Behaviormetrika (2018) 45:565–596
https://doi.org/10.1007/s41237-018-0064-x

ORIGINAL PAPER

# The design, development, and validation of a multimedia-based performance assessment for credentialing confined space guards

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

Received: 6 March 2018 / Accepted: 18 August 2018 / Published online: 28 August 2018


## Abstract

In this article, we present a study on the design, development, and validation of a multimedia-based performance assessment (MBPA) for measuring the skills of confined space guards in Dutch vocational education. An MBPA is a computer-based assessment that incorporates multimedia to simulate tasks. It is designed to measure performance-based skills. A confined space guard (CSG) supervises operations that are carried out in a confined space (e.g., a tank or silo). In the Netherlands, individuals who want to become certified CSGs must participate in a one-day training program, and pass both a multiple-choice knowledge test and a performance-based assessment. In the first part of this article, we focus on the design and development of the MBPA, using a specific framework for design and development. In the second part of the article, we present a validation study. We use the argument-based approach to validation to validate the MBPA (Kane in Educational measurement. American Council on Education and Praeger Publishers, Westport, 2006 and J Educ Meas 50(1):1–73, 2013). More specifically, the extended argument-based approach to validation is used (Wools et al. in CADMO 18(1):63–82, 2010 and Stud Educ Eval 48:10–18, 2016). The approach suggests using multiple sources of validity evidence to build a comprehensive validity case for the proposed interpretation of assessment scores (Kane 2006, 2013) and to evaluate the strength of the validity case (Wools et al. 2010, 2016). We demonstrate that MBPA scores can be used for their intended purpose; students' performance in the MBPA can be used as the basis for making a CSG certification decision.

Communicated by Ronny Scherer and Marie Wiberg.

This research was supported by eX:plain.

✉ Sebastiaan de Klerk
s.dklerk@explain.nl

Extended author information available on the last page of the article

**Keywords** Design · Development · Validation · Multimedia-based performance assessment

## 1 Introduction

The growing capabilities and wide availability of technology have enabled a whole new generation of technology-driven assessments, which are far more elaborate than mere computer-based versions of earlier item-based pen-and-paper tests (Quellmalz and Pellegrino 2009; Clarke-Midura and Dede 2010). The new generation of technology-based assessment both expands and deepens the domain of assessment (Levy 2013). Technology ensures that more flexible and context-driven presentations of tasks and environments are possible in computer-based assessment (CBA), which can lead to a broader and better understanding of what students have learned. With technology, assessment designers are enabled to design assessments that measure complex aspects of student knowledge and skills that were difficult, if not impossible, to measure using traditional paper-based tests or performance-based assessment (PBAs) (Clarke-Midura and Dede 2010).

The use of technology in assessment has grown rapidly (Clarke-Midura and Dede 2010). Although most CBAs are still based on pen-and-paper tests, in recent years, more emphasis has been placed on innovations in technology-based assessment (cf., Koenig et al. 2010; Quellmalz et al. 2010). The innovations began with the introduction of innovative item types (Scalise and Gifford 2006; De Klerk 2012), but have now progressed to simulation and game-based assessments (Rupp et al. 2012; Levy 2014; Mislevy et al. 2014; De Klerk et al. 2015). In simulation-based assessment (SBA), test takers are confronted with dynamic or interactive features in the tasks they are set (Levy 2013). This can be in the form of an animation or movie that accompanies a task, but it can also be an interactive feature within a task (Parshall et al. 2002). Levy considers tasks in SBA to embody the same conception of complexity as the definition of complex tasks provided by Williamson et al. (2006). In short, a task is complex if: (a) multiple processes are required; (b) multiple elements of task performance are captured; (c) there is a potential of high variability in the data; and (d) it is dependent on other tasks or actions in the assessment.

A second group of highly innovative assessment techniques is found in the so-called game-based assessment (GBA) (Mislevy et al. 2014). In GBA, test takers play a real video game, while all actions are logged in the background. This can play a role in performance evaluation. The rationale is that game environments provide opportunities for students to demonstrate their skills in highly contextualized and interactive situations (Klopfer et al. 2009). In this way, it is possible to measure new aspects of the students' skills and to measure other aspects better. An example of GBA is provided by Shute (2011). She uses the term stealth assessment for the measurement of student competencies in video games. As students are playing and progressing through the levels of a game, their actions, decisions, use of tools, navigation skills, and so on are being logged and then used to update beliefs about student competency in a particular skill, without students even noticing that they are in an assessment situation.

Technology-based assessment is an umbrella term, suggesting that there is not one type of technology-based assessment, but rather that technology can manifest itself in assessment on a continuum of complexity, interactivity and fidelity (Parshall et al. 2002; De Klerk 2012). Complexity refers to the number of actions a student can perform in an assessment; interactivity indicates the extent to which an assessment permits manipulation of elements in the assessment; and fidelity represents the degree to which the assessment corresponds with a real-world setting. On the left side of the continuum, computer-based transformations of item-based pen-and-paper tests score lowest on complexity, interactivity and fidelity. These assessments only require students to click on one of the item's alternatives (low complexity); do not change during administration of the assessment (no interactivity); and the items are not embedded in a context (low fidelity). The most common example of this type of CBA is a multiple-choice test presented via a PC. Virtual reality or serious games are found at the right of the continuum for assessment. These assessments score highest on complexity, interactivity and fidelity. Usually, such assessments require students to perform complex and interactive tasks in situations that represent a real-world setting.

In this article, we discuss the design, development, and validation of a technology-based assessment that is positioned close to the right of the continuum of complexity, interactivity, and fidelity. The study presented in this article is particularly important because it is one of the first scientific endeavors of the evidence-based design, development, and validation of a computer-based assessment used to assess practical skills that are currently assessed in a hands-on performance-based assessment. The main research question is: Do test takers' results from a multimedia-based performance assessment (MBPA) provide a valid estimation of their skills in being a confined space guard?

The assessment we present incorporates images, animations, and videos for creating complex and interactive tasks in a simulation of a real-world setting. We call this type of technology-based assessment MBPA, because the tasks in the assessment are for a large part constructed of multimedia and are used to measure student skills that were previously measured by a PBA. The purpose of the MBPA we discuss here is to measure the skills of CSGs after they have completed vocational training (De Klerk et al. 2014, 2018). The skills that a student must demonstrate during the assessment were determined by a commission of experts.

A CSG supervises operations that are carried out in a confined space. A confined space is any space which by design has limited or restricted means of entry or exit, is large enough for a person to enter to perform tasks, has the potential for a significant hazard to be present, and is not intended for continuous occupancy. An example of a confined space is a fluid storage tank. Most confined spaces are found in petrochemical plants. Different kinds of operations take place in confined spaces—for instance, cleaning or welding. By Dutch law, these operations have to be carried out under the supervision of an individual who is certified as a CSG. In the Netherlands, certification of CSGs takes place after a one-day training program, which concludes with a multiple-choice test to measure students' knowledge of theory, and a PBA to measure students' skills in a simulated practical setting.

Although PBA has been discussed and supported as a valuable tool for formative and diagnostic assessment of students (Gulikers et al. 2004; Roelofs and Straetmans 2006), the research is less supportive in cases where PBA is used as a summative assessment. This is foremost because PBAs are prone to measurement errors resulting from several sources, including task, occasion and rater sampling variability (Shavelson et al. 1993; Cronbach et al. 1997; Dekker and Sanders 2008). Shavelson et al. (1999) provide an example in which task and occasion sampling are confounded. In such cases, their combined effect strongly increases the rate of measurement error. These findings indicate that students' scores in a PBA do not solely represent students' proficiency in a particular skill, but are influenced by the specific task that they are assigned, the occasion of the assessment, and the raters who judge their performance. In addition, this study found that it was difficult to define the exact source of measurement error because of the complex relationship between task and occasion sampling. In addition, PBAs are expensive and time-consuming when compared to CBA. Although much research has been carried out on (innovative) technology-based assessments in many educational fields, considerably less research has been devoted to the use of technology-based assessment as an equivalent to PBA in a more practical educational field (i.e., vocational education and training). The purpose of the current study, therefore, is to design, develop, and evaluate an MBPA for credentialing CSGs in Dutch vocational training.

We aim to contribute to the theory and practice of innovative assessments by presenting, developing and validating a new type of assessment: the MBPA. We think that this computer-driven measurement instrument does not need to be a game or high-fidelity simulation, by definition. Instead, we argue that MBPA, a simulation type assessment, which is based only on audio/video material, animation, and interactive interface features, may suffice for measuring vocational constructs. Next, we discuss the design and development of an MBPA for CSGs in Dutch vocational training, before presenting the results of the empirical validation study in which we applied an extended version of the argument-based approach to validation (Kane 2006, 2013; Wools et al. 2010, 2016).

