Education Tech Research Dev (2018) 66:147–171
https://doi.org/10.1007/s11423-017-9559-5

DEVELOPMENT ARTICLE

# A framework for designing and developing multimedia-based performance assessment in vocational education

**Sebastiaan de Klerk**[1] · **Bernard P. Veldkamp**[2] · **Theo J. H. M. Eggen**[3]

Published online: 20 December 2017


**Abstract** The development of any assessment should be an iterative and careful process. Ideally, this process is guided by a well-defined framework (see for example Downing in: Downing and Haladyna (eds) Handbook of test development, Lawrence Erlbaum Associates, Mahwah, 2006; Mislevy et al. in On the roles of task model variables in assessment design (CSE Technical Report 500), Educational Testing Service, Princeton, 1999; AERA et al. in Standards for educational and psychological testing, AERA, Washington, DC, 2004), but such a framework is not always available when the instrument to be developed is new or innovative. Frameworks for the development of traditional computer-based tests have been published and experimented with since the late 1990s, by which time CBT had already existed for more than a decade. In an earlier empirical pilot study, we described a new type of assessment for Dutch vocational education, called multimedia-based performance assessment (MBPA) (self-revealing reference 2014). This CBT uses multiple media formats and interactive tasks to measure skills that are currently measured by performance-based assessment. In conducting that pilot study, deficits in the existing literature made it difficult to ground all developmental steps in sound scientific theory. To remedy those deficits, this article presents and validates a framework for the design and development of MBPA, combining a search of the relevant literature from several subfields of educational assessment and consultation with assessment experts. The framework unites assessment development and multimedia development theory, focus solely on vocational education,

✉ Sebastiaan de Klerk
s.dklerk@explain.nl

Bernard P. Veldkamp
b.p.veldkamp@utwente.nl

Theo J. H. M. Eggen
theo.eggen@cito.nl

1 eX:plain, P.O. Box 1230, 3800 BE Amersfoort, The Netherlands

2 Department of Research Methodology, Measurement and Data Analysis, Faculty of Behavioral Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

3 Cito, P.O. Box 1034, 6801 MG Arnhem, The Netherlands

and answers the call for a framework from the scientific community. The first step in validating the prototype framework involved five semi-structured interviews with Dutch assessment and multimedia experts to produce a final version of the framework. Second, the pilot MBPA was reconstructed in accordance with this finalized framework, resulting in an improved MBPA and demonstrating that the proposed framework is a useful and applicable tool for the design and development of MBPA in vocational education.

**Keywords** Assessment design · Assessment development · Multimedia-based performance assessment · Computer-based assessment · Vocational education

## Introduction

Multimedia-based performance assessment (MBPA) is an innovative type of assessment that blends computer-based testing (CBT) and performance-based assessment (PBA) (self-revealing reference 2014). We have introduced the term multimedia-based performance assessment for two reasons. First, MBPA is enabled by technological and digital innovations employing *multimedia*. In contrast to more traditional forms of CBT, multimedia is used in MBPA to simulate a real-world environment in which tasks are administered as the student navigates through that environment. Using multimedia (e.g., animation, video, virtual reality), students are "immersed" in a virtual environment in which they must complete tasks or achieve objectives. Second, *performance-based assessments* require students to perform an activity or construct an original response, and is used to assess higher-order cognitive skills or, as is usually the case in vocational education, procedural and manual skills (Lane and Stone 2006; Herman et al. 1992). Examples of PBAs are a practical driving test and writing an essay. MBPA is used to measure student skills that are currently measured in a performance-based assessment, and so we have coined the term *multimedia-based performance assessment*. In this virtual environment, students may be provided with tools to help them complete their tasks, and because they have, (to varying degrees) more freedom to operate, there is usually more interaction between the student and the computer in this environment than in other types of CBT.

MBPA offers a different approach to the measurement of certain constructs currently measured with difficulty by performance-based assessment. For example, the skills or competencies that students demonstrate during a PBA are rated by one or more raters, usually resulting in a categorization of competency mastery as, for instance, insufficient/sufficient. However, it has often been demonstrated that PBAs score low on generalizability, reliability, and standardization, and that rater effects can influence students' scores (Kane 1990; Linn et al. 1991; Ruiz-Primo et al. 1993; Shavelson et al. 1993; Yen 1993; Messick 1995; Dekker and Sanders 2008). MBPA might be used as an alternative and improved approach to the measurement of vocational skills.

### The current status of MBPA design and development

In a previous study (self-revealing reference 2014), we developed and tested a pilot version of an MBPA in the context of Dutch vocational education to measure the vocational skills of *confined space guards* (CSG). A CSG supervises operations that are carried out in a confined space—for example, in a tank at a petrochemical plant. In the Netherlands, every CSG must complete a training program and pass a PBA in a simulated work environment to demonstrate that they can perform all the CSG's tasks. The performance is rated by a

rater on several criteria. We have tried to capture all tasks of the PBA in the MBPA, using multimedia and interactive tasks. In designing and developing this innovative MBPA, we identified an unmet need for a well-defined framework to guide this complex and multi-faceted process. Although there are guidelines and frameworks available for the design and development of assessments in general, very few actually focus on innovative assessments, especially in the domain of vocational education.

Additionally, research and development on the measurement of vocational skills through CBT is rising (e.g., Iseli et al. 2010; Rupp et al. 2012a, b; Levy 2013), which indicates that a lot of researchers and practitioners are currently using an unfit developmental framework, no framework at all, or a self-developed/intuitive framework. Rupp et al. (2012a, b, p. 5) for example indicate that: "the advances that such environments provide require a similar match in advancements in assessment design". Shute et al. (2010, p. 305) note that: "We maintain that not only is it important (…) to identify particular methods for designing and developing assessments that are valid and reliable and can help us meet the educational challenges confronting us today". Quellmalz et al. (2012, p. 367) "drew key principles for designing the assessment tasks" during the development of their SimScientist assessment program. And Levy (2013, p. 191) subtly remarks that: "reflecting on the research conducted and lessons learned in the last 100 years with traditional assessment formats (…) begs the question whether a similar level wisdom is needed about how to best design simulations of assessment". Finally, still other research (e.g., Quellmalz et al. 2013; Halverson et al. 2012; Wainess et al. 2011; Vendlinski et al. 2010; Shute et al. 2009) presents parts of a methodology or principles for designing and developing MBPA and related types of assessment, but these were never integrated into a full design and development framework.

Nevertheless, literature does provide several fully integrated and comprehensive design frameworks that can also be used for the development of innovative assessments. For example, Downing's twelve steps for effective test development (2006), the Standards for Educational and Psychological Testing (AERA et al. 2004) and the evidence-centered design (ECD) framework (Mislevy et al. 1999). However, although they can be used, all three frameworks are not entirely suitable for designing and developing MBPAs in vocational education. The first two of these focus strongly on traditional testing formats and can be used as a structured, step-by-step approach to assessment development. Mislevy et al.'s framework is also applicable for simulation-based assessments but offers a more abstract approach to assessment development than the framework presented in the current article. For example, the central element in Mislevy's model is the conceptual assessment framework (CAF). The CAF consists of three models: the student model (what is going to be assessed?), the task model (what tasks are going to be used for assessment?) and the evidence model (how can we (statistically) connect the student model with the task model?). These models provide a conceptual framework to structure the reasoning of the assessment from construct to task, and back to construct. It does not provide a sort of step-by-step approach for design and development of a (computer-based) assessment.

## The need for a framework for MBPA design and development

Considering the above delineated current status of MBPA design and development, we think that there is a need for a comprehensive framework for the design and development of MBPA, particularly in vocational education. First, the framework presented in this article can integrate the isolated efforts of researchers discussing a specific (sub)process of the design and development of simulation-based assessment, and MBPA in particular.

