A Framework for Designing and Developing Multimedia-Based Performance Assessment in Vocational Education

39th IAEA Annual Conference
Tel Aviv, Israel, October 20 – 25, 2013

Sebastiaan de Klerk [a,b,*] (s.dklerk@ecabo.nl),
Bernard P. Veldkamp [b] (b.p.veldkamp@utwente.nl),
Theo J.H.M. Eggen [b,c] (theo.eggen@cito.nl)
[a] Centre of Expertise ECABO, [b] University of Twente, [c] Cito

Abstract

Professional assessment development takes place according to a developmental framework (e.g. Downing, 2006; Mislevy, Steinberg, & Almond, 1999) due to the complex nature of assessment development. The development of assessments is an iterative and careful process of trial-and-error. A developmental framework is lacking during the earliest development of new assessment methods. De Klerk, Eggen, and Veldkamp (*in press*) discuss the emergence of a new assessment method in Dutch vocational education - multimedia-based performance assessment (MBPA). An MBPA is a CBT used for assessing specific traits, attributes or competencies of students. De Klerk et al. also remarked that it is important to have a framework for the design and development of MBPA. Therefore, in the present paper, a framework consisting of two general stages, which in turn consist of a total of thirteen steps for the design and development of MBPA will be presented and validated. The framework is constructed on basis of a literature synthesis about assessment development from several subfields of educational assessment and consultation of assessment experts. The presented framework is a first step towards empirical investigation into the use of MBPA.

Keywords: assessment in vocational education and training, multimedia-based performance assessment, assessment development

Author Note

This research was supported by Centre of Expertise ECABO.

*Corresponding author.
Sebastiaan de Klerk,
Department of Vocational Examination, Centre of Expertise ECABO, Amersfoort, The Netherlands.
Tel: +31337501005
Email: s.dklerk@ecabo.nl

A Framework for Designing and Developing Multimedia-Based Performance Assessment in Vocational Education

Performance-based assessment (PBA) is the most prevalent assessment method in vocational education and training (VET) (Baartman, 2008). Performance-based assessments may take place during work placement (i.e. internship of a student) in the authentic work setting, or in simulated form in a representation of the authentic work setting. The skills or competencies that students demonstrate during PBA are graded by one or more raters and this usually results in a categorization of competency mastery (e.g. insufficient/sufficient). The rationale behind PBA is that it offers the possibility to have students perform real tasks in an authentic work environment where competencies can be observed and evaluated which cannot be measured using more traditional measures (e.g. paper and pencil tests) (Linn, Baker, & Dunbar, 1991; Gulikers, Bastiaens, & Kirschner, 2004).

However, there is a strong repelling force between the features that characterize PBA - authenticity, complexity of tasks, raters, etc. – and quality criteria that are imposed on assessments (Linn & Baker, 1996). Assessments are required to be standardized, representative of a domain, reliable, and above all to produce valid scores and inferences (Kane, 1990; Messick, 1995). An assessment taking place during work placement can often not adhere to the criteria mentioned above because work environments are by definition not standardized and do not always provide representative tasks of the whole job. For example, students are not allowed to do all tasks related to their job because some tasks have a high risk to damage the company where they do their work placement. De Klerk, Eggen, and Veldkamp (*in press*) have reported in great detail on the measurement issues related to PBA. Additionally, researchers, guided by the digital revolution and the growing influence of technology on educational measurement, also make a case for a multimedia-based equivalent of PBA (multimedia-based performance assessment or MBPA) that might offer a solution to the measurement issues related to PBA. MBPA is a highly contextualized computer-based test (CBT) that may incorporate sound, pictures, video, animation, interactivity, serious gaming elements, and that is administered via a technological application (Mayrath, Clarke-Midura, Robinson, 2012).

Because technology improves exponentially and becomes more cost-efficient there may be a bright future ahead for MBPA. At the least, MBPA appears to be more efficient than the expensive and logistically challenging PBA. But can MBPA be, psychometrically speaking, as effective as PBA, or even more effective than PBA? The answer to this question is not within the scope of this article. But we do provide the first step towards empirical investigation into the use of MBPA in vocational education by presenting a validated framework for designing and developing MBPA.