### 1.1 Design and development of a multimedia-based performance assessment

The measurement instrument was designed and developed following a two-stage, twelve-step framework for designing and developing an MBPA (De Klerk et al. 2018). The focus of the framework is on the integrative, iterative and adaptive character of the design and development of an MBPA. In the framework, design and development are regarded as parallel operations, which are carried out simultaneously from the earliest moments of MBPA design and development (see Fig. 1). In addition, design and development are part of a continuous process of monitoring developmental progress in relation to the MBPA's desired final state. If needed, the design and development process can be adapted to retain the alignment between the current state of development and the original purpose of the assessment.

We now discuss the framework in greater detail and explain how we used it to build our MBPA. The framework encompasses two parallel stages—design and

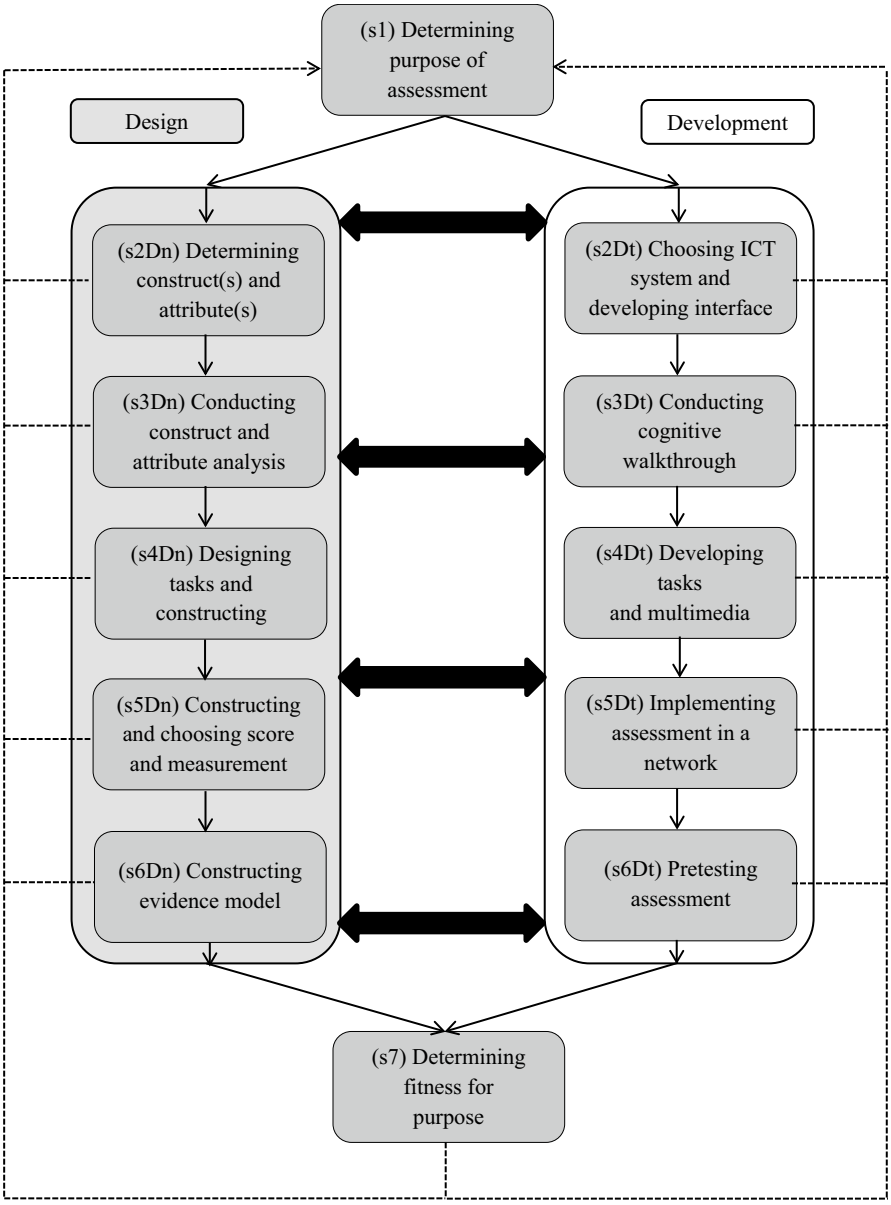


**Fig. 1** Flow schematic of the final framework for the design and development of multimedia-based performance assessment

development—that consist of five steps each. Including the two general steps of purpose determination and fitness for purpose determination, which go beyond the two stages, the framework comprises twelve steps in total. The 12-step framework revolves around the concepts of integrated, iterative and adaptive design and

development. These concepts define MBPA design and development as simultaneous rather than sequential processes, as a cyclical process of continuous evaluation of the current status and the purpose of the assessment, and as a process of prototype realization, testing, and refining.

To emphasize the integrated, iterative and adaptive character of the framework, we alternate continually between the steps that are part of the design stage and the steps that are part of the development stage (i.e., from left to right in the framework). First, a summary of each step's content is given. We then discuss the step with reference to the development of our measurement instrument.

Before we began designing and developing, we first laid out the purpose of the assessment. In other words, we defined the proposed interpretation of the assessment scores. The purpose of the MBPA is to measure CSG skills, as defined in the "final attainment objectives," so that it can be used as a tool for making a CSG certification decision. The desired interpretation can only be met if we can make a chain of inferences, from actual student performance to the certification decision. Because this chain of inferences is part of the interpretive argument in the extended argument-based approach to validation (Kane 2006, 2013), we will come back to this during the discussion of the validation study.

The purpose of the MBPA and the strategy to attain this purpose were documented in a detailed project description. The project description included a systematic developmental plan, which followed the framework steps. A risk analysis was carried out to hypothesize possible pitfalls in the project and how to avoid them or to handle them if they did occur. The project organization was described to assign the project team roles and to ensure clear communication between members of the team. In addition, a time table and budget were established.

We then started the design phase by determining the constructs and attributes that we intended to measure at the finest grain size. This was done by analyzing constructs and attributes in collaboration with subject matter experts (SMEs), through multiple rounds of consultation. Of course, the nature of CSG tasks was already known through the instruction material and final attainment objectives of the PBA. In addition, the first author took part in a one-day course and performed the PBA to become a certified CSG. We used this material and knowledge for further specification of the constructs and attributes of the MBPA.

Based on this analysis, the tasks in the assessment were designed and developed in collaboration with the SMEs. We first built what we called an assessment skeleton, in which the general flow of the assessment was laid out, including the multimedia needed and the items or tasks at different sections in the MBPA. Because the purpose of the MBPA is to make a summative certification decision, based on student performance, we decided not to include feedback in the MBPA. The only feedback that students received was their total score at the end of the MBPA. The assessment skeleton was made on paper and was a relatively abstract delineation of the MBPA. However, it ensured that all constructs and attributes, based on the analysis from the previous step, were incorporated in the MBPA tasks. Because an assessment skeleton is still a rather coarse-grained representation of the assessment, it was not sufficiently defined for developing the MBPA.

We, therefore, further elaborated the assessment skeletons into assessment templates. The assessment templates were also made on paper and showed—screen by screen—what would be presented to the students during the MBPA. To help with building the templates, we also performed a cognitive walkthrough, which means that we mentally went through the MBPA. This gave us the opportunity to experience the MBPA before it was developed and ensured that we did not overlook essential elements of the MBPA.

We were then able to complete the templates that show which buttons are presented at different stages of the MBPA, which multimedia are presented, what instructions are given to the student, and what possible actions can be carried out by the student. Based on the assessment templates, we were able to develop the multimedia (video and photo material) in 1 day at a reconstructed CSG job site in the Netherlands, which is used for professional practice and PBAs. The screen by screen assessment templates also served as primary input for the graphical designer. We hired a professional graphical designer who was experienced in designing intuitive, usable and efficient interfaces for interactive websites. Based on the assessment templates, he was able to build the buttons needed to operate the MBPA (e.g., to proceed to the next item, to request extra information, to zoom in or out, etc.). The general interface and design of the MBPA were made industrial so that it would fit the CSG profession.

We had already decided to work with GMP-X, our own online-based ICT system and assessment platform. Once the graphical designer had delivered the MBPA interface, we imported it into our assessment platform. The programmer then built the structure of the assessment, as laid out in the assessment templates, and to make sure that the MBPA could be accessed online. After the MBPA was implemented on our network, we performed several test rounds. During the test rounds, we clicked on all buttons to see whether they worked, and we checked that all answers were written correctly in the scoring database. After the testing phase, all initial bugs were repaired. We then had a fully functioning MBPA, which students could access via the internet. In Sect. 2, we discuss the items and tasks in the MBPA and provide screenshots.