Second, the framework can fill the knowledge gap left by the existing frameworks in several ways.

First, we present a framework that strongly and solely focuses on the design and development of interactive and virtual assessments in a vocational education setting where practice-oriented constructs are subject of measurement. The framework therefore emphasizes analyses of qualification profiles, job descriptions, final attainment objectives, and the curricula as a whole, which other frameworks do not. Second, the frameworks shortly discussed above all exist for more than a decade. With a digital revolution progressing at an ever-increasing pace, now may be a good time to unify MBPA design and development practices in a framework. Third, although the content of our proposed and to be presented framework is for some part based on other literature, we present a wholly different structure of reasoning, and we try to incorporate the best of multiple 'worlds' in our framework. For example, in contrast to other frameworks, a central feature of our framework is the interaction between multiple processes, design and development stages, and multi-disciplinary teams. Fourth, we actually use two strategies, semi-structured interviews of experts and building an MBPA according to the design and development principles described in the framework, to *validate* our framework. This has not been done for other frameworks.

The specific benefits of our new framework, as compared to the literature, theories, and frameworks discussed above thus are: (1) a united vision and methodological procedure for the design and development of a multimedia-based equivalent of performance-based assessment, (2) the unique emphasis on vocational education, an in scientific research underexposed type of education that strongly diverges from other types of education through its emphasis on learning by doing and performance-based assessment, and (3) an answer to the call for a methodology for designing and developing innovative, simulation-based assessments for the measurement of practical skills as made by the scientific community.

## A framework for MBPA design and development

Above, we have argued that there is a need for a framework for the design and development of MBPA in vocational education. If so, why would the proposed framework, as presented in this article, be the one that is needed by both researchers and practitioners? First, our framework fills a void left by the other frameworks. Where other frameworks are either centered around rather abstract reasoning processes (e.g., ECD) or, to the contrary, are too linear in nature (e.g., Downing's 12 steps for effective test development), our framework finds the right balance in presenting a step-by-step approach within the interactive, adaptive, and iterative processes of design and development. Secondly, the framework to be presented is the first one that explicitly focuses on a vocational education setting. This does not mean that the framework cannot be used in other (educational) settings, but it does mean that specific features that are unique for assessment design in vocational education have been incorporated in the framework. In that regard, our framework does also differ from the other frameworks mentioned. Thirdly, the framework has been developed in collaboration with practicing assessment experts and leading assessment researchers. Bringing together experience from both fields makes it more likely that the framework will suffice for its purpose. Fourthly, we validate the framework using two strategies. First, after a first prototype of the framework had been developed, we organized 5 semi-structured expert interviews. Improvements to the prototype were made on basis of the outcomes of these interviews. Second, we have built an MBPA using the

presented framework. Although it is beyond the scope of this article to discuss all the (empirical) details of this MBPA, we do discuss how we have used the framework for design and development of the MBPA.

To summarize, research on the design and development of MBPA is dispersed, and although elements of the developmental process have been discussed in literature, these were mostly isolated and no unified framework for MBPA development has been presented. Furthermore, the frameworks for designing assessments that do exist do not provide enough support for building an MBPA in vocational education. Finally, we have argued that the framework that will be presented in this article unites previous research and has unique characteristics that make it the first comprehensive framework for the design and development of MBPA in vocational education. The next section describes how the proposed framework was constructed.

## Method

The framework for the design and development of MBPA was constructed and validated in five consecutive steps: (1) a literature search relating to relevant aspects of assessment design and development; (2) construction of a first prototype, based on the first step and following consultation with three Dutch assessment experts; (3) validation of the first prototype on the basis of five semi-structured interviews with assessment experts other than those consulted during step 2; (4) finalization of the prototype on the basis of validation results; and (5) empirical testing of the final version by development of an MBPA. Below, we will discuss how each step was carried out and we will provide an example of how steps are linked as a chain.

### Step 1: literature search

To begin construction of the prototype, sources that included Web of Science, Scopus, and Google Scholar were searched using relevant terms (e.g., "assessment design", "assessment development", "assessment guidelines", "assessment framework", "test design", "test development", "test guidelines"). Following the literature review strategy of Petticrew and Roberts (2006), items were selected if (1) the main topic of the article or chapter related to assessment/test development and (2) the article or chapter provided a structured set of rules or guidelines (i.e., a framework) for assessment development. For example, the ECD framework (Mislevy et al. 1999) provided relevant and valuable input and their evidence model has therefore been adopted here.

### Step 2: construction of the prototype

The literature study was followed by consultation with three professional experts in the field of assessment, working respectively for the Dutch national institute for test development (which is called Cito), the University of Twente in the Netherlands, and a private Dutch assessment development company. All three had more than 10 years of experience in the design and development of assessments. During construction of the prototype, four rounds of expert consultation were organized to ensure that development remained on track and to avoid tunnel vision. In the first round, the literature from the previous step was discussed and categorized. On basis of this information, a rough sketch of the framework

was made. In the second round, two general stages were constructed and possible steps within the framework were selected (as explained below). In the third round, the prototype framework was built: steps were placed in the right stages, and connected. Team set-up for the stages was defined, and the iterative character of the framework was emphasized by feedback loops. Finally, in the fourth round the framework was graphically refined and a manuscript was written in which both stages and all steps were explained. Having first constructed the two general stages of the framework, both stages were structured as sequential steps for design and development of MBPA, and those stages and steps were then connected to explicate their interrelationships. Finally, multiple sub-steps were added to the task design and development steps; these cannot stand on their own as separate steps because they are strongly connected to their parent step, but as they define quite specific processes during task design and development, it is reasonable to add them to the framework. For example, in accordance with the experts' view, the evidence model was placed appropriately within the framework, linking it to the relevant steps.

## Step 3: validation of the prototype

### *Participants*

The prototype was validated by means of five semi-structured with experts in either assessment, multimedia design and development, or both, and the prototype was finalized on this basis. To keep construction and validation of the prototype separate processes, the participating experts were not involved in construction of the prototype. The experts were selected on the basis of their experience in innovative assessment design and development. All experts have theoretical (e.g., publications in scientific journals) and practical experience on the topic. As all experts accepted to be interviewed we consider the sample to be representative. Each expert had more than 10 years of (leadership) experience in the design and development of innovative CBT. Three of the participants had a doctoral degree and two had a master's degree in the areas of assessment or multimedia development. Their backgrounds were very diverse. One was primarily researching the incorporation of serious gaming elements in assessment for the purpose of personnel selection. Another was responsible for the implementation of technology-based assessment at Cito, while a third was involved in the development of multimedia for CBT. The two remaining experts were primarily involved in research on the innovative use of CBT in higher education.

### *Materials and procedure*

The five identified experts were appoached via e-mail and we asked them whether they would be willing to read and study a manuscript describing the proposed prototype, and then be systematically interviewed about it; all five replied positively. A semi-structured interview schedule was then constructed, based on a study in which an evaluation system for performance assessments had been validated on the basis of expert interviews (see Wools et al. 2011). Although semi-structured interviews only provide qualitative evidence, they can be a valuable source of data for validating processes or products (Barriball and While 1994; Wools et al. 2011). For example, in (educational) design-based research, in which the design and development of educational processes and products are subject of research, systematic expert interviews are often used to identify the strength of a process or product (McKenney and Reeves 2012; McKenney and Van den Akker 2005).

Specifically, the interviews revolved around two concepts: the content and usability of the prototype. Content was characterized in terms of four categories: general quality, completeness, correctness, and coherence. Usability was characterized in terms of two categories: general usability and fitness for purpose. During the interviews, the experts were systematically questioned about all elements in the framework in respect of those concepts and categories. All interviews were conducted face-to-face at the experts' work location; the interviews were recorded, making it possible to carefully re-listen to and interpret the experts' statements without losing the context in which those statements were made. An example of a question would be: "To what extent is the content of the first step of the first stage correctly discussed in relation to the design and development of multimedia-based performance assessment?".