Assessment design and development

Assessment development is the cornerstone of assessment quality. This means that sound and coherent assessments originate in a structured and well-defined approach to the development of the assessment. This will also ensure that sufficient evidence for the validation of the future assessment scores is collected during development (Downing, 2006). Weak assessments are often the result of unstructured and guideless or ‘out of the blue’ endeavors to assessment development. Assessment design and development is a time-consuming, intensive, and laborious process of trial and error. Multiple specialists from different fields are often working collectively to address different parts of assessment development (Mislevy, Steinberg, & Almond, 1999). For the design and development of MBPA, for example, a team of educational measurement experts, psychometricians, educational experts, subject matter experts, multimedia specialists, and programmers is needed. Carrying out such complex and

highly interactive processes is very difficult without guidance by a framework. Above that, a framework enables designers and developers to collect evidence for future validation of the assessment.

We stress the importance of using a framework for assessment design and development because of the complex nature of the process. A framework is inevitable, especially if it concerns a relatively new type of assessment, like MBPA. Next, we will present and validate a framework for designing and developing a multimedia-based performance assessment in vocational education and training.

Method

The framework for the design and development of MBPA was constructed on basis of a profound analysis and synthesis of literature on assessment design and development, and assessment experts' input. The literature was selected on Web of Science, Scopus and Google Scholar by searching on relevant terms (e.g. "assessment design", "assessment development", "test development", etc.). Experts were selected on basis of their experience with educational assessment development (10 years or more).

Validation

The framework was validated on assessment development literature, frameworks and guidelines by demonstrating that all steps in the framework are related to and designed upon an integration of previous literature from multiple fields (e.g. educational measurement and technology). After presentation of the framework we will discuss every step and ground them in a wide variety of literature, which together provides the validation of the framework. However, we have specifically linked the steps of the framework to Downing's (2006) twelve steps for effective test development, The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2004) – from here on referred to as the Standards -, and the evidence-centered design (ECD) framework of Mislevy, Almond, and Steinberg (1999). The validation strategy is interwoven with the discussion of the different steps of the framework.

Results

The framework is composed of two general stages, *analysis and design* and *development and administration*, and both stages reflect different processes. The analysis and design stage is mainly guided by assessment experts and subject matter experts and is for the most part executed mentally and on paper. The development and administration stage, on the other hand, is mainly guided by multimedia and ICT experts and is for the most part executed practically in an ICT environment. Though the framework physically reflects a linear type process, in reality the steps and stages may be executed parallel and they exert mutual influence.

Analysis and design

The first stage constitutes seven steps and results in a blueprint for the development and administration stage. The general rationale behind this stage is to design assessment tasks that are grounded in theory, measurable, and elicit those behaviors or traits in students that reflect the construct (i.e. competencies, skills, knowledge, etc.) to be measured. The steps of the first stage and their interrelations are represented in the upper box of Figure 1.

Insert Figure 1 about here

The first step is *(1) determining the purpose(s) and construct(s) of the assessment*. In this step, the purpose of the assessment, the construct under measurement and the rationale behind the assessment is determined. Also, during the first step an extensive overall plan for systematic guidance of the developmental process should be made (Step 1: Downing, 2006). The Standards emphasize the interpretation of assessment scores that strongly relate to the purpose of the assessment. In the case of MBPA in vocational education the purpose is usually certification (assessment of learning). Other purposes can be outplacement of a course, or job selection (RCEC, 2011; Baker, O'Neil, & Linn, 1993; Drasgow & Olson-Buchanan, 1999; Schmeiser & Welch, 2006). Thus, it should be clear what the purpose of the assessment is and what interpretations have to follow from produced scores (see Standard 1.1, 1.2, 3.2, and 14.1). Mislevy et al. (1999) refer to this step as one of the key ideas in educational measurement: "identifying the aspects of skill and knowledge about which inferences are desired".