The next step was to determine whether the MBPA is fit for purpose. To answer this question, we refer to the entire bottom part of the framework presented in Fig. 1. In other words, to answer this question, we first have to do a pretest, in which a representative sample of students performs in the MBPA. We then have to decide how to score student responses in the MBPA, and we apply a measurement model to analyze student scores. Finally, evidence has to be collected that supports the supposed interpretation of the MBPA scores, as defined in the very first step of the framework. The validation study, which is presented below, can be regarded as the evidence model because it will determine whether the MBPA is fit for purpose.

### 1.2 Validation of the multimedia-based performance assessment

We applied the argument-based approach to validation (Kane 2006, 2013) to validate the interpretations assigned to the MBPA's scores. More specifically, an

extended version of the approach, which includes an evaluation argument, was used (Wools et al. 2010, 2016). The argument-based approach to validation, developed by Kane (2006, 2013), presents a standardized framework for the validation process, consisting of a developmental stage, in which an interpretive argument is formed, and an appraisal stage, in which a validity argument is formed.

### 1.3 Interpretive argument

The interpretive argument can be regarded as a chain of reasoning from student performance on the tasks in the MBPA to the decision whether to certify (Kane 2006, 2013). Figure 2 shows the chain of reasoning in the interpretive argument of the MBPA. First, students' performance of the MBPA tasks is converted into a numerical score. Students get one point for each correct action or answer in the MBPA. The sum of the points is the total score. We consider the total score to be a representation of the test domain score, which consists of all possible tasks that could be administered in the MBPA. In the following step of reasoning, we extrapolate the test domain score as a representation of the skills or competences to be measured. In other words, we expect that a high MBPA score will correspond with a strong mastery of CSG skills and that a low MBPA score will correspond to weak mastery of CSG skills. Finally, the scores are extrapolated from the competence domain to the practice domain. In other words, can we regard students' MBPA scores not only as a representation of skill or competence within the assessment context, but also outside the assessment context, in their future professional life (Gulikers 2006)? Only if we can make this extrapolation, can we take a meaningful decision about certifying CSG students, based on MBPA performance. The argument-based approach to validation has also been used in certification testing (Kane 2004).

### 1.4 Validity argument

The claims within the interpretive argument (Kane 2006, 2013), which is the extrapolation from performance to decision, need to be supported and evaluated for the MBPA. That is, we need to provide both analytical and empirical evidence for evaluating the assumptions that the interpretive argument provides. The analytical evidence follows naturally from the framework that was used to design and develop the MBPA. The empirical evidence was gathered through a representative sample of students that took part in the pretest of the MBPA.

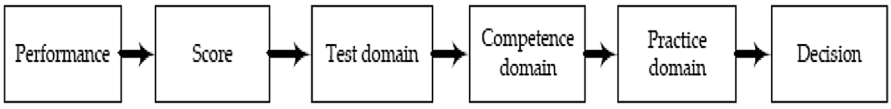


**Fig. 2** Chain of reasoning in the interpretive argument (Adapted from Wools (2015))

### 1.5 Analytical validity evidence

The first element of analytical evidence is that the use of the framework provides a tool for sufficient coverage of the content domain of CSG knowledge and skills. The logical structure of the framework has ensured that the tasks and assignments in the MBPA require the student to apply all knowledge and skills necessary to become certified as an entry-level CSG. Fortunately, the construct analysis demonstrated that the content domain (constructs and attributes) is relatively small, which made it possible to include all the knowledge and skills needed in the CSG profession. This process was done with the input of SMEs and provides sufficient evidence of the content validity of the MBPA.

The second element of analytical evidence is that the use of the framework provides a tool for an appropriate element of (cognitive) complexity with regard to the tasks and assignments in the MBPA. The MBPA should not only cover a domain in terms of content, but also in terms of complexity. The assessment should be neither too easy nor not too difficult. The tasks and assignments in the assessment need a sufficient level of complexity to ensure valid assumptions about and interpretations of the assessment scores. The assessment skeleton and the templates, based on a cognitive walkthrough of the assessment, ensured that the entire process—from the construct analysis stage to the task and assignment design stage—incorporated a complexity analysis. For example, the construct analysis revealed that it is not only essential that students are able to read different sections of a work permit, but that they are also able determine whether a work permit is valid and can perform the correct actions to make the work permit valid where required. In the MBPA, therefore, having students check whether a work permit is valid would not be complex enough. Instead, we had to find a way to test whether students would also report this to the operator so that the necessary changes or additions to the work permit could be made. Using the following screen, students had to report to the operator that the work permit was not valid and needed to be revised. Using this step-by-step approach, from construct analysis to task design through assessment templates and a cognitive walkthrough, allowed us to ensure the right level of complexity in the MBPA's tasks and assignments. In the Sect. 2, we first discuss the empirical evidence. We then look at the empirical evidence for the validity argument of the MBPA (Kane 2006, 2013). Finally, we evaluate the validity argument (Wools et al. 2010, 2016).

## 2 Method

### 2.1 Participants

The participants in the empirical study were 55 CSG students (1 female and 54 male). Participants ranged in age from 19 to 64 years, with a mean age of 40.4 years ($\sigma = 11.5$). They were asked to participate in the MBPA after they had completed training. Participation was voluntary, and students did not receive a payment. All participants were recruited at one of the two locations where they had completed

their vocational training, the PBA and the pen-and- paper multiple-choice knowledge test. We had hoped to include more participants in the study, but this was not possible because of the low yearly number of students and the widespread distribution of students across the country. Yet, the 55 participants were a representative sample of the population for two reasons. First, all participants had just completed the course. Secondly, there is no reason to assume that the sample differs from the target population in age, ethnicity, or education. We can, therefore, assume that the sample of students included in this experiment is a representative sample.

## 3 Materials

*Knowledge test* Immediately after training, the students performed a knowledge-based pen-and-paper test, consisting of 21 multiple-choice questions, each offering 3 alternatives. According to the assessment regulations, as determined by the assessment commission, the students had 30 min to complete the test and had to answer 14 or more questions correctly to pass the test. The test is composed of randomly selected items from an item bank and the reliability of the test is $\alpha = 0.64$.

*Performance-based assessment (PBA)* In the PBA, students performed CSG tasks in a reconstructed, yet realistic situation. Before the PBA started, students were randomly assigned to one of the four scenarios that would be played during the PBA. A scenario always started with receiving the work permit from the operator (a role that is played by the rater/examiner). The student then had to collect more information regarding the work permit and the operations to be carried out. Students had to ask for a walkie-talkie and had to ensure that the right channel was selected and that the walkie-talkie was functioning properly. An accomplice of the rater played the role of a worker who was going into the confined space to carry out operations (e.g., cleaning a tank). Students had to discuss how to communicate with the worker when he or she was in the confined space. Several aspects regarding the confined space did not match work permit specifications (e.g., tools lying around in an unsafe manner). The student was supposed to notice these issues and report them to the operator. In addition, using a wind direction flag and several emergency gathering points (indicated by icons), the student had to indicate the direction to the gathering point in the case of a gas alarm. The rater also judged the extent to which the student took the time to proactively inspect the environment around the confined space. The worker then entered the confined space and made one or more intentional mistakes, which the student was supposed to identify and correct. Finally, after the worker had spent some time in the confined space, an alarm went off. The student then had to follow the correct emergency procedures. The assessment ended when the student and the worker were both at the emergency gathering point and had notified the operator that they were safe. Figure 3a, b gives an idea of what the PBA looks like.

The rater used a rubric consisting of 19 criteria to evaluate the student ($\alpha = 0.35$). All 19 criteria were marked as insufficient or sufficient by the rater. From the 19 criteria (e.g., "tests the walkie-talkie"), 9 are gating items (Judd 2009) (e.g., "recognizes and reacts to an alarm"). If a student's performance in any one of these items is insufficient, he or she will fail the PBA. Because the rubric consists of rather


**Fig. 3** The confined space guard is checking the environment of the confined space for potentially dangerous features (**a**) and the confined space guard (with white helmet) discusses communication with worker (**b**). The rater observes from a distance

narrowly defined actions, we decided to develop an additional rubric of 12 criteria ($\alpha = 0.8$) that focused on the upper-level (behavioral) constructs of the CSG vocation. These constructs comprise communication, proactivity, environmental awareness, and procedural efficiency. Raters were asked to assess students on a scale ranging from 0 (e.g., "Student does not demonstrate any communication skills") to 3 (e.g., "Student communicates very well"). Hence, students could get between 0 and 12 points for the new criteria, and between 0 and 19 points for the original criteria. Both rubrics were marked by the rater. We also calculated the combined score of both rubrics ($\alpha = 0.73$). Translated versions of these rubrics can be found in the Sect. 6.