A verbatim transcript was made of each interview. To keep experts' statements in their proper context, cues were written in the margin of the transcript to indicate any special circumstance that led to this statement (e.g., an example, anecdote, or personal experience). Text fragments that referred specifically to the content concept, the usability concept, or to one of the underlying categories were then filtered and selected from the full transcripts. This selection of text fragments was done on an individual and independent basis by each of the authors of this article. Subsequently, the authors of this article collectively discussed what fragments were useful for the revision of the prototype and which were not. A high degree of correspondence was found between the three authors in the fragments selected, and any fragments selected by all were automatically included (i.e., 77% of the statements). Statements that were selected by one or two of the three authors were collectively discussed for possible inclusion (i.e., 18% of the statements—the remaining 5% of statements were not used for the revision of the prototype); overall, we were quickly able to reach agreement about these statements. The rationale behind this strategy was that the experts' views as expressed in these text fragments would be both meaningful and useful for the further development of the prototype into a final framework. To follow up on our example, the evidence model of course formed part of the subject matter of the five interviews; the experts were questioned on the positioning of the evidence model within the framework and whether it was usable and correctly described.

### Step 4: adjustment of the prototype and final framework

In the fourth step, statements made by the assessment experts were used to transform the prototype into a final framework. In total, 28 text fragments were extracted from the interviews that relate directly to the framework. This may seem as a rather limited number of text fragments, but there was a high degree of correspondence between the statements made by the five experts. As can be seen in the results section below, the role of the evidence model changed in the final framework as compared to the prototype, demonstrating how significantly the interview data impacted the finalization process.

### Step 5: validation of the final framework

In the fifth and final step, the final version of the framework was used in a real situation. As discussed above, a pilot MBPA had already been developed for a Dutch vocational course for CSGs, and the degree of difficulty experienced in the design and development of that pilot MBPA was one reason for building this framework (self-revealing reference 2014). In the last step of this research, the final framework was used to redevelop, that is to build a new CSG MPBA from scratch, which can be seen as an empirical step in validating our

framework; if the end product (the new MBPA developed according to the such demonstration that a justifiable evidence model could be built for this MBPA would constitute a strong argument that the evidence model was correctly positioned within the framework. There follows a discussion of the results from each of the five steps.

## Results

### Step 1: literature search

The literature search returned 14 articles or book chapters used in construction of the prototype, and each step in the prototype is grounded in a relevant area of the literature. The steps in both stages of the framework (design and development) were specifically linked to Downing's (2006) twelve steps for effective test development, The Standards for Educational and Psychological Testing (AERA et al. 2004) (from here on referred to as the Standards), and the evidence-centered design (ECD) framework of Mislevy et al. (1999).

We refer in particular to these three frameworks for a number of reasons. First, Downing's *Handbook of Test Development* (Downing 2006) provides a step-by-step approach to assessment development, which is also an aim of the proposed framework. Secondly, while the Standards are the most influential guidelines on test analysis, they are in practice often used for test development. Because these guidelines are so comprehensive and underwritten by the most influential experts in the area, it would be a missed opportunity not to use this standard work for the development of the present framework. Third, while the ECD framework supports more abstract reasoning about designing and developing assessments, it is also often used to build simulation-based assessments, which are closest in kind to the proposed multimedia-based approach.

### Step 2: construction of the prototype

Based on the literature search and input from the participating assessment experts, two processes were identified in building an MBPA: a design phase and a developmental phase. The framework therefore comprises two general stages: *analysis and design* and *development and administration*, both involving different processes. The analysis and design stage is guided mainly 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 guided mainly by multimedia experts and practitioners and is for the most part executed practically in an ICT environment. For efficiency, we will refer in the following sections to the "assessment developer" as representing the whole team engaged in the design and development of the MBPA. Decisions made at the first stage influence the second stage, and conversely, the first stage is also influenced by the second stage, when possible hiatuses in the first stage may be detected. The full prototype framework is presented in Fig. 1.

#### *Analysis and design*

Turning to a systematic discussion of all parts of the framework, the first stage involves seven steps, resulting in a detailed report for use at the development and administration stage. In this first stage, the general rationale is to design assessment tasks that are

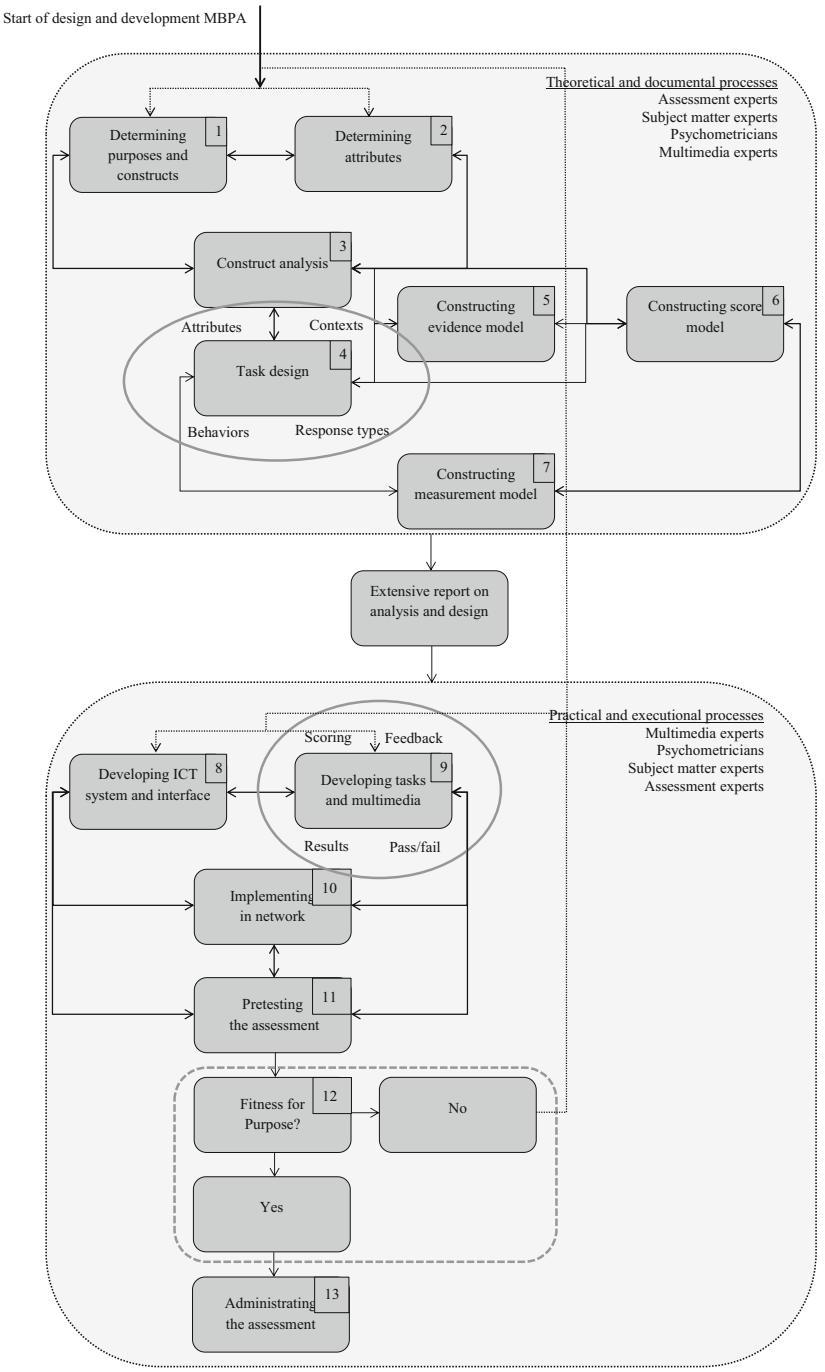


**Fig. 1** Flow schematic of the prototype framework

grounded in theory, are measurable, and elicit student behavior that reflects the construct (competencies, skills, knowledge, etc.) to be measured. These steps were identified on the basis of the literature informing each step, along with the expert inputs during the several rounds of consultation. *At the end of the discussion of each step we give a short, one sentence directional summary for practitioners who wish to develop an MBPA.*