The second step is *(2) determining the attribute(s) of the construct under measurement*. Some constructs are composed of several attributes, for example, in vocational education it is not uncommon to assess competencies (Baartman, 2006). Competencies are usually composed of knowledge, skills, and attitude (Klieme, Hartig, & Rauch, 2008). Sometimes students have to demonstrate that they have mastered one of the attributes, but sometimes they also have to demonstrate the combination of attributes in one setting (commonly referred to as 'competency'). This usually calls for a performance-based assessment setting (Linn, Baker, & Dunbar, 1991; Baartman, 2006). For the development of the assessment then, it is very important to define which attributes of the construct are part of the assessment (and therefore operationalized) and which are not. For example, if the construct is writing, the attribute can be knowledge on writing or style (latent), however it can also be the writing skill of students or the use of style in a writing assignment (manifest). Step 2 of Downing's (2006) twelve steps for effective test development stresses the importance of carefully delineated constructs. Thus, it is important to consider which attribute of the construct particularly is under measurement, and the appropriateness of the content of the assessment for the particular attribute under measurement should be justified (see Standard 1.6). The second step of the framework can again be related to the key idea explicated by Mislevy et al. (1999): "identifying the aspects of skill and knowledge about which inferences are desired".

The third step is *(3) analyzing the construct under assessment*. After the first two steps it has become clear what the purpose of the assessment is, what the construct under measurement is, and which attributes of the construct are going to be part of the assessment. Now, detailed information about the construct should be collected. The information about the construct is collected from a content domain. In education, generally, the content domain is everything that can possibly be part of the assessment (also referred to as the universe of tasks) and in which the construct is grounded. The content domain should be defined as explicitly and thoroughly as possible (see Standard 14.9). Qualifications in VET are constructed on basis of competency-based vocation profiles, which are the result of a profound analysis of a vocation conducted by educational institutions and the labor market and reflect what an experienced employee knows and does. Based on the competency-based vocation profile, the qualification profile describes in great detail what an entry employee should know, and should be able to perform if certified. The information in these profiles can be used to further start fencing off the content domain of the construct.

Content outside of the content domain cannot be part of the assessment. Within the domain there is a universe of tasks that could be designed and incorporated into the assessment (Mislevy et al., 1999; Mislevy, 2011). Through profound analysis of actual job

behavior, it is possible to design tasks that are part of the assessment (Weekley, Ployhart, & Holtz, 2006). By carefully observing qualified job incumbents, typical job behaviors that are the pillars of the vocation can be isolated. Generally, this stage is characterized by collaboration between multiple specialists: subject matter experts (SMEs), and assessment experts (Downing, 2006; Weekley et al., 2006).

Another part of this stage is the cognitive analysis of the construct, which explain the cognitive steps students take in completing actual job behaviors, and those should be strongly aligned with the tasks that are part of the assessment (Mislevy, et al., 1999). If alignment is missing, sound statements about the generalization from an assessment setting to a real-world setting can never be made. Think aloud methods are generally used to analyze the cognitive strategies individuals follow while performing specific tasks (Van Someren, Barnard, & Sandberg, 1994; Messick, 1995).

Finally, the construct analysis delineated above, using multiple perspectives leads to a comprehensive and exhaustive content domain. The task content is going to be based upon and selected from this domain. The third step in the framework is also related to another key idea of educational measurement as discussed by Mislevy et al. (1999): "identifying the relationships between targeted knowledge and behaviors in situations that call for their use".

The fourth step is *(4) designing assessment task(s) and operationalization of student behavior*. This step is defined by an exchange relationship with the third stage, which means that the task content is defined by the construct analysis from the previous step, and that task design may uncover possible shortcomings in construct analysis. The tasks should elicit behavior in students which can be interpretable to make claims about student skills, competencies or knowledge. Before the actual task can be designed the content of the task should be selected, in cooperation with SMEs, from the content domain. This step is comprised of four different parts: *task attribute, task context, student behavior,* and *response type*.

The first part of task design is determining which type(s) of attribute(s) is/are going to be part of to be designed tasks. An entire multimedia-based performance assessment is a construction of a multitude of tasks, and all tasks tap on specific *task attributes.* Task attributes are for example, knowledge, attitude, skill, cognition, competency, or behavior (Frederiksen & Collins, 1989; Mislevy et al., 1999). The task attributes can also differ in their level of complexity. Tasks in assessment in vocational education are usually composed of multiple attributes (Baartman, 2006; Klieme, Hartig, & Rauch, 2008).