*MBPA* Another primary instrument in the study was the MBPA itself. The case that students were tested on involved the cleaning of a tank on a petrochemical plant by two workers. This case was built in the online environment using multimedia. Students started in an office setting where the contractor handed work permit to the CSG and one of the workers. In this setting, students had to ask the contractor for an explanation of the work permit, check the work permit for blanks or errors, ask for a walkie-talkie, and then test the walkie-talkie. The setting then changed to the confined space itself. In this setting, students were required to determine the correct escape route in case of an emergency. Students had to ensure that the environment was safe to work in, and that there were no irregularities with regard to the work permit and the actual situation at the confined space. In the next phase, students had to supervise two workers who were cleaning the interior of the confined space. Finally, students had to react to a plant alarm.

Students were required to watch multimedia elements and to answer questions in the MBPA. For example, students were presented with a digital work permit that they could inspect using the zoom and navigation buttons in the MBPA. They then had to answer the question accompanying the work permit (e.g., "Is the first column of the work permit completed correctly?"). In this case, both the work permit and the question were presented simultaneously. Where students had to watch a film fragment, the film fragment was presented first, followed by the question. In this way, students were confronted alternately with multimedia and different types of questions. The MBPA consists of 35 questions: 18 yes/no questions, 5 multiple-choice questions (with 4 options), 4 fill-in-the-blank questions, 1 multiple-response question, 1 rank order question, and 6 so-called intervention questions. The intervention questions required students to watch two videos of workers cleaning a tank and to intervene whenever their actions were incorrect. Students could intervene by clicking on a big red stop button that was located beside the video screen. Students were told that they could only click the stop button three times. In other words, if they clicked the stop button when there were no faulty actions, they had one less chance to press the button when it was required. Figure 4 illustrate the MBPA.

*Questionnaire* After students had performed in the MBPA, they were asked to complete a questionnaire comprised of 15 items ($N = 15$) addressing the following: (1) their background characteristics (e.g., "What is the highest level of education you have completed?"); (2) computer experience (e.g., "On a scale ranging from 1 (never) to 5 (every day)—How often do you play videogames on a computer?"); and (3) MBPA usability (e.g., "On a scale ranging from 1 (strongly disagree) to 5 (strongly agree)—I was comfortable with the interface of the MBPA"). Reliability for computer experience and usability were $\alpha = 0.64$ and $\alpha = 0.75$, respectively. The questionnaire was based on a translated version of the System Usability Scale (Bangor et al. 2008) and a questionnaire on the use of the Internet and computers at home, developed by Cito (2014). As a result, students' computer use and the usability of the MBPA could be classified as subscales of the questionnaire. A translated version of the questionnaire can be found in the Sect. 7.

### 3.1 Procedure

Students participated in their training and completed the pen-and-paper test immediately afterwards. Then, depending on the experimental condition to which they were randomly assigned, students either first performed in the PBA and then in the MBPA ($N = 27$), or the other way around ($N = 28$). Students were not allowed to confer with each other between both assessments, so that it was impossible for them to exchange knowledge regarding the MBPA. For the MBPA, students were seated behind a laptop or PC. All assessments were administered under supervision of the first author. Students logged in with a personal login on the assessment platform (GMP-X). Before they began, students were shown how the assessment functioned using pictures and text. No time limit was imposed on students either for completing individual items or the entire assessment. Student questions were answered by the supervisor, but only if the questions were related to the way the assessment


**Fig. 4** MBPA screenshots

functioned. After students finished the assessment, they had to complete a questionnaire that was upside down on their table.

## 4 Results

### 4.1 Empirical validity evidence

The first element of empirical validity evidence was constructed around the internal structure of the MBPA. The internal structure of the assessment was assessed through a psychometric analysis [based on classical test theory (CTT)] of the

**Table 1** Mean, standard deviation, and 95% confidence interval for measures (1000 sample bootstrapping performed)

| Measure | Mean | $\sigma$ | Rel. | 95% CI | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| MBPA (35p) | 22.54 | 3.45 | 0.94 | 21.59 | 23.57 |
| PBA (19p) | 17.35 | 1.52 | 0.35 | 16.87 | 17.72 |
| PBA (12p) | 9.11 | 2.45 | 0.80 | 8.39 | 9.87 |
| PBA (total) | 26.25 | 3.78 | 0.73 | 25.15 | 27.25 |
| MC test (21p) | 17.89 | 1.85 | – | 17.35 | 18.39 |
| MBPA time (min) | 29.2 | 8.33 | – | 26.91 | 31.85 |
| Q-computer exp. | 2.88 | 0.91 | 0.64 | 2.6 | 3.12 |
| Q-MBPA usability | 2.93 | 0.76 | 0.75 | 2.72 | 3.14 |

The questionnaires on computer experience and the usability of the MBPA used a five-point Likert scale

**Table 2** MBPA test characteristics (1000 sample bootstrapping performed)

| *M* | GLB | $S^2$ | SEM | Skewness | Kurtosis |
|---|---|---|---|---|---|
| 22.54 ($\sigma = 3.45$) | 0.94 | 11.1 | 0.83 | 0.014 | 0.488 |

*M* the mean score on the test, *GLB* greatest lower bound, $S^2$ the variance of the scores, *SEM* the standard error of measurement

answers that the students gave. We decided to use CTT rather than item response theory (IRT) because of two reasons. First, the number of participants was too low to perform IRT analyses. Secondly, we believe that the goal of validation can also be reached using CTT, as the CTT indices (e.g., reliability and item indices) provide enough information about the quality of a test. As mentioned earlier, the assessment was composed of 35 items. In total, students could get one point for each correct answer. The mean score on the test was 22.5 ($\sigma = 3.44$), with a 95% confidence interval [21.6, 23.6], which indicates that, compared to the PBA (for which students on average received more than 17 out of 19 points), the test was more difficult for the students. The maximum score (obtained by two students) was 30, whereas the minimum score was 14 ($N = 1$). The standard deviation is rather low, which means that most students achieved a score around the mean. The average time that students needed to complete the assessment was 29 min ($\sigma = 8$). The minimum amount of time spent on the assessment was 19 min and the longest was 58 min. The high standard deviation and the wide bandwidth between minimum and maximum indicate that there is a lot of variance between the time students spent on the assessment. Table 1 provides mean, standard deviation, and confidence intervals for the MBPA.

Table 2 provides other characteristics of the MBPA. The variation of the scores is relatively high (11.9). The distribution of the scores is not skewed (0.014), but the kurtosis is high (0.488). This indicates that a relatively large portion of the variance is caused by extreme values on both sides of the distribution, with most students' scores being clustered around the mean. In addition, the standard error of measurement (SEM), which was calculated by multiplying the standard deviation of the assessment scores by the square root of 1 minus the reliability, is 0.83. The SEM was defined as the standard deviation of the mean of a hypothetical normal

distribution of many administrations of the same test. In other words, it represents the possible distance between the observed score and the true score in a CTT context. For the MBPA, the SEM was relatively low, which means that the students' true score was relatively close to their observed score. To be precise, 95% of the scores on the hypothetical distribution for a student with a mean score of 22.5 fall between 20.8 and 24.2. The reliability of the MBPA is high—with a Greatest Lower Bound (GLB) of 0.94. Of course, the high reliability is in accordance with a low SEM. We looked at the GLB as this was the index automatically calculated by the software used (TiaPlus (Cito 2006) and because this index is regarded as one of the most accurate estimations of the reliability of a test (Verhelst 2000; Sijtsma 2009).

To establish further support for the internal structure of the MBPA, we also looked at CTT indices for the individual tasks and assignments in the MBPA. Table 3 displays the CTT indices for the 35 items in the test. The second and third columns indicate what part of the content was assessed with the item and which item format was used. The *p* value of the test is the proportion of students that answered the item correctly. That is, a high *p* value is associated with a relatively easy question, whereas a low *p* value points to a difficult question for this group of students.

The GLB reliability index is presented for the MBPA. This was not possible for the other measures because of the number of items; for these measures, Cronbach's alpha is presented. We cannot report the reliability of the multiple-choice test, because the items are randomly selected from the item bank and, therefore, every student has a different test.

The mean p value of the test was 0.62, which indicates that the assessment was rather difficult for this group of students. The $r_{it}$, the point-biserial correlation between the item scores and the total test score in which the item score is not excluded from the total score, gives an indication of the discriminative power of an item. The higher the value, the better the item can discriminate between good and poor performers. A low $r_{it}$ means that students that score high on the overall test score low on the item, whereas students that score low on the overall test score high on the item. Conversely, a high $r_{it}$ means that good performers do well on the item and poor performers do worse on the item. The mean $r_{it}$ of the MBPA is 0.22. There is reason to believe that the mean $r_{it}$ could be improved, as the quality of the individual items in the MBPA fluctuates (see Table 3). To summarize, we have provided evidence that the MBPA has a strong internal structure. Although there is room for improvement, all indices fall within acceptable levels.