The first step, then, is *(1): determining the purpose(s) and construct(s) of the assessment*. Here, the assessment developer elaborates a comprehensive argument concerning the purpose of the assessment—what precise construct is to be assessed and why there is a need for that assessment. During this first step, an extensive overall plan should be made for systematic guidance of the developmental process (Step 1: Downing 2006). The Standards emphasize the interpretation of assessment scores that strongly relate to the purpose of the assessment, which may, for example, include certification of individuals (e.g., a yes/no decision), course placement, or curricular reform (RCEC 2015; Baker et al. 1993; Drasgow and Olson-Buchanan 1999; Schmeiser and Welch 2006). The assessment developer should state clearly the purpose of the assessment and what interpretations should follow from the scores produced (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".

Bring together a group of subject matter experts to systematically discuss and document the purpose of the assessment and the underlying constructs to be measured.

The second step is *(2): determining the attribute(s) of the construct under measurement*. Some constructs comprise several attributes—in vocational education, it is not uncommon to refer to competencies (Baartman et al. 2006), usually composed of knowledge attributes, skill attributes, and attitude attributes (Klieme et al. 2008). Sometimes, students must demonstrate that they have mastered one of the attributes; on other occasions, they may be required to demonstrate a combination of attributes in a single setting, usually in a performance-based assessment (Linn et al. 1991; Baartman et al. 2006). For 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 might be knowledge on writing or style, but it might equally be the student's writing skill or the use of style in a writing assignment. Step 2 of Downing's (2006) twelve steps for effective test development stresses the importance of carefully delineated constructs. The assessment developer needs to consider which attribute in particular of the construct is to be measured, and the appropriateness of the assessment content for that particular attribute 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".

In the subject matter experts meeting, systematically discuss and document how constructs can be deconstructed in the attributes they are composed of.

The third step is *(3): analyzing the construct under assessment*. From 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 to be included in the assessment. Following the analysis of steps 1 and 2, it may become clear that it is more efficient and effective to develop, for example, a traditional CBT (e.g., a multiple-choice test) rather than an MBPA. If so, then the most efficient method (in this case, a multiple-choice CBT) takes priority and should be developed; MBPA should be used only if it improves measurement.

If it is decided to continue development of an MBPA, then the assessment developer should collect as much information as possible about the construct from a content domain. The content domain is everything that can possibly be part of the assessment. The assessment developer should try to define the content domain as explicitly and thoroughly as possible (see Standard 14.9). Qualifications in vocational education and training (VET) are constructed on the basis of competency-based vocation profiles, which result from the analysis of a vocation as conducted by educational institutions and the labor market, reflecting what an experienced employee knows and does. Based on the competency-based vocation profile, a qualification profile describes in great detail what an entry employee should know and be capable of in order to be certified. The assessment developer can use the information in these profiles to define the limits of the domain of the construct.

Logically, the assessment developer cannot include in the assessment anything outside of the domain. Within the domain, there is a universe of tasks that might be designed and incorporated into the assessment (Mislevy et al. 1999; Mislevy 2011). Through systematic analysis of actual job behaviors, the assessment developer is able to design tasks that will form part of the assessment (Weekley et al. 2006). For example, by carefully observing the performance of qualified job incumbents, the assessment developer can isolate typical job behaviors that are the pillars of the vocation. This stage is generally characterized by a synthesis between subject matter experts (SMEs), and assessment experts (Downing 2006; Weekley et al. 2006).

This stage also includes cognitive analysis of the construct, indicating which cognitive steps students must take in completing actual job behaviors, and these should be strongly aligned with the assessment tasks (Mislevy et al. 1999). In the absence of this alignment, we can never make sound statements that generalize from an assessment setting to the real world. Think aloud methods are generally used to analyze individuals' cognitive strategies while performing specific tasks (Van Someren et al. 1994; Messick 1995).

Finally, using multiple perspectives (e.g., a competency-based profile, a qualification file, an analysis of job behavior, data from SMEs, and a cognitive analysis) the construct analysis delineated above informs a comprehensive argument explaining which factors of vocational behavior should be included in the tasks. In the task design step, then, the assessment developer should follow a strategy to select tasks that cover either the whole domain or the most important tasks within the domain. The latter strategy, which is fairly often used in vocational education, is also called the critical incidents technique (Flanagan 1954)—selecting the tasks that best predict future job behavior or are characterized as high risk, either for the student or for the organization, based on the construct analysis. This third step in the framework relates 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".

Document, from as many sources as possible or available (e.g., qualification profiles, job analyses, cognitive analyses, and subject matter experts), what constitutes the vocation, from the highest level of function profiles to the deepest level of thinking patterns.

The fourth step is *(4): designing assessment task(s) and operationalization of student behavior*. This step is defined by an exchange relationship with the previous stage *(3)*, in that the assessment developer should continually monitor whether tasks cover the domain of the construct and whether the task design uncovers gaps in the construct analysis (i.e., specific parts of the construct that did not surface during construct analysis but are important for assessment). The tasks should elicit student behavior that can be logged in support of claims about student skills, competencies, or knowledge. This step comprises four elements: *task attribute, task context, student behavior,* and *response type*.

The first element of task design is determining which attributes should form part of the tasks to be designed. An entire multimedia-based performance assessment is a construction of multiple tasks, and all tasks entail specific *task attributes*—for example, knowledge, attitude, skill, cognition, competency, or behavior (Frederiksen and Collins 1989; Mislevy et al. 2002). Task attributes can also differ in their level of complexity. Tasks in vocational education assessments usually comprise multiple attributes (Baartman et al. 2006; Klieme et al. 2008).

Next, the *task context* can be designed. To enhance authenticity, factors at play in a real-world context should also form part of the task context (Gulikers et al. 2004). Logically, this begins from designing an environment that resembles the real-world environment. 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. Clearly, then, task context incorporates more than just the physical context of the task.

The assessment developer can now define *student behavior*, which is the behavior students must actually demonstrate in the assessment task/s. The behavior that the task elicits in students provides evidence about the targeted construct (Mislevy et al. 1999), and the assessment developer should define student behavior in the smallest components that can be incorporated into a scoring model.

The final part of Step 4 is the *response type* that characterizes the tasks. MBPA includes a whole range of new response types for logging actual student behavior in the tasks—for example, speed, clicking behavior, navigational behavior through the virtual environment, typing, eye-tracking, and accuracy. Of course, both innovative and traditional item types can be incorporated in the MBPA (for an overview of innovative scoring in CBT, see also Williamson et al. 2006; Mayrath et al. 2012a, b; self-revealing reference 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). Logically, 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 Steps 2 and 3), and the intended students (see also Standard 3.6). The fourth step in the framework also relates 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". We can also recognize some basic models from the ECD framework (Mislevy et al. 1999) in this step: the student model and the task model.

Synthesize all information from the previous steps into meaningful tasks, which can be defined as tasks that require students to demonstrate behavior that truly provide the most informative inferences about their knowledge, skills, and abilities.

