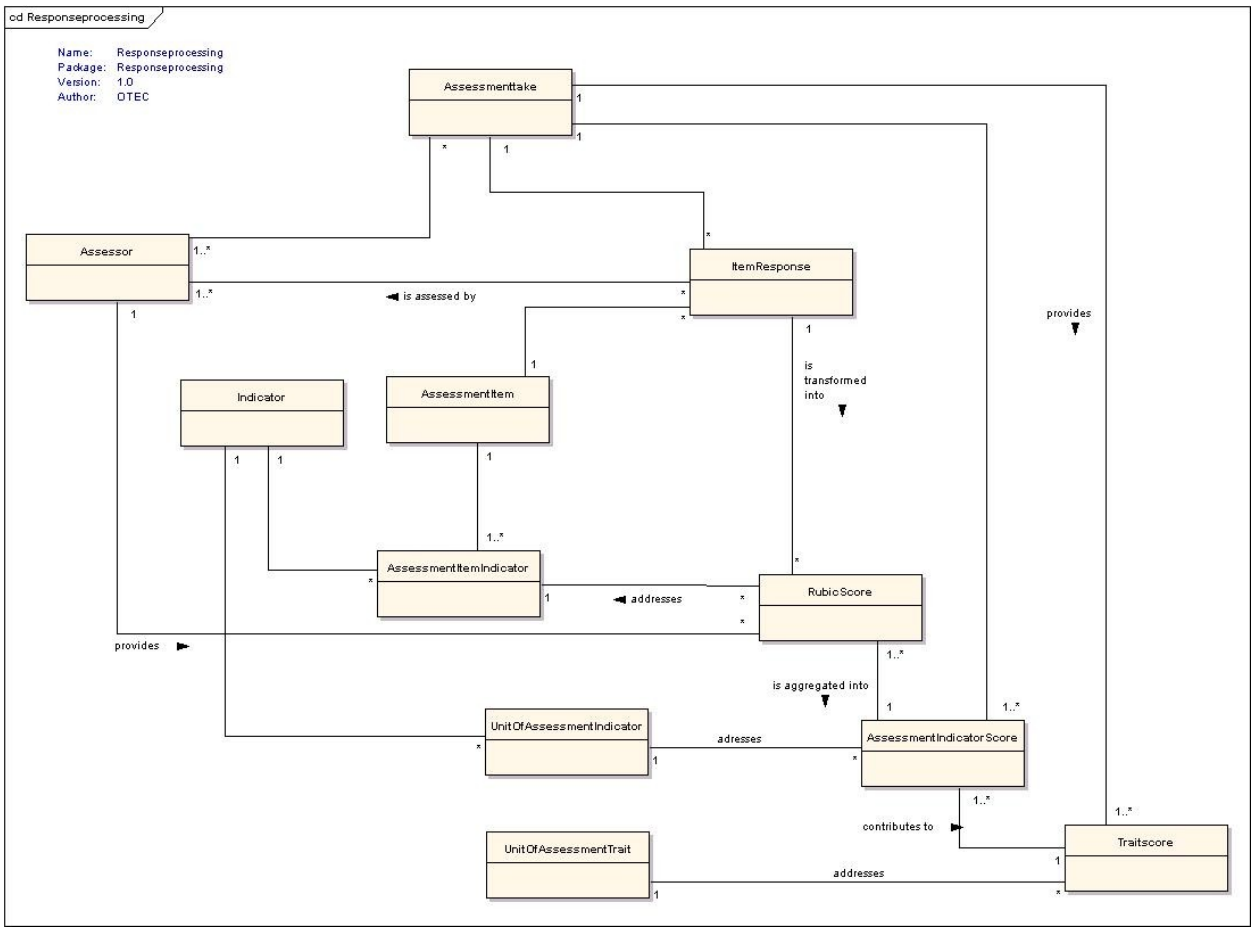

Figure 7. Decision making

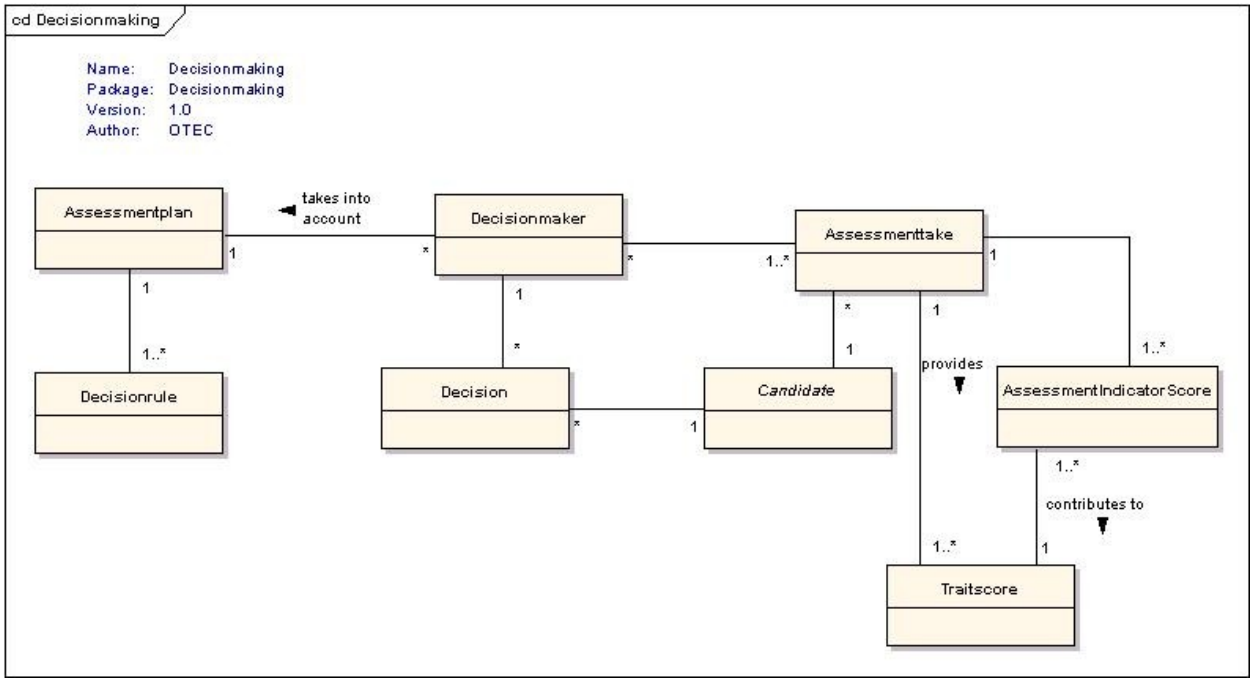

Figure 8. Assessment plan, assessment scenario and unit of assessment modeled in more depth.