The next element of empirical validity evidence is used to support the external structure of the MBPA, in particular the convergent validity (based on the PBA scores) and the discriminant validity (based on the questionnaire and the multiple-choice knowledge test). In other words, the MBPA scores are expected to correlate with the PBA scores, but not with the questionnaire and to a lesser extent with the multiple-choice test. Although the multiple-choice test also measures aspects of the CSG construct, it does so on a different level. The correlations (Spearman's rho) between the measures of the experiment are presented in Table 4. Spearman's rho is used because there is a monotonic relationship between the variables and because the measures do not meet the assumptions of normality and linearity. For example, on the 19-point rubric, most students score 17–19 of the criteria as correct. It is,

**Table 3** CTT indices of 35 items in the multimedia-based performance assessment

| Item | Content | Type | $p$ value | $r_{it}$ |
|---|---|---|---|---|
| 1 | Explain WP | MC-4 | 0.36 | 0.38 |
| 2 | Ask for addition WP | Yes/no | 0.76 | 0.26 |
| 3 | Ask for addition WP | Yes/no | 0.82 | 0.40 |
| 4 | Ask for addition WP | Yes/no | 0.44 | 0.29 |
| 5 | Ask for addition WP | Yes/no | 0.89 | 0.31 |
| 6 | Ask for addition WP | Yes/no | 0.72 | 0.37 |
| 7 | Ask for addition WP | Yes/no | 0.13 | 0.01 |
| 8 | Ask for addition WP | Yes/no | 0.89 | 0.24 |
| 9 | Check WP | Yes/no | 0.36 | 0.33 |
| 10 | Check WP | Yes/no | 0.86 | 0.05 |
| 11 | Check WP | Yes/no | 0.87 | − 0.03 |
| 12 | Check WP | Yes/no | 0.44 | 0.37 |
| 13 | Explain WT | MC-4 | 0.87 | 0.22 |
| 14 | Channel WT | MC-4 | 1 | – |
| 15 | Test WT | Fill in | 0.95 | 0.23 |
| 16 | Battery check WT | Fill in | 0.15 | 0.01 |
| 17 | Escape plan | MC-4 | 0.49 | 0.53 |
| 18 | Work preparation | Yes/no | 0.96 | − 0.07 |
| 19 | Work preparation | Yes/no | 0.95 | − 0.16 |
| 20 | Work preparation | Yes/no | 0.18 | 0.23 |
| 21 | Work preparation | Yes/no | 0.86 | 0.05 |
| 22 | Work preparation | Yes/no | 0.60 | 0.12 |
| 23 | Work preparation | Yes/no | 0.33 | 0.46 |
| 24 | Work preparation | Yes/no | 0.89 | 0.07 |
| 25 | Environment check | Multiple sel. | 0.34 | 0.37 |
| 26 | Report to operator | MC-4 | 0.96 | 0.05 |
| 27 | Agree communication | Fill in | 0.53 | 0.13 |
| 28 | Error intervention | Intervention | 0.18 | 0.26 |
| 29 | Error intervention | Intervention | 0.63 | 0.40 |
| 30 | Error intervention | Intervention | 0.31 | 0.24 |
| 31 | Error intervention | Intervention | 0.27 | 0.56 |
| 32 | Error intervention | Intervention | 0.29 | 0.35 |
| 33 | Error intervention | Intervention | 0.87 | 0.06 |
| 34 | React to emergency | Rank order | 0.81 | 0.16 |
| 35 | Report to operator | Fill in | 0.89 | 0.16 |
| Mean | | | 0.62 | 0.22 |

*p value* the proportion of the students who have answered the item correctly, $r_{it}$ the correlation between the item score and total score

therefore, better to look at the rank order of the scores on the different measures than at the linear correlation. As can be seen, the correlations between the MBPA and the rubrics used in the performance assessment are 0.37 ($p < 0.01$) and 0.38 ($p < 0.01$), respectively, for the 19-point rubric and the 12-point rubric, which are both significant. We have also combined students' scores on both rubrics to get a total rubric

**Table 4** Correlations, means, and standard deviations of measures (1000 sample bootstrapping performed)

| Measure | Mean | $\sigma$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| 1. MBPA | 22.5 | 3.5 | | | | | | | |
| 2. PBA (19) | 17.4 | 1.5 | 0.39** | | | | | | |
| 3. PBA (12) | 9.1 | 2.5 | 0.38** | 0.68*** | | | | | |
| 4. PBA (total) | 26.3 | 3.8 | 0.43** | 0.84*** | 0.96*** | | | | |
| 5. MC Test | 17.9 | 1.9 | 0.30* | 0.2 | 0.21 | 0.23 | | | |
| 5. MBPA (time) | 29.2 | 8.3 | 0.01 | − 0.13 | − 0.2 | − 0.22 | − 0.05 | | |
| 6. Q-Computer exp. | 2.9 | 0.9 | 0.09 | 0.12 | 0.15 | 0.16 | − 0.01 | 0.1 | |
| 7. Q-MBPA usability | 2.9 | 0.8 | 0.18 | 0.15 | 0.09 | 0.16 | − 0.06 | − 0.18 | 0.42** |

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

score. The correlation between the total rubric score and the MBPA score is strongly significant ($r_s$=0.43 ($p < 0.01$)). We also applied a correction for attenuation and found that the correlation then improves considerably (respectively to 0.92, 0.59, and 0.70). This indicates that the correlation is strongly diluted by measurement error. Considering the low reliability that was established for the performance-based assessment, it was very likely that the correlation would increase.

Of course, there is also a strong significant correlation between both rubrics used in the assessment ($r_s$=0.68, $p < 0.001$). We also performed a linear regression analysis to see the extent to which performance in the MBPA could predict performance in the PBA. Because of the negative skew of the distribution of the rubrics, especially the 19-point rubric, we first subtracted each score from the highest score obtained, plus one, and then performed a log transformation (see Field 2009).

We did this for the 12-point rubric, the 19-point rubric, and the total rubric score to get a reliable comparison. The regression analysis for the 19-point rubric showed a significant effect ($F(1,53) = 4.365$, $p < 0.05$), which indicates that the MBPA score can account for 7.6% of the variation in the PBA score. We performed the same analysis for the 12-point rubric, which was also significant ($F(1,46) = 5.544$, $p < 0.05$), with an explained variance of 10.1%. Finally, we performed a regression analysis for the total rubric score, which was also significant ($F(1,46) = 5.905$, $p < 0.05$), with an explained variance of 11.4%. The total rubric score was the best predictor for performance in the MBPA. Unfortunately, the rater forgot to complete the 12-point rubric on one assessment, which explains the lower number of students in the second analysis.

To provide further evidence, if MBPA performance is related to PBA performance, then we would expect students who had failed their PBA to score significantly lower on the MBPA than students who had passed their PBA. Unfortunately, because of the high scores on the PBA, the group of students who failed their PBA was very small ($N = 8$). This makes it quite difficult to interpret the results and the following conclusions should therefore be viewed with some caution. However, because the analyses seem to indicate interesting results we do decided to include them in the article. The group of students who passed the PBA had a mean score of 23.2 ($\sigma = 0.46$) and the group of students who failed had a mean score of 20.1 ($\sigma = 1.1$). We used an independent samples t test to check whether the groups differed significantly, which was the case ($t(53) = -2.563$,

$p < 0.001$, $d = 0.70$). We then performed a logistic regression analysis to check the extent to which the MBPA score could predict whether a student will pass or fail their PBA. The MBPA score is treated as a continuous predictor in the logistic regression analysis. The dependent variable (success in PBA) is a dichotomous outcome variable (0 = failed, 1 = passed). The results of the analysis can be found in Table 5. The analysis demonstrated that the MBPA score made a significant contribution to predicting whether students failed or passed the PBA ($\chi^2(1, 55) = 5.09$, $p < 0.05$). The odds ratio ($e^{\beta}$), as a measure of effect size, for the MBPA score is 1.39 with a 95% confidence interval [1.04, 1.86]. This suggests that a one unit increase in the MBPA score increases the probability of being successful in the PBA (i.e., passing the PBA), with 1.39. To summarize, the overall correlations and regression analysis provide evidence for the convergent validity of the MBPA.

The absence of a correlation between students' MBPA scores and their background characteristics, the questionnaire ratings (computer experience and usability of the MBPA) and multiple-choice test results should provide evidence for discriminant validity. The background characteristics are age, education, and ethnicity. Age was not correlated with assessment scores ($r_s = 0.00$, $p > 0.05$). We calculated the biserial correlation coefficient for education. The biserial correlation coefficient is used when one variable is a continuous dichotomy (Field 2009).