The second part of task design is *task context*. Factors that are present in, and influence, a real-world context should also be part of the context created in the tasks to enhance authenticity (Gulikers, Bastiaens, & Kirschner, 2004). Logically, this starts with designing an environment that resembles the real-world environment. In MBPA, this is a virtual environment. A central concept in task context is the authenticity of the task. Gulikers et al. (2004) distinguish five dimensions of authenticity: the assessment task, the physical context, the social context, the assessment result or form, and the assessment criteria. It is thus important to stress that task context incorporates more than just the physical context of the task. Furthermore, in this part of task design the general 'flow' of the MBPA is designed. This means that context scenarios are written that clarify how students move through MBPA from task to task.

The third part of task design is defining *student behavior*, which is the behavior that students have to demonstrate for performance of the assessment tasks. The behavior that the task evokes in students provides evidence about the targeted construct (Mislevy et al., 1999). Student behavior should therefore be defined in the smallest components possible because it

also determines the responses in the MBPA which are ultimately incorporated in the score model.

The fourth and final part of task design is the *response type*. In MBPA there is a whole range of response types that can be logged. For example, speed, clicking behavior, navigational behavior through the virtual environment, typing, eye-tracking, and responses on innovative and traditional items types (for an overview see Mayrath, Clarke-Midura, Robinson, & Schraw, 2012; de Klerk, 2012). Downing (2006) argues that the creation of effective assessment tasks with the right context and the appropriate cognitive level is one of the most difficult tasks in assessment development (see Step 4). The type of item and the response formats should be selected for the purposes of the assessment (see step 1), the domain to be measured (see step 2 and 3), and the intended test takers (see also Standard 3.6). The fourth step in the framework is also related to another key idea of educational measurement as discussed by Mislevy et al. (1999): "identifying features of situations that can evoke behavior that provides evidence about the targeted knowledge".

After task design the fifth step in the framework is *(5) constructing the evidence model.* This step is schematically located between stages three and four, and relates to the exchange relationship between the former two steps. In the evidence model a comprehensive and extensive argument is presented that vindicates and explains how the constructed tasks, including attributes, context, student behavior, responses and ultimately scoring result in psychometrically sound statements about students. In other words, evidence should prove that we can actually say something about students in real life (i.e. the criterion) based on performance on the tasks in the assessment (i.e. the predictor) (see Standard 14.12). Often, the strength of the relationship can be determined after the administration of the assessment has yielded results. However, it important to systematically analyze to what extent it seems plausible to expect valid results from the performance on designed assessment tasks as to statements about the construct under measurement. Downing (2006) remarks that systematic, thorough, and detailed documentation for validity arguments should be collected during development. Mislevy et al. (1999) discern two parts within the evidence model; the statistical model and the evidence rules. The evidence model in our framework refers to and builds upon the evidence rules specified in the ECD framework because the assessment developer should provide evidence on the relationship between student behavior in assessment tasks and the construct.

The sixth step is *(6) constructing the score model*. Student responses in the MBPA have to be scored to be able to construct a measurement model that will lead us from collected observed variables to claims about the construct. All student responses collected during administration that contribute to a score are part of the score model. Scoring can be quantitative as well as qualitative, and scoring rubrics assist in attaching weights to the scores and combining scores into an overall score or result (Shepherd & Mullane, 2011). According to Downing (2006) perfectly accurate scoring results in valid meanings, as they are anticipated by the assessment developer. Furthermore, scoring criteria and the procedures for scoring should be presented in sufficient detail and clarity for making the scoring as accurate as possible (see Standard 3.22). Mislevy et al. (1999) classify scoring mainly under the tasks model in their ECD framework but it also relates to their student model, and evidence model because of the link between performance and evaluation.

The seventh step is *(7) constructing the measurement model*. Mislevy and Riconscente (2006) define the measurement model as a mechanism to define and quantify to what extent students' responses, as combined in the score model, provide information for statements we want to make about students. The administration of an assessment yields a certain amount of data, depending on the amount of responses and the type of responses students have to

produce. Scoring ultimately results in claims on targeted knowledge or competency of students. By applying a measurement model on collected observed variables we can infer from data to a scale of (a) latent variable(s). Psychometric models, for example Item Response Theory (IRT), are part of the measurement model (see Standard 3.9). Mislevy et al. (1999) discuss the statistical model, which largely corresponds with our measurement model, and they define it as part of the evidence model. Although the measurement model represents the relationship between students' degree of construct mastery (e.g. a latent characteristic) reflected in their performance and scores produced on basis of the performance, we specify the construction of the measurement model as a final step of the first stage because that seems most realistic for designing MBPAs which may constitute a multitude of different task types.