The fifth step is *(5): constructing the evidence model.* This step is schematically located between steps three and four, and relates to the exchange relationship between the former two steps. The evidence model implies that the assessment developer should construct and present a comprehensive and extensive argument that vindicates and explains why the constructed tasks (including attributes, context, student behavior and responses) should result in sound statements about students. In other words, there should be evidence that we can actually say something about students in real life (i.e., the criterion) based on their performance of the tasks in the assessment (i.e., the predictor) (see Standard 14.12). Often, the strength of the relationship can be determined after administration of the assessment has yielded results. However, it important to systematically analyze to what extent it seems plausible to expect valid results from performance of designed assessment tasks. For this reason, Downing (2006) remarked that systematic, thorough, and detailed documentation

for validity arguments should be collected continuously (Steps 3 and 12). Mislevy et al. (1999) discern two models within the evidence model: the statistical model and the evidence rules. The evidence model in the proposed framework refers to and builds upon the evidence rules specified in the ECD framework, as the assessment developer should provide evidence of the relationship between student behavior in assessment tasks and the construct.

Use a strong methodology (e.g., the extended argument-based approach to validation (Wools 2015)), to make a strong validity case that proves that the inferences to be drawn on basis of task performance, can hold under all circumstances.

The sixth step is *(6): constructing the score model*. Student behavior in the assessment has to be scored in order to construct a measurement model that will lead us from collected observed variables to claims about the construct. All observed student behavior during administration that contributes to an overall score forms part of a score model. Scoring may be quantitative as well as qualitative, and scoring rubrics assist in attaching weights to the scores and combining them into an overall score or result (Shepherd and Mullane 2008). According to Downing (2006), perfectly accurate scoring results in valid meanings, as they are anticipated by the assessment developer. Furthermore, the assessment developer should specify the scoring criteria and procedures for scoring in sufficient detail and clarity to make scoring as accurate as possible (see Standard 3.22). In their ECD framework, Mislevy et al. (1999) classify scoring mainly under the task model, but it also relates to their student model and evidence model because of the link between performance and evaluation.

Build a score model in which all types of evidence are identified, which means that scoring is not just about counting rights and wrongs, but also about investigating and identifying potential performance indicators that can be found in the log files.

The seventh step is *(7): constructing the measurement model*. Mislevy and Riconscente (2006) defined the measurement model as a mechanism to define and quantify the extent to which students' responses, as combined in the score model, inform statements we wish to make about these students. The administration of an assessment yields a certain amount of data, depending on the number and type of responses students must produce. Scoring ultimately supports claims of targeted knowledge or competency among students. By applying a measurement model to collected observed variables, we can infer from data to a scale of (a) latent variable(s). Psychometric models such as Item Response Theory (IRT) are part of the measurement model (see Standard 3.9). Mislevy et al. (1999) discussed the statistical model, which largely corresponds with our measurement model, defining it as part of the evidence model. The measurement model represents the relationship between students' degree of construct mastery (i.e., a latent characteristic) reflected in their performance and scores produced on the basis of performance. We specify the construction of the measurement model as a final step in the first stage because that seems most realistic for designing MBPAs, which may involve multiple different task types.

Include psychometricians, and together construct a hierarchical structure in which the observable variables (i.e., the scores from the previous step) are connected to the right constructs and attributes, then impose that structure on the chosen measurement model (e.g., a multivariate IRT model or a Bayesian Network) to calculate reliable proficiency estimates.

The assessment developer concludes the first stage with a detailed report of each step, this report functions as the guide for the second stage in which other parties are involved in the actual development of the MBPA.

*Development and administration*

What has been decided and reported in the first stage will be incorporated in an MBPA during the second stage, which consists of six steps and results in a functioning assessment. It is possible that specific observations in the second stage may require the assessment developer to return to the first stage, even after completion.

The eighth step, and the first of the second stage, is *(8): developing the ICT system and interface*. Logically, the development of an ICT infrastructure holds only if one is not already in place. The infrastructure should be able to incorporate and present multimedia and innovative items. We emphasize that this need not necessarily refer to immersive virtual environments, as in (serious) games or simulations (e.g., a flight simulator for the training of pilots). However, we do mean to include a virtual interface that, for example, can incorporate movies, animations, and avatars.

Make a decision regarding the platform to use for the presentation of the MBPA.

The ninth step is *(9): developing tasks and multimedia and implementing them in the ICT system*. Multimedia experts create the multimedia content to be incorporated in the tasks, based on the first stage task design. Now, the assessment developer can start filming, creating animations, avatars, and innovative item types. This is an iterative process of creating, evaluating, and adjusting, leading finally to the first version of the assessment. If the developed tasks indicate gaps between construct analysis and task design, or where specific parts of the designed tasks cannot be incorporated into the virtual environment, the assessment developer should return to the first stage to reconsider the designed tasks and the analysis of the construct for refit. This step also includes programming of the assessment and assignment of scores to tasks. Another decision that must be made during this step is whether, how, and when feedback will be provided.

Use the information from the previous stage, especially the fourth step, to build tasks and to have them resemble the task description as close as possible.

The tenth step is *(10): implementing in network*. The assessment can now be implemented in a network of computers, or installed or uploaded on single computers. Extensive guidelines exist on the development of computer-based assessment (e.g., ATP 2002; ITC 2005), and the assessment developer should use such guidelines in executing the previous three steps of assessment development.

Implement the assessment in a network (e.g., via internet or locally).

The eleventh step is *(11): pretesting the assessment*. The assessment should be pretested before administration, using a relatively small sample of students from the target population. However, the sample should be large enough to be able to draw meaningful inferences about the functioning of the tasks in the assessment.

Design a study set-up and have a representative sample of students perform the MBPA and use the data collected to build the evidence model as defined in the fifth step of the first stage.

The twelfth step is *(12): evaluating fitness for purpose*. If the assessment functions correctly, the next step is to start using it for its intended purpose. If the assessment does not function correctly, the eleventh step loops back to either the first step of the first stage or the first step of the second stage, and the assessment developer should repeat all steps from that point on. This highlights the relationship between the first and the second stages and the iterative character of assessment development.

Evaluate the findings from the pretest study and draw conclusion on the fitness for purpose of the MBPA.

The thirteenth and final step is *(13): administering the assessment*. The assessment can be administered when pretesting delivers the desired interpretation of assessment scores. This does not mean that the assessment is complete and can be endlessly reused. The quality of the assessment and its fitness for purpose should be constantly monitored by educational and assessment experts (see also Standard 3.25).

Administer the assessment live in the (high-stakes) educational setting.

## Step 3: validation of the prototype

We have validated the prototype discussed above on the basis of five semi-structured assessment expert interviews. We were able to filter and select 28 text fragments (i.e., statements) from the verbatim interview transcripts that specifically referred to the content concept, the usability concept, or one of the underlying categories. The distribution of the statements is shown in Table 1. In addition, we have added the questions that are answered by the statements in the first column of the table. These are not questions that were part of the interview, but they make the concepts and categories more explicit.

All the text fragments are listed in Table 2. Using these text fragments, it was possible to determine which steps or stages needed to be adjusted or refined for transformation of the prototype into the final framework. For efficiency, duplicate text fragments have been deleted; this only holds if the duplicate text fragments were used to refer to the same concept or category. For example, if two identical fragments from the same interviewee (or from two or more interviewees) referred to the completeness of the framework, then one was deleted. If one of the two statements referred to the completeness of the framework and the other to the usability of the framework, none was deleted.