First, we divided the students into two groups (low education vs. high education). The low education group consisted of students who had continued education up to high school or lower vocational education ($N = 26$, $M_{\text{MBPA}} = 21.83$, $\sigma = 3.07$), whereas the high education group consisted of students who have had continued education from middle level vocational education and upwards ($N = 27$, $M_{\text{MBPA}} = 23.08$, $\sigma = 3.60$). We calculated the point-biserial correlation [which is for true dichotomies (Field 2009)], and then converted it into the biserial correlation. Although education and student MBPA score were positively correlated, this effect was not significant ($r_b = 0.19$, $p > 0.05$). For ethnicity, we were especially interested in two groups: students with Dutch ethnicity ($N = 40$, $M_{\text{MBPA}} = 22.8$, $\sigma = 3.35$), and students with another ethnicity ($N = 15$, $M_{\text{MBPA}} = 22.78$, $\sigma = 3.41$). We calculated the point-biserial correlation between ethnicity (0 = Dutch, 1 = other) and the students' MBPA scores. Again, we did not find a significant correlation ($r_{pb} = -0.01$, $p > 0.05$). Overall, student's background characteristics were not related to their MBPA performance, which supports the discriminant validity of the MBPA.

We found further support for discriminant validity, because there is no significant correlation between student MBPA scores and their computer experience ($r_s = 0.09$, $p > 0.05$). Additionally, the MBPA score and student rating on the usability of the MBPA are not correlated ($r_s = 0.14$, $p > 0.05$). It is interesting to note that there is a

**Table 5** Logistic regression analysis of passing performance-based assessment

| Predictor | $\beta$ (SE) | Wald's $\chi^2$ ($df = 1$) | $p$ | $e^{\beta}$ | $e^{\beta}$ (95% CI) | |
|---|---|---|---|---|---|---|
| | | | | | L | U |
| Constant | − 5.4 | 3.05 | 0.08 | 0.00 | | |
| MBPA score | 0.33 | 5.09 | 0.02 | 1.39 | 1.04 | 1.86 |

The dependent variable in this analysis is performance-based assessment success coded so that 0 = failed and 1 = passed

significant correlation between students' computer experience and their rating of the usability of the MBPA ($r_s = 0.42$, $p < 0.01$), but that there is no significant correlation between the time spent on the MBPA and the score obtained ($r = 0.07$, $p > 0.05$).

However, there is a significant correlation between the multiple-choice knowledge-based test and the MBPA ($r_s = 0.3$, $p < 0.05$), which may indicate that, at least to some extent, the multiple-choice test and the MBPA do measure the same construct(s). Interestingly, there is no significant correlation between the PBA scores and the multiple-choice test scores ($r_s = 0.09$, $p > 0.05$).

Finally, we determined the number of misclassifications at six different levels of MBPA cutoff scores (50, 55, 60, 65, 70 and 75%). No misclassifications would mean that all students ($N = 8$) that failed their PBA would also fail the MBPA, and that all that passed the PBA would also pass the MBPA ($N = 47$). The results are presented in Table 6. Although the lowest cutoff percentage (50%) results in the least number of misclassifications, which can be explained by the fact that we have a small group of students who failed their PBA, it is most interesting to note the difference in fail–fail classifications between the cutoff points at 55 and 60%. At the 55% cutoff point, only two students who failed their PBA would also fail the MBPA, whereas this number increased to 7 at the 60% cutoff score. Therefore, a cutoff score at approximately 60% would be most defensible empirically. In addition, we looked at the number of misclassifications at different levels of cutoff scores using Cronbach's alpha in TiaPlus (Cito 2006). The analysis indicates that the least misclassifications take place when the cutoff score is placed at 50% (see also Table 6). In TiaPlus, the GLB reliability coefficient is the point of departure to estimate the misclassifications at the different cutoff levels.

### 4.2 Validity evaluation

In the previous paragraph, we presented validity evidence. In this paragraph, the validity evidence is used and evaluated. The argument-based approach is applied to prove the proposed interpretation of the MBPA (Kane 2006, 2013). Wools et al. (2010, 2016) distinguishes three criteria to evaluate the validity and the process of validation for an assessment. The first criterion evaluates the interpretive argument, the second criterion evaluates the different elements of validity evidence, and the third criterion evaluates the validity argument as a whole.

With regard to the first criterion, we can say that there are a substantial number of inferences. Following the chain of inferences, we have to go from a student's performance and accompanying raw scores to meaningful statements regarding performance in a practice domain and then to a final certification decision. This is an indication of the complexity of the inferences that we wish to make with the MBPA. Nevertheless, these inferences are required to ensure that the MBPA can be used for its intended purpose. The question concerns whether the interpretive argument addresses the correct inferences and assumptions (Wools et al. 2010, 2016). We specified the interpretive argument in sufficient detail so that the chance on possible voids or inconsistencies in our inferred reasoning is kept at a minimum. According to the extended argument-based approach to validation, each inference in the chain (the arrows in Fig. 2) should have at least one

**Table 6** Number of misclassifications MBPA–PBA at different cutoff score levels

| Cutoff percentage | | Performance-based assessment | | *N* misclassifications |
|---|---|---|---|---|
| | | Fail | Pass | |
| MBPA-50% | Fail | 2 | 2 | 8 |
| | Pass | 6 | 45 | |
| MBPA-55% | Fail | 2 | 3 | 9 |
| | Pass | 6 | 44 | |
| MBPA-60% | Fail | 7 | 12 | 13 |
| | Pass | 1 | 35 | |
| MBPA-65% | Fail | 7 | 30 | 31 |
| | Pass | 1 | 17 | |
| MBPA-70% | Fail | 7 | 35 | 36 |
| | Pass | 1 | 12 | |
| MBPA-75% | Fail | 8 | 39 | 39 |
| | Pass | 0 | 8 | |

warrant, a supporting warrant (or backing), and rejected rebuttals. A rebuttal indicates a circumstance in which the warrant or backing would not hold (Wools et al. 2010). We can demonstrate this by looking at each inference in the chain individually. The first inference is from performance to score, or the scoring inference. The same performance always leads to the same score (warrant), but this will only hold if the MBPA is correctly programmed (rebuttal). Furthermore, there has to be an objective scoring system (backing), which needs to be used objectively (rebuttal). In our case, the MBPA has a standardized and objective scoring structure. Scoring has already been addressed in the assessment skeleto*n*, which was made in collaboration with SMEs.

The second inference is from score to test domain, or the generalization inference. The tasks in the MBPA provide a representative sample of tasks of the whole test domain (warrant), but this only holds true if there are enough tasks in the MBPA (rebuttal). The use of the framework and collaboration with SMEs ensures that there are enough tasks in the MBPA. This is also demonstrated by the validity evidence, because we have shown that the reliability of the assessment is high.

The third inference is from test domain to competence domain, or the first extrapolation inference. The tasks in the MBPA provide an adequate measure of CSG skills (warrant), but this will only be the case if the MBPA does not suffer from construct underrepresentation or construct irrelevant variance (rebuttals). The MBPA is a good representation of content, authenticity and complexity (backing). The fact that we did not find variables that correlated with the MBPA score, except for the rubric scores which are allowed to correlate with MBPA scores, means that we can reject construct underrepresentation or construct irrelevant variance for this MBPA.

In addition, the tasks in the MBPA have been designed on the basis of a very extensive construct analysis, in collaboration with SMEs, which ensures that the MBPA is representative regarding authenticity and complexity.

The fourth inference is from the competence domain to the practice domain, or the second extrapolation inference. The practice domain is correctly operationalized within the competence domain (warrant), but only if all relevant aspects of CSG performance are represented in the competence domain (rebuttal). Again, evidence is provided by the fact that the design and development followed a well-defined and structured process, in which the steps from the framework were followed, in collaboration with SMEs. All tasks and assignments that currently take place in the PBA were transformed into a computer-based equivalent, which indicates sufficient representation.

The last inference is from the practice domain to the final certification decision. There should be a cutoff score (warrant) that is correct (rebuttal). We have provided several possible cutoff scores with an accompanying number of misclassifications. We did not apply a formalized standard setting procedure. However, we provided different cutoff scores, which SMEs can use in their decision for a cutoff score.

The second criterion for validity evaluation relates to the validity evidence (Kane 2006, 2013). Is the presented validity evidence plausible and representative for the assumptions that we wish to make with the MBPA scores? In other words, are the inferences justified by our validation study (Wools et al. 2010, 2016)? Each element of validity evidence should relate to and substantiate one or more inferences in the chain of reasoning. Wools et al. 2010, 2016 indicate that an evaluation status should then be assigned to the inference as a whole. The status is justified when warrants and backings on the validity elements are accepted, and possible rebuttals are rejected. With the evidence presented above, we argue that the validity elements give enough support for all the inferences in the interpretive argument.