The first stage concludes with a blueprint of the assessment and a collection of validity evidence. The blueprint is the point of departure for the second stage.

Development and administration

The second stage of the framework is more practical than the first stage. Development of what has been decided and reported in the first stage will be incorporated in an MBPA. The second stage consists of six steps and results in a functioning MBPA. However, it may be that shortcomings of the first stage are recognized in the second stage and in that case it is necessary to return to the first stage before completing the second stage. The steps of the second stage and their interrelations are represented in the lower box of Figure 1.

The eighth step, and the first step of the second stage, is *(8) developing ICT system and interface*. The development of an ICT infrastructure only holds if one is not present yet. The infrastructure should be able to incorporate and present multimedia and innovative items. We emphasize that we are not necessarily discussing immersive virtual environments here, as in the case of those common in (serious) games or simulations (e.g. a flight simulator for the training of pilots). However, we do mean a virtual interface that, for example, can incorporate movies, animations, and avatars in a flow like architecture. This means that the ICT system and interface should be able to present (a) scenario(s) that the student can 'walk through' and complete in the assessment. The exact structure and composition of the ICT system and interface is beyond the scope of this paper.

The ninth step is *(9) developing tasks and multimedia and implementing in ICT system*. Multimedia experts create the multimedia which can be programmed in the tasks, based on task design in the first stage. Now, the developers start filming, and creating animations, avatars, and innovative item types. This is a laborious and iterative process of creating, evaluating, and adjusting, but finally leads to the first form of the assessment. This step also includes the programming of the assessment and attaching responses and accompanying scores to tasks. Another part of this step is programming possible feedback and scoring rules.

The tenth step is *(10) implementing in network*. The assessment can now be implemented in a network of computers, or they can be locally installed or uploaded on single computers. Extensive guidelines exist on the development use of computer-based assessment (e.g. ATP, 2002; ITC, 2005), and the assessment developer should use a set of guidelines during the execution of the previous three steps of assessment development. The MBPA should be tested to make sure it is free of bugs and that it functions smoothly.

The eleventh step is *(11) pretesting the assessment*. The assessment should be pretested before administration for high-stakes use or for use of large groups of students. Here, in contrast to the previous step where the ICT functionality was tested, pretesting refers to determining the psychometric functionality of the assessment. The pretesting should take place within a relatively small sample of students from the target population. However, the

sample should be large enough to make statements about the functioning of the tasks in the assessment.

The twelfth step is *(12) deciding upon the fitness for purpose*. If the assessment functions psychometrically correct, then the next step is to start using it for its actual purpose in practice. If the assessment does not function psychometrically correct, the eleventh step loops back to either the first step of the first stage or the first step of the second stage. This exemplifies the relationship between the first and the second stage, and the iterative character of assessment development.

The thirteenth step is *(13) administrating the assessment*. The organizational institution can start to administer the assessment to large groups of students for its ultimate purpose if the pretest results in the desired outcomes. This does not mean that the assessment is finished and can be endlessly used. The quality of the assessment, and its fit for purpose should constantly be monitored by educational and assessment experts (see also Standard 3.25).

The complete framework for the design and development of MBPA, including interrelations between stages one and two and all steps, is depicted in Figure 1.

Discussion and conclusion

The point of departure of this article was to provide a first step towards empirical investigation into the use of MBPA in VET. We provided that first step by presenting and validating our framework for the design and development of MBPA. In this article we have described the development of our framework for the design and development of MBPA and we have validated the framework on a synthesis of literature. In the future the framework will be further validated on basis of experts' appraisal of the framework that will be gathered in semi-structured interviews. Also, we are planning to develop an MBPA on basis of the presented framework.