## Step 4: adjustment of the prototype and final framework

The text fragments refer either to general factors that are applicable to the complete prototype framework (e.g., "more iterativeness") or to specific factors that are applicable only to an element of the framework (e.g., "eliminate the final step"). In finalizing the prototype, we have addressed both types of text fragments, adjusting the general flow of

**Table 1** Classification of number of text fragments in concepts and categories

| Questions | Categories | N |
|---|---|---|
| To what extent is the quality of the framework sufficient according to experts? | Content: general quality | 6 |
| To what extent are steps and/or stages in the framework complete according to experts? | Content: completeness | 8 |
| To what extent are steps and/or stages in the framework correct according to experts? | Content: correctness | 5 |
| To what extent are steps and/or stages in the framework coherent according to experts? | Content: coherence | 4 |
| To what extent is the usability of the framework sufficient according to experts? | Usability: general usability | 3 |
| To what extent does the framework fulfill a specific purpose in a practical setting? And who are the end users of the framework? | Usability: fitness for purpose | 2 |
| Total | | 28 |

**Table 2** Classification of verbatim text fragments in concepts and categories (the percentages indicate the proportion of experts making this statement during the interview)

| Questions | Text fragment |
|---|---|
| Content: general quality | The framework needs to display a more dynamic process. (60%)<br>The framework needs to display a more fluid process rather than a sequential or linear process. (60%)<br>The framework needs to display more of an iterative process. (40%)<br>The framework needs to incorporate more loops. (80%)<br>The framework needs more balance between the first and second stages. (100%)<br>The framework needs to be more of a cyclical, concentrically designed process, in which every step loops back to the previous steps and in which a prototype is continually updated throughout design and development. (40%) |
| Content: completeness | The framework lacks a step that facilitates the process of suppressing bias that results from the interface of the MBPA (20%)<br>The framework lacks a step that refers to a cognitive walkthrough of the MBPA (60%)<br>The framework lacks a step that refers to a paper-based walkthrough of the mockup of the MBPA (40%)<br>The framework lacks sufficient information about the feedback of performance in the MBPA to the students (60%)<br>The framework lacks information about the scoring of MBPA (e.g. is it fully automatic or blended) (60%)<br>The design of tasks and feedback should be incorporated into one step of the framework (80%)<br>The second stage of the framework needs to be elaborated to maintain the balance between both stages (100%)<br>The framework lacks an exit step before the developmental phase starts (60%) |
| Content: correctness | The first stage of the framework needs to emphasize in what way MBPA design differs from traditional test design (80%)<br>The first step of the framework should be "determining purposes" rather than "purposes and constructs" (40%)<br>The framework should constantly update a prototype of the MBPA after every step (20%)<br>The eighth step of the framework, in the second stage, should be "choosing an ICT interface" rather than "developing" one (80%)<br>The first stage should be "design" rather than "analysis and design", and the second stage should be "development" rather than "development and administration" (60%) |
| Content: coherence | The first and second stages of the framework need to be more parallel processes in order to make the framework more coherent (60%)<br>The final step of the framework, administration, is not part of design or development (40%)<br>ICT is much more important in the first stage of the framework in order to make the connection with the second stage (100%)<br>The framework needs to emphasize the relationship between design and development by incorporating more backward looping, also between stages (60%) |
| Usability: general usability | The framework needs to remain practical, in a sense that usability and the fulfillment of purposes is more important than a shiny layout (80%)<br>To improve usability, the framework needs a go/no go step before actual, costly development can start (80%)<br>The framework is useful if ICT experts are part of the design stage as well as the development stage; otherwise the gap between design and development becomes too large (100%) |

**Table 2** continued

| Questions | Text fragment |
|---|---|
| Usability: fitness for purpose | The framework is useful for experts leading a group of subject matter experts, however the framework needs to be simplified to make it useful for practitioners as well (80%)<br>The framework needs to place more emphasis on the developmental phase as well to make the framework useful for multimedia experts (60%) |

the framework as well as specific elements within it. A schematic flow diagram of the amended prototype (the final version of the framework) is depicted in Fig. 2.

*General adjustments*

First, we have tried to make the framework more dynamic and fluid by transforming the sequential form of the framework into a more parallel form and by adding more backward loops between steps and stages. The steps in the left part of the framework relate to assessment design while the steps in the right part of the framework relate to development. Second, the backward loops between each step and the original purpose determination also exemplify the iterative nature of the design and developmental process. Now, progress is constantly monitored by relating each step to the original purpose that the assessment should fulfill. In that way, we have also placed more emphasis on ICT from the earliest moments of assessment development. Third, we have removed the sub-steps of task design and development to make the framework more efficient and practical for its users. Fourth, we have removed the numbering of steps to emphasize the dynamic, fluid, and parallel nature of the framework. Fifth, we have slightly extended the development stage by adding a cognitive walkthrough step, which also relates to the interface of the assessment. In this way, we also provide more balance between both stages. Sixth, we have renamed the stages; the first stage is *design*, and the second stage is *development*. Finally, we have removed types of process and the constitution of the development team from the framework. Based on the interviews, we believe it is possible to develop MBPAs with relatively small teams, in which members take on different roles and work collectively through all steps of the framework.

*Specific adjustments*

We have also made some specific adjustments to the prototype framework. First, we have added steps to the final framework: a cognitive walkthrough and a separate step involving the construction of a feedback model. Second, we have also removed or changed steps. The eighth step of the second stage has been renamed *choosing ICT system and developing interface*, rather than *developing an ICT interface*. We have added the attributes determination to the first step of design—first, because it relates directly to the purpose of the assessment, and second, because it makes the framework more efficient and user-friendly. The score model and measurement model are also combined in one step in the final framework because the scores produced in an assessment strongly influence the choice of measurement model, and vice versa. Second, it also strengthens the efficiency and user-friendliness of the final framework. We have renamed *constructing a measurement model* to *choosing a measurement model*. The final stage, *administration*, has been removed, and