Finally, the third criterion focuses on the outcome of the validation process or the validity argument as a whole. The question to be answered is: Is the validity argument as a whole plausible (Wools et al. 2010, 2016)? The validity argument can only be plausible when both the first and second criteria are met. As with the second criterion, the third criterion is somewhat subjective but boils down to taking all elements of validity evidence into account and then deciding whether the argument is strong enough to substantiate the validity of the assessment scores and final interpretations. In our case, we can say that all criteria have been met. The validity evidence provided in this article is plausible because every inference in the chain of reasoning, from performance to decision, can be substantiated by evidence.

## 5 Discussion and conclusion

The aim of this study was to investigate the design, development and validation of an MBPA for credentialing CSGs in Dutch vocational education. The first part of the article focused on design and development. In particular, a 12-step framework was used. This was specifically built for the design and development of the MBPA. The second part of the article focused on the validation of the developed MBPA, using the argument-based approach to validation (Kane 2006, 2013). More specifically,

an extension of the approach was used, as presented by Wools et al. (2016), which includes an evaluation of the interpretive and validity argument.

Design and development were simultaneous processes; some of the team's project members worked on the ICT-development side, whereas others focused on content analysis and design, constantly providing feedback to each other in regular project meetings. We had the advantage that we already knew much about the vocation to be assessed and that we had our own assessment infrastructure. However, we found the framework a useful and efficient guide during development of our MBPA.

After development, a random and representative sample of 55 CSG students performed in the PBA and the MBPA. The goal of the validation study, using the extended argument-based approach to validation, was to build strong interpretive and validity arguments and to evaluate the strength of both arguments. The interpretive argument is a chain of inference which we can develop from raw student performance data to meaningful statements about their future functioning in the practice domain. The validity argument is composed of different elements of validity evidence, both analytical and empirical. If the evidence was convincing, then we could extrapolate from student performance to the practice domain. In this case, the MBPA would prove to be a sound and adequate measurement instrument for credentialing CSGs.

The first analytical element of validity evidence referred to the content validity of the instrument. We demonstrated that the structured design and development process, through the use of the framework, ensured that the content of the MBPA was a full and representative sample of the content domain. We have also shown how different steps in the framework secured the correct complexity of tasks and assignments in the MBPA, which was the second analytical validity element. More specifically, the process from construct analysis to task design in assessment templates, through a cognitive walkthrough of the assessment, offered the chance to design tasks at the correct level of (cognitive) complexity. These elements of validity evidence followed naturally from using a comprehensive framework for the structured design and development of the MBPA. The other three elements are based on the empirical data produced by students performing in the MBPA.

Our first empirical element of validity evidence related to the internal structure of the assessment. Based on a psychometric analysis and indices from classical test theory, we discussed the MBPA's general test characteristics and the characteristics of individual items in the assessment. The results show that the overall indices fall within acceptable levels, although the evidence could be stronger. Some tasks in the MBPA function very well, whereas others have insufficient item characteristics. The correlation between some items and the MBPA score is low or even negative, and some items are too easy or too difficult. This can be explained by the fact that this is a first version of the MBPA. In test development, it is not uncommon to have multiple rounds of pretesting and revising before the assessment reaches its final form. We anticipate finding an explanation for the worst functioning items, either in terms of content or in the quality of the item. Revising or replacing these items in the MBPA would further strengthen the internal structure.

A second empirical element of validity evidence was based on the external structure of the assessment. The convergent and discriminant validities of the MBPA were used as indicators of the strength of the external structure. Students' MBPA

scores were correlated with their PBA scores, as evidence for the convergent validity of the instrument. Students' scores on the PBA (by independent rubric and total rubric score) moderately correlated with their scores on the MBPA. The fact that the correlation is not stronger may be because of several reasons.

First, the rubrics used for rating student performance on the PBA do not show much variance in sum score. We had foreseen this problem for the 19-point rubric and, therefore, developed the 12-point rubric to induce more variation in student PBA scores. And, indeed, it does produce slightly more variance in students' scores, but not enough to make a real difference. It is statistically difficult to establish strong relationships between two variables when one of the variables has almost no variance. Secondly, the correction for attenuation on the correlations indicated that there is a stronger relationship between the measures, but that it is diluted by measurement error. Finally, although the PBA is the best external criterion, it may not be the ultimate criterion, because PBA might have (psychometric) limitations and we cannot be certain about the quality (i.e., validity and reliability) of the scores. Future research in this area should also try to find criteria that are outside of the assessment domain. One external criterion could, for example, be the students' future job appraisals, made by their managers. A future study on the subject could include a strong analysis of the quality of the PBA, for example, through generalizability theory (Brennan 2001).

The group of students who failed their PBA was relatively small, which makes it difficult to draw firm conclusions. However, this group did score significantly lower on the MBPA than the group of students who passed their PBA. The regression analysis substantiated this finding by demonstrating that students' scores on the MBPA were a significant predictor for failing or passing the PBA. This provides important evidence for the convergent validity of our MBPA.

To establish evidence for the discriminant validity of the MBPA, we looked at the correlation between students' MBPA results and their background characteristics. We did not find any significant relationships, which is favorable evidence for the MBPA. More evidence follows from the non-significant relationship between MBPA score and computer experience, MBPA usability, and time spent on the assessment. Another critical limitation for the MBPA is the fact that the scores moderately correlate with the knowledge-based multiple-choice test, whereas the MBPA is based on the PBA. On the other hand, knowledge is always a prerequisite for successful performance in a PBA. This finding, therefore, is explicable and does not inevitably diminish the quality of our validity argument.

The third empirical and fifth overall element of validity evidence is related to the number of misclassifications at different cutoff score levels. Of course, the number of misclassifications is related to the reliability of the MBPA, which is high. We used the GLB index (Verhelst 2000; Sijtsma 2009), which provides the best estimate of reliability by giving the greatest lower bound of the reliability. That means that the reliability of the test is at least as high as the GLB indicates. In this case, the GLB is 0.94.

There are some general limitations to our study. First, the sample size is small. It was difficult to get a substantial number of students to participate in the study, because many assessment locations do not have internet connections or computers and the locations themselves are spread all over The Netherlands. The assessment

itself takes place, on average, 15 times per year, per location. Sometimes, a group can consist of less than five students, which suggests how difficult it can be to get a sufficient number of students to participate. On the other hand, because there are not many students per year and we have collected data for 7 months, we can say that we have included a substantial amount of data in our study. Furthermore, if we look at background, the sample does not systematically differ from the general population.

Secondly, the quality of the PBA is a limitation. Although the PBA is professionally organized, only one rater is used. The rater also plays the part of the operator in the assessment. The 19-point rubric, used for rating a students' performance, shows little to no variance, which makes it difficult to draw firm conclusions regarding a comparison between the MBPA and the PBA.

Thirdly, although the number of participants was not sufficient and CTT provides enough support for the validation of the MPBA, a future study should also focus on analyzing data with more complex IRT-based models.

A final limitation to our study is that the authors were involved in the design of the MBPA as well as the validation, which might have (unconsciously) influenced certain decisions made during both processes. With regard to this limitation, we ask other researchers to independently perform a validation study on the MBPA. On a general level, the limitations discussed above indicate that our study shows that it is possible to test procedural practical skills through a computer-based test. In no way, this means that this is possible for each and every (type of) skill or for each and every MBPA test form. In our opinion, each MBPA should be built and evaluated with the same rigor as the MBPA discussed in this article.

There are several implications of our study. First, from a practical point of view our study shows that policy makers and test developers can decide on using MBPA for assessing procedural or practical skills that previously were assessed in practical performance-based assessments. From a theoretical point of view, this is one of the first scientific endeavors on the validation of an MBPA. In that sense, this is only the beginning and there is a whole new era of innovative assessment research to look forward to. As already mentioned, future research should focus on using more complex psychometric modeling (e.g., IRT or Bayesian network modeling). Concurrently, technology will also progress, which enables future practitioners to develop more advanced, interactive, and complex assessments. It is, therefore, essential to put equal effort in scientific research and theory building.

To conclude, using the extended argument-based approach to validation, we have built a comprehensive validity case, composed of analytical and empirical validity evidence. The validity argument was constructed to substantiate the interpretive argument. Through the validity argument, we have demonstrated that the interpretive argument is plausible and appropriate, which means that the MBPA scores can be used according to their proposed interpretation. In other words, a CSG student performance in the MBPA can be used to decide on his or her accreditation.