We have shown that the framework is grounded in theory and literature by relating every step to both the twelve steps of test development by Downing (2006), Mislevy's (1999) evidence-centered design framework for the design of assessments, and the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2004). However, the present research is limited because actual value of the framework is only shown by the development of an MBPA. Future research should therefore focus on the design and development of MBPA according to the framework, on the psychometric functioning of the designed and developed MBPA, and on an empirical and psychometric comparison of PBA and MBPA.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2004). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Association of Test Publishers (ATP: 2002). *Guidelines for cmoputer-based testing*: ATP.

Baartman, L.K.J., Bastiaens, T.J., Kirschner, P.A., & Van der Vleuten, C.P.M. (2006). The wheel of competency assessment: Presenting quality criteria for Competency Assessment Programmes. *Studies in Educational Evaluation, 32*, 153-170.

Baartman, L.K.J. (2008). *Assessing the assessment: Development and use of quality criteria for competence assessment programmes* (Doctoral dissertation, Utrecht University, The Netherlands). Retrieved from http://hdl.handle.net/1820/1555.

Baker, E.L., O'Neil, H.F., & Linn, R.L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*(12), 1210-1218.

De Klerk, S. (2012). An overview of innovative computer-based testing. In T.J.H.M. Eggen & B.P. Veldkamp (Eds.), *Psychometrics in practice at rcec* (pp. 137-150). Enschede: RCEC.

De Klerk, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2013). *A Blending of Computer-Based*

Assessment and Performance-Based Assessment: Multimedia-Based Performance Assessment (MBPA). The Introduction of a New Method of Assessment in Dutch Vocational Education and Training (VET). Unpublished manuscript.

Downing, S.M. (2006). Twelve steps for effective test development. In S.M. Downing and T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Drasgow, F., & Olson-Buchanan, J. (1999). *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327-358.

Frederiksen, J.R, & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*, 27-32.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52(3)*, 67-86.

International Test Commission (ITC: 2005). *International guidelines on computer-based and internet delivered testing*: ITC.

Kane, M.T. (1990). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, Klieme, E., & Leutner, D. (Ed.), *Assessment of competencies in educational contexts* (pp. 3-22). Göttingen: Hogrefe.

Linn, R.L., & Baker, E.L. (1996). "Can performance-based student assessments be psychometrically sound?" In J.B. Baron & D.P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth Yearbook of the National Society for the Study of Education, Part 1* (pp. 84-103). Chicago: University of Chicago Press.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology, 14*, 1-38.

Mayrath, M.C., Clarke-Midura, J., & Robinson, D.H. (2012). Introduction to technology-based assessments for 21st century skills. In M.C. Mayrath, J. Clarke-Midura, D.H. Robinson & G. Schraw (Eds.), *Technology-based assessments for 21st century skills* (pp. 1-11). Charlotte, NC: Information Age.

Mayrath, M.C., Clarke-Midura, J., Robinson, D.H., & Schraw, G. (Eds.) (2012). *Technology-based assessment for 21st century skills*. Charlotte, NC: Information Age.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Mislevy, R.J., & Riconscente, M.M. (2006). Evidence-centered assessment design. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (1999). *On the roles of task model variables in assessment design*. (CSE Technical Report 500). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (2011). *Evidence-centered design for simulation-based assessment.* (CRESST Report 800). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Research Center for Examinations and Certification. (2011). *Beoordelingssysteem voor de kwaliteit van toetsen en examens* [Evaluation model for the quality of tests and examinations]. Unpublished manuscript.

Schmeiser, C.B., & Welch, C.J. (2006). Test Development. In R.L. Brennan (Ed.), *Educational Measurement* (pp. 307-353). Westport, CT: Praeger.

Shepherd, C.M., & Mullane, A.M. (2008). Rubrics: The key to fairness in performance-based assessments. *Journal of College Teaching & Learning, 5*(9), 27-32.

Van Someren, M.W., Barnard, Y.F., & Sandberg, J.A.C. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.

Weekley, J.A., Ployhart, R.E., & Holtz, B.C. (2006). On the development of situational judgment tests: issues in item development, scaling, and scoring. In J.A. Weekley & R.E. Ployhart (Eds.) *Situational Judgment Tests: Theory, Measurement, and Application* (pp. 157-182). Mahwah, NJ: Lawrence Erlbaum Associates.

Figure 1. Flow schematic representation of the complete framework