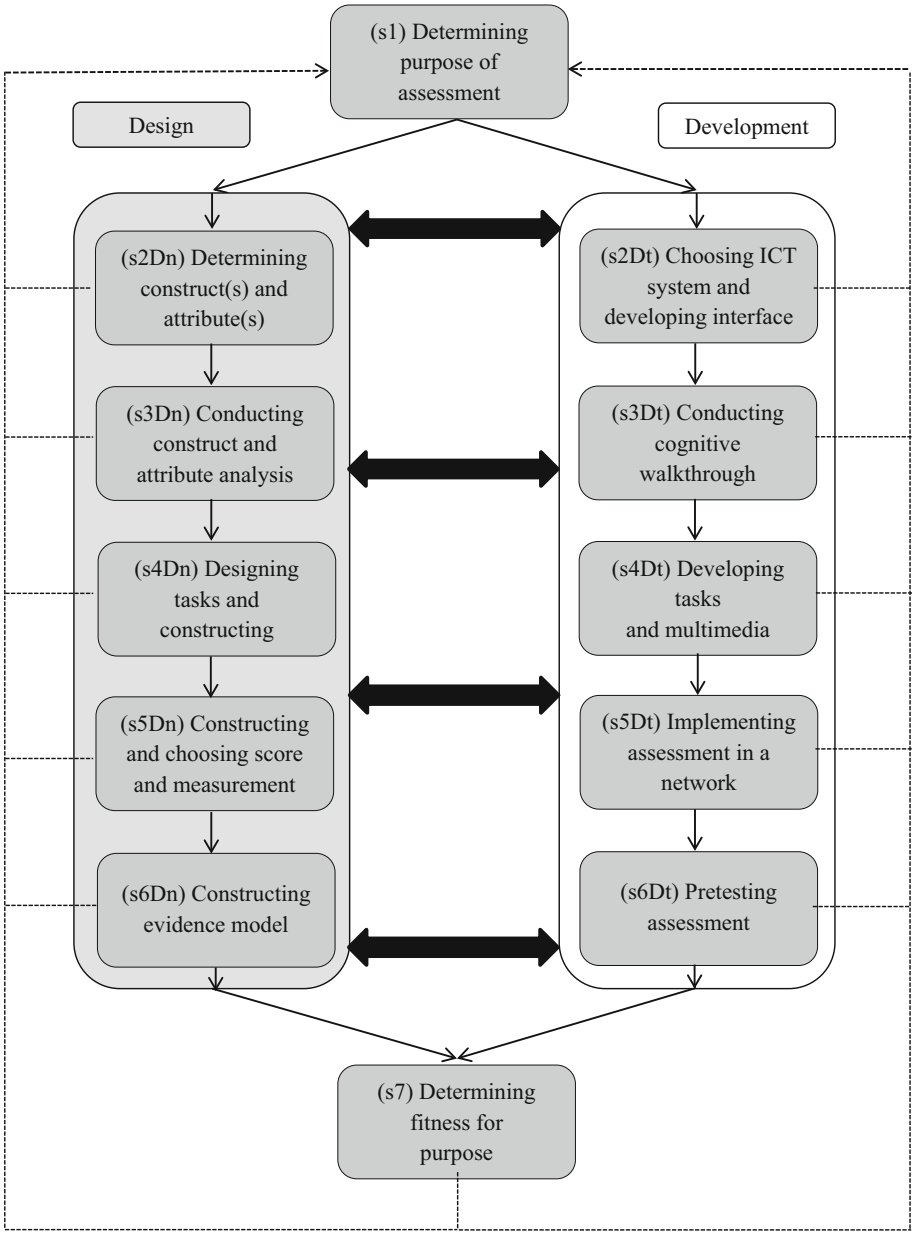


**Fig. 2** Flow schematic of the final framework for the design and development of multimedia-based performance assessment

we have removed the extensive reporting between first and second stages because the final version of the framework represents an integrated process involving both stages. By incorporating the adjustments discussed above in the prototype framework, we believe we have addressed most of the statements made by the experts during the interviews.

### Step 5: validation of the final framework

During the fifth step, the final framework was used to rebuild the pilot version of the MBPA (self-revealing reference 2014). Following the steps of the framework, the developmental process from pilot MBPA to final product can be seen as a validation strategy for the framework. The developmental process confirmed that the framework is functioning as intended, as the MBPA has improved considerably by comparison with the pilot version. Next, we will discuss the developmental process for the final MBPA and indicate to what extent the framework helped to improve the final MBPA by comparison with the pilot version.

We started by defining the purpose of the CSG assessment, which was defined in the "final attainment objectives" for certification of CSGs (s1). The purpose of the MBPA and the strategy for achieving that purpose was documented in a detailed project description. In general, it is advisable to write a project plan; especially in a practical setting where multiple specialists are interacting to create the product (i.e., the MBPA). The project description also included a systematic developmental plan and choice of ICT system for the MBPA (s2Dt), which followed the framework's steps. Although not specifically detailed in the framework, a risk analysis was conducted to document possible pitfalls in the project and how to avoid them, or how to handle them if they occurred. A risk analysis can be a useful tool in complex and multifaceted projects to hypothesize what might go wrong and to prepare for such events. The project organization was described in order to assign project team members' roles and to ensure clear communication between members. By comparison with our pilot endeavor (self-revealing reference 2014), the start of the project was much more structured, which in itself improved the chances of a successful outcome. For example, careful calculation of possible risks in the future steps of design and development can prevent mistakes and delays in building the assessment, which increases the possibility of an assessment that is fit for its original purpose. Furthermore, the iterative and dynamic nature of the framework is already reflected in the first step because it is important to constantly monitor, during the following steps, to what extent the instrument still adheres to its intended purpose defined in this step.

The design phase commenced by determining the constructs and attributes to be measured and analyzing them for translation into the MBPA tasks (s2Dn). This was done in collaboration with subject matter experts (SMEs) through multiple rounds of consultation. Of course, a lot was already known about the CSG's tasks from the instructional material and the final attainment objectives of the performance-based assessment. Additionally, the first author took part in a one-day course and performed the PBA to become a certified CSG. This material and knowledge was used to further develop a structure of constructs and attributes for the MBPA (s3Dn). After this step, developers should first return to the first step to make sure that the construct analysis aligns with the intended purpose of the MPBA, before actual products are being developed in the following steps.

Our framework ensured that attributes were mapped at the finest possible grain, enabling the design and development of assessment tasks that would yield the most interesting and relevant information about students' skills (s4Dn). Of course, this was done in collaboration with the SMEs, first building what we called an *assessment skeleton,* in which the general flow of the assessment was laid out, including required multimedia and items or assignments. Generally speaking, the assessment skeleton helps in avoiding the risk of mistakes in the actual development, as the MBPA is first built on paper. We therefore advise practitioners to always first build an assessment skeleton.

Although this is done on a relatively abstract level, it ensures that all constructs, final attainment objectives, and primary observables are incorporated in the tasks. Because the assessment skeleton is still a relatively coarse-grained representation, it is not sufficient for actually building the assessment. For that reason, we further elaborated the assessment skeletons into *assessment templates*, showing (screen by screen) what was to be presented during the course of the assessment. These templates were based on a cognitive walk-through of the assessment (s3Dt) and describe which buttons are present and which are not, which multimedia is presented, instructions for the student, what possible actions can be attempted and how these actions are scored (s5Dn). A cognitive walkthrough can also be done by a sample of students, using think aloud protocols. Practitioners may benefit from having students perform this step, because students might think or behave differently than SMEs do or how they can foresee how a student would behave in the assessment situation. Furthermore, it helps in making design choices. For example, do students expect a button or cue at the right hand side of the screen or left hand side? All these design features play an important role in the functioning of the MBPA. It is therefore important to connect to previous stages of construct analysis and assessment purpose. Do these still align, or are there variables that might cause construct irrelevant variance in the process of measurement?

The assessment templates enabled collection of multimedia (video and photo) material in 1 day, at a reconstructed job site in the Netherlands that is used for practice and performance-based assessments (s4Dt). Of course, depending on the size of the domain and MBPA this may take a considerably longer time. In this case, for example, we have only used one scenario in the MBPA. Increasing this number may of course result in more multimedia needed. It is advisable to involve professionals for the development of the multimedia. As with the previous step, vague multimedia which can be multi interpretable may lead to construct irrelevant variance. In that case, students correctly or incorrectly perform MBPA tasks because of the design of the multimedia and not because of their underlying proficiencies. Following the dynamic nature of the framework (depicted by the flow of arrows around the stages in Fig. 2), developers should continue to monitor whether the developed multimedia still meets the purpose delineation and construct analysis from the first steps, and the proposed assessment lay-out from the latter steps (assessment skeleton and templates).

In addition, the templates served as a primary input for design of the buttons needed in the assessment. For this step, we also hired a professional designer who was very experienced in designing intuitive, usable, and efficient interfaces for interactive websites. Furthermore, in combination with the buttons, the templates provided the necessary material for the programmer to build the structure of the assessment into our own assessment platform (s5Dt). It is important that an assessment developer, during this step, constantly monitors progress. Programming may take a considerable amount of time and effort. When the programmer goes astray, he should be brought back on the right path within foreseeable time, as it can become a costly endeavor otherwise. In fact, the interaction between the design and development stage may be most important during this step, as what has been designed now becomes a functioning product.

The next step was to test the assessment—first for its technical functioning and then for its psychometric functioning, in a pilot study (s6Dt). The assessment was administered via the Internet, and multiple test rounds enabled any remaining errors to be resolved, so ensuring that the assessment was technically functional. To give an impression of the MBPA, we present two screen captures in Figs. 3 and 4.