## Appendix A

| No. | Criterion | Insufficient | Sufficient |
|---|---|---|---|
| 19-point rubric | | | |
| 1 | Carries out last minute risk analysis correctly | | |
| 2 | Wears the prescribed personal protective equipment (PPE) | | |
| 3 | Asks the operator for explanations about the work permit | | |
| 4 | Asks the operator about the functioning of walkie-talkies | | |
| 5 | Discusses communication with workers in the confined space | | |
| 6 | Discovers and reports deviations in the work permit | | |
| 7 | Carries out actions in accordance with the work permit | | |
| 8 | Determines wind direction and the correct escape route | | |
| 9 | Explores the confined space environment before operations | | |
| 10 | Asks about the dangers of the last substance stored in the confined space | | |
| 11 | Tests the walkie-talkies at the confined space | | |
| 12 | Registers the workers entering and leaving the confined space | | |
| 13 | Recognizes and corrects incorrect PPE of workers | | |
| 14 | Recognizes and reacts to an alarm | | |
| 15 | Alerts workers in the confined space during an alarm | | |
| 16 | Verifies the number of workers leaving the confined space during an alarm | | |
| 17 | Reports the alarm to the operator via the walkie-talkie | | |
| 18 | Leaves the confined safe and tidy | | |
| 19 | Attaches a “do not enter” sign to the confined space | | |

| Construct/points | 0 | 1 | 2 | 3 | Rater judgment |
|---|---|---|---|---|---|
| 12-point rubric | | | | | |
| Communication | Student shows (almost) no communication skills | Student shows poor communication skills | Student shows sufficient communication skills | Students shows good communication skills | |
| Proactive attitude | Student has (almost) no proactive attitude | Student has poor proactive attitude | Student has sufficient proactive attitude | Student has good proactive attitude | |
| Environment awareness | Student (almost) doesn’t observe environment | Student observes environment poorly | Student observes environment adequately | Student observes environment well | |
| Procedural efficiency | Student (almost) doesn’t follow procedures | Student follows procedures poorly | Student follows procedures adequately | Student follows procedures well | |
| Total number of points: | | | | | |

## Appendix B

Questionnaire

General questions (6 questions)

1. *Sex.*

   O M

   O F

2. *Place of residence.*

   _____

3. *Age*

   ____

4. *What is your highest level of education?*

   O None

   O Elementary school

   O Lower level high school

   O Lower vocational/technical education

   O Middle level high school

   O Middle level vocational/technical education

   O Higher level high school

   O Higher level vocational/technical education

   O University

   O Postgraduate education

5. *What is your ethnic origin?*

O Dutch

O Turkish

O Moroccan

O Surinam

O Antillean

O Other. Please specify:_____

Computer use (3 questions)

6. *Do you have a computer?*

O No

O Yes, but no internet

O Yes, with internet

7. *Do you use a computer frequently?*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

8. *Do you use a computer for games frequently?*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

Multimedia-based Performance Assessment (7 statements)

9. *I found the computer assessment easy to use.*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

10. *I need more support to handle the computer assessment.*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

11. *I would like to do computer assessments more often.*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

12. *I could easily find the buttons on the screen.*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

13. *I think computer use has negatively impacted my assessment result.*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

14. *I quickly felt familiar with the computer assessment.*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

15. *I have participated seriously in the experiment.*

O Strongly disagree

O Disagree

O Neutral

O Agree

O Strongly agree

## References

Bangor A, Kortum PT, Miller JT (2008) An empirical evaluation of the system usability scale. Int J Hum Comput Interact 24:574–594

Brennan RL (2001) Generalizability theory. Springer, New York

Cito (2006) TiaPlus. Test and item analysis. Cito, Arnhem

Cito (2014) The use of internet and the computer at home questionnaire. Dutch version retrieved from http://toetswijzer.kennisnet.nl/html/internetvaardigheid/vragenlijst.pdf

Clarke-Midura J, Dede C (2010) Assessment, technology, and change. J Res Technol Educ 42(3):309–328

Cronbach LJ, Linn RL, Brennan RL, Haertel EH (1997) Generalizability analysis for performance assessments of student achievement or school effectiveness. Educ Psychol Measur 57(3):373–399

De Klerk S (2012) An overview of innovative computer-based testing. In: Eggen TJHM, Veldkamp BP (eds) Psychometrics in practice at RCEC. RCEC, Enschede, pp 137–150

De Klerk S, Eggen TJHM, Veldkamp BP (2014) 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). CADMO 22(1):39–56

De Klerk S, Veldkamp BP, Eggen TJHM (2015) Psychometric analysis of the performance data of simulation-based assessments: a systematic review and a Bayesian network example. Comput Educ 85:23–34

De Klerk S, Veldkamp BP, Eggen TJHM (2018) A framework for designing and developing multimedia-based performance assessment in vocational education. Educ Tech Res Dev 66(1):147–171

Dekker J, Sanders PF (2008) Kwaliteit van beoordeling in de praktijk [Quality of rating during work placement]. Kenniscentrum handel, Ede

Field A (2009) Discovering statistics using SPSS, 3rd edn. SAGE Publications Inc., Thousand Oaks

Gulikers JTM (2006) Authenticity is in the eye of the beholder: Beliefs and perceptions of authentic assessment and the influence on student learning (Doctoral dissertation). Open University, Heerlen

Gulikers JTM, Bastiaens TJ, Kirschner PA (2004) A five-dimensional framework for authentic assessment. Educ Tech Res Dev 52(3):67–86

Judd W (2009) Gating items: definition, significance, and need for further study. Pract Assess Res Eval 14(9). Available online: http://pareonline.net/getvn.asp?v=14&n=9

Kane MT (2004) Certification testing as an illustration of argument-based validation. Measurement 2:135–170

Kane MT (2006) Validation. In: Brennan RL (ed) Educational measurement, 4th edn. American Council on Education and Praeger Publishers, Westport, pp 17–64

Kane MT (2013) Validating the interpretations and uses of test scores. J Educ Meas 50(1):1–73

Klopfer E, Osterweil S, Salen K (2009) Moving learning games forward. Education Arcade, Cambridge

Koenig AD, Lee JJ, Iseli MR, Wainess R (2010) A conceptual framework for assessing performance in games and simulation (CRESST Report 771). University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles

Levy R (2013) Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. Educ Assess 18(3):182–207

Levy R (2014) Dynamic Bayesian network modeling of game based diagnostic assessments (CRESST Report 837). University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles

Mislevy RJ, Oranje A, Bauer MI, Von Davier A, Hao J, Corrigan S, Hoffman E, DiCerbo KE, John M (2014) Psychometric considerations in game-based assessment. GlassLab Rep. Retrieved from http://www.instituteofplay.org/work/projects/glasslab-research/

Parshall CG, Spray JA, Kalohn J, Davey T (2002) Practical considerations in computer-based testing. Springer, New York

Quellmalz ES, Pellegrino JW (2009) Technology and testing. Science 323:75–79

Quellmalz ES, Timms MJ, Buckley B (2010) The promise of simulation-based science assessment: the Calipers Project. Int J Learn Technol 5(3):243–263

Roelofs EC, Straetmans GJJM (eds) (2006) Assessment in actie [Assessment in action]. Cito, Arnhem

Rupp AA, Nugent R, Nelson B (2012) Evidence-centered design for diagnostic assessment within digital learning environments: integrating modern psychometrics and educational data mining. J Educ Data Mining 4:1

Scalise K, Gifford B (2006) Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. J Technol Learn Assess 4(6). Retrieved [March 20, 2012] from http://www.jtla.org

Shavelson RJ, Baxter GP, Gao X (1993) Sampling variability of performance assessments. J Educ Meas 30(3):215–232

Shavelson RJ, Ruiz-Primo MA, Wiley E (1999) Note on sources of sample variability in science performance assessments. J Educ Meas 36(1):56–69

Shute VJ (2011) Stealth assessment in computer-based games to support learning. In: Tobias S, Fletcher JD (eds) Computer games and instruction. Information Age Publishing, Charlotte, pp 503–523

Sijtsma K (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika 74(1):107–120

Verhelst ND (2000) Estimating the reliability of a test from a single test administration. Measurement and Research Department Reports 98–2. National Institute for Educational Measurement, Arnhem

Williamson DM, Bejar II, Mislevy RJ (2006) Automated scoring of complex tasks in computer-based testing: an introduction. Lawrence Erlbaum, Mahwah, NJ

Wools S (2015) All about validity. an evaluation system for the quality of educational assessment. Published doctoral dissertation. https://doi.org/10.13140/RG.2.1.3201.3288

Wools S, Eggen TJHM, Sanders PF (2010) Evaluation of validity and validation by means of the argument-based approach. CADMO 18(1):63–82
Wools S, Eggen TJHM, Béguin AA (2016) Constructing validity arguments for test combinations. Stud Educ Eval 48:10–18

## Affiliations

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

[1] Department of Vocational Examination, eX:plain, Amersfoort, The Netherlands

[2] University of Twente, Enschede, The Netherlands

[3] Cito, Arnhem, The Netherlands