**Fig. 3** MBPA screen capture

Finally, construction of the evidence model consisted of building an argument for the validity of the assessment (s6Dn). In this case, to some extent, we already build an evidence model by using the framework itself, and professionals from several fields contributed to the process, with practical IT design and development of the assessment by an experienced web designer, multimedia expert, and programmer. The content was specified by subject matter experts and based on our previous experience of performance-based assessment. Another important aspect of evaluating the assessment is the empirical analysis of its performance properties by use of a measurement/statistical model, ultimately determining whether the MBPA has really met its goal (s7). The s6Dt, s6Dn, and s7 steps


**Fig. 4** MBPA screen capture

will be discussed in greater detail in a future publication (self-revealing reference, *manuscript submitted for publication*).

In particular, systematic reasoning about the assessment by use of the framework has improved considerably by comparison with the pilot version; for example, more and improved tasks in the assessment ensured sufficient reliability and validity of the MBPA. Furthermore, because the most important aspects of CSG performance were better understood by virtue of the extensive construct analysis in collaboration with SMEs, these aspects could really be emphasized in the assessment tasks and the score model. Furthermore, although many professionals collaborated in the project, communication and planning remained positive and on track.

## Discussion and conclusion

The point of departure for this article was to provide a framework for designing and developing multimedia-based performance assessment in vocational education. We have reported on the construction of a prototype framework for the design and development of MBPA, validating the framework through five semi-structured assessment expert interviews. We have reworked the prototype into final form on the basis of assessment experts' input, and we have used the framework to rebuild a new and improved version of an earlier pilot MBPA for measuring the skills of confined space guards.

The framework was grounded in theory and previous analyses by relating each of its steps to the most widely-accepted assessment development frameworks: the twelve steps for effective test development by Downing (2006), Mislevy et al.'s (1999) evidence-centered design framework for the design of assessments, and the Standards for Educational and Psychological Testing (AERA et al. 2004) as well as relevant other literature. Second, the framework was validated through interviews with five assessment experts, which indicated that the prototype needed to be adjusted in relation to several general aspects of the framework as well some specifics. These adjustments were made, and a final version of the framework was presented. Finally, the framework was used to develop a complete and operational MBPA.

We would like to emphasize that the current article focuses on the design and development of MPBA through the framework, and that the MBPA presented is only used to exemplify the application of the framework in a real-world setting. Research on the empirical functioning of MBPA is beyond the scope of this article. In a future publication, we will therefore focus on the psychometric functioning of the MBPA presented for the sake of the argument in this article (self-revealing reference *manuscript submitted for publication*). Furthermore, we would also like to emphasize that, although empirical functioning of the MBPA may say something about the quality of the framework, one cannot state that improper functioning of one MBPA immediately disqualifies the framework. One might say that a framework's real value comes to light during the next decade, when practitioners and researchers can work and experiment with it. Nevertheless, a limitation to this study is the fact that the framework has only been validated through the semi-structured assessment expert interviews and one case study (the MBPA that we have built). Future research and practice should be concerned with testing and validating the framework for multiple assessments in multiple (educational) settings. There is a special need for (quasi-)experimental research in which the functioning of assessments built by our framework is studied. Because, although semi-structured interviews, as we have already

remarked, have been used for validation purposes; the fact remains that there is, at least to some extent, subjectivity involved in this validation strategy. This is another limitation to our study. Maybe the prototype framework would have been adjusted differently if other experts had been invited for the semi-structured interviews.

The final framework has also been simplified by comparison with the prototype, making it easier to use and understand, not only for practitioners as well as researchers. We have demonstrated that our framework can be used for designing and developing MBPA in (Dutch) vocational education and training. Future studies could focus on using the framework in other educational settings as well (e.g., primary education, higher education, etc.). Thereby, the framework's validity could grow into other educational environments. Also, it might be used for building multimedia-based assessments for personnel selection or for other psychological disciplines. Finally, researchers and practitioners working on related types of technology-based assessment, for example simulation-based assessments in a more general sense, or game-based assessments, are asked to study and test the framework in their surroundings.

We believe that the coming decades will be characterized by a growing emphasis on multimedia-based performance assessment and related types of assessment in vocational education, to which this framework can be hoped to contribute.

**Acknowledgements** The authors are grateful for the insightful comments made by the assessment experts during the development of the prototype. The authors also thank the external assessment experts for their voluntary participation in the interviews. This research was supported by eX:plain.

**Funding** There was no specific funding for this research.

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.



## 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 computer-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.

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.

Barriball, K., & While, A. (1994). Collecting data using a semi-structured interview: A discussion paper. *Journal of Advanced Nursing, 19*(2), 328–335.

Dekker, J., & Sanders, P. F. (2008). *Kwaliteit van beoordeling in de praktijk [Quality of rating during work placement]*. Ede: Kenniscentrum Handel.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & 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.

Halverson, R., Owen, E., Wills, N., & Shapiro, R. B. (2012). *Game-based assessment: An integrated model for capturing evidence of learning in play*. ERIA working paper.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automated assessment of complex task performance in games and simulations (CRESST Research Rep. No. 775)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/R775.pdf.

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, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 3–22). Göttingen: Hogrefe.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (pp. 387–431). Westport, CT: Praeger.

Levy, R. (2013). Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educational Assessment, 18*(3), 182–207.

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

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

McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research*. New York, NY: Routledge Education.

McKenney, S., & Van den Akker, J. (2005). Computer-based support for curriculum designers: A case of developmental research. *Educational Technology Research and Development, 53*(2), 41–66.

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. (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).

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., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden, MA: Blackwell.

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C., et al. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology, 105*(4), 1100–1114.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.

RCEC. (2015). *Het RCEC beoordelingssysteem voor de kwaliteit van examens [The RCEC evaluation system for the quality of assessment]*. Enschede: Research Center for Examinations and Certification.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41–53.

Rupp, A. A., DiCerbo, K. E., Levy, R., Benson, M., Sweet, S., Crawford, A., et al. (2012a). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4,* 49–110.

Rupp, A. A., Nugent, R., & Nelson, B. (2012b). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining, 4*(1), 1–10.

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

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

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.

Shute, V. J., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y.-J., Jeong, A. C., et al. (2010). Modeling, assessing, and supporting key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 281–309). Boston: Springer.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.

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.

Vendlinski, T. P., Delacruz, G. C., Buschang, R. E., Chung, G. K., & Baker, E. L. (2010). Developing high-quality assessments that align with instructional video games. CRESST Report 774. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Wainess, R., Koenig, A., & Kerr, D. (2011). Aligning instruction and assessment with game and simulation design. CRESST Report 780. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*

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.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Hillsdale, NJ: Erlbaum.

Wools, S. (2015). *All about validity: An evaluation system for the quality of educational assessment (Doctoral dissertation)*. Enschede: Ipskamp Printing.

Wools, S., Sanders, P. F., Eggen, T. J. H. M., Baartman, L. K. J., & Roelofs, E. C. (2011). Evaluatie van een beoordelingssysteem voor de kwaliteit van competentie-assessments [Testing an evaluation system for performance tests]. *Pedagogische Studiën, 88,* 23–40.

Yen, W. M. (1993). Performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

**Sebastiaan de Klerk** is a researcher at Research Center for Examinations and Certification (RCEC), and a scientific advisor at eX:plain. The RCEC is a collaboration between the University of Twente and Cito (the national institute for test development in the Netherlands). His work focuses on designing, developing, and analyzing multimedia-based performance assessment in Dutch vocational education.

**Bernard P. Veldkamp** is a professor at the University of Twente, and director of the RCEC. He conducts research in the areas of educational, psychological, and health assessment. His interests focus on computerized assessment and data mining.

**Theo J. H. M. Eggen** is a professor at the University of Twente and a senior research scientist at Cito. His research interests focus on computerized adaptive testing and item response theory. He is also president of the International Association for Computerized Adaptive Testing.