Computers in Human Behavior 60 (2016) 264–279

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian Network☆

Sebastiaan de Klerk [a, b, *], Theo J.H.M. Eggen [b, c], Bernard P. Veldkamp [b]

[a] eX:plain, The Netherlands
[b] University of Twente, The Netherlands
[c] Cito, The Netherlands

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form
22 January 2016
Accepted 17 February 2016
Available online xxx

Keywords:
Multimedia-based performance assessment
Scoring interactive student performance
Evidence identification
Evidence accumulation
Log file analysis
Bayesian network

ABSTRACT

Computer-based simulations are increasingly being used in educational assessment. In most cases, the simulation-based assessment (SBA) is used for formative assessment, which can be defined as *assessment for learning*, but as research on the topic continues to grow, possibilities for summative assessment, which can be defined as *assessment of learning*, are also emerging. The current study contributes to research on the latter category of assessment. In this article, we present a methodology for scoring the interactive and complex behavior of students in a specific type of SBA, namely, a Multimedia-based Performance Assessment (MBPA), which is used for a summative assessment purpose. The MBPA is used to assess the knowledge, skills, and abilities of *confined space guard* (CSG) students. A CSG supervises operations that are carried out in a confined space (e.g., a tank or silo). We address two specific challenges in this article: the *evidence identification challenge* (i.e., scoring interactive task performance), and the *evidence accumulation challenge* (i.e., accumulating scores in a psychometric model). Using expert ratings on the essence and difficulty of actions in the MBPA, we answer the first challenge by demonstrating that interactive task performance in MBPA can be scored. Furthermore, we answer the second challenge by recoding the expert ratings in *conditional probability tables* that can be used in a Bayesian Network (a psychometric model for reasoning under uncertainty and complexity). Finally, we validate and illustrate the presented methodology through the analysis of the response data of 57 confined space guard students who performed in the MBPA.

1. Introduction

The use of interactive computer-based simulations in educational assessment is growing together with research on the topic (see for example, Clarke-Midura & Dede, 2010; Mislevy et al., 2014; Quellmalz & Pellegrino, 2009; De Klerk, Veldkamp, & Eggen, 2015; Shute, 2011). In the research literature, different terms are being used for these simulations. In a recently published comprehensive research report, Mislevy et al. (2014) discuss so-called game-based assessments (GBA). Other researchers discuss simulation-based assessments (SBA) (e.g., Levy, 2013), technology-based assessments (TBA) (e.g., Bennett, Persky, Weiss, & Jenkins, 2007), or computer-based assessments (CBA) (e.g., Parshall, Spray, Kalohn, & Davey, 2002). In essence, TBA is the overarching term for the other types of assessments (CBA, GBA and SBA). At this point, researchers and practitioners regard TBA as any assessment in which technology is used to administer the assessment (Baker, Chung, & Delacruz, 2008) and CBA mostly as a traditional paper-based test that has been converted into a computer-based equivalent (De Klerk, 2012). One step down the ladder, SBA can be considered as a higher abstraction that encapsulates GBA. By definition, a computer game is always a simulation. This can be a cognitive process (Kerr & Chung, 2012) as well as a real-world environment (Iseli, Koenig, Lee, & Wainess, 2010). It is only the means through which this process is simulated that makes GBA a specific subgroup of assessments within SBA. Although no two games are the same (Schrader & McCreery, 2012), there are universal features of games

☆ This research was supported by eX:plain.
* Corresponding author. Department of Vocational Examination, eX:plain, Amersfoort, The Netherlands.
E-mail address: s.dklerk@explain.nl (S. de Klerk).

http://dx.doi.org/10.1016/j.chb.2016.02.071
0747-5632/

and GBAs (Mislevy et al., 2014; Prensky, 2001; Squire, 2003). For example, a game is characterized by playing which makes it entertaining to do, at least to the extent that it is more entertaining than performing a traditional assessment (Shute & Ventura, 2013). Furthermore, a game has an advancement rate, which means that fulfilling assignments in combination with reaching goals brings the player to a higher level in the game, and gives players the feeling of winning and losing. To some extent, a game also gives players a free virtual space to act in, which contrasts with other types of SBA that may significantly restrain student actions.

Although much of the research on these types of assessment focuses on using it for a formative purpose, there is also a growing body of research on using it in a summative assessment situation (e.g., Rupp, DiCerbo, 2012, Rupp, Nugent, 2012). Computer simulations have some interesting features that make them suitable for use in SBAs. For example, an SBA can provide an authentic assessment environment, especially compared with paper-based tests. In addition, skills that cannot be assessed using traditional assessments (e.g., how to act in dangerous situations) can be more realistically tested in an SBA. Furthermore, computers can be fully objective and standardized, which results in fair scoring and unvarying interaction with the student being assessed. This is particularly interesting in comparison with a Performance-based Assessment (PBA) with human raters. A PBA is used to assess the performance-oriented knowledge, skills, and abilities of students in a realistically simulated real-world environment. In this environment, students perform actions and operations as they would in the real work environment. Their performance is observed and rated by one or more raters on a rubric. The goal of the PBA is to determine whether the student has demonstrated sufficient skill to be certified.

SBA can be placed along a continuum of interactivity, immersion and freedom to act within the simulation. For example, on the right side of the continuum, some simulations allow a high degree of interaction between the student and the assessment, which means that the state of the simulation changes on the basis of what the student does in the simulation. The same holds for immersion and freedom to act: some simulations provide a full computer-based environment in which a student can roam freely as a virtual character. On the left side of the continuum, some simulations are (much) more restricted and offer a strongly simplified representation of the subject of simulation. Most recently, simulations are starting to move towards the right side of the continuum, and these simulations can be defined as GBAs.

In most cases, assessment designers are not interested in students' competencies, which comprise students' knowledge, skills, and abilities (KSAs), *within* the computer environment, but in a generalization of these KSAs *outside* the computer environment. The question then is: how do you score interactive student performance *within* the computer environment? And how can these scores be used to say something about KSAs *outside* the computer environment? In any case, capturing the raw data that students produce while performing the simulation (e.g., mouse clicks, time stamps, navigational path, etc.) is common practice today (Koenig, Lee, Iseli, & Wainess, 2010). On the other hand, finding meaningful relationships, patterns, and clusters in the performance data is still a difficult task. The first challenge therefore is the process of analyzing the performance data to identify the most meaningful elements, a process which is referred to as *evidence identification* (Levy, 2013). In our case, as we explain below, the identification of meaningful elements in student performance data logs was already part of the design phase of the MBPA. A second challenge is to combine, weigh, and aggregate these pieces of evidence in the student performance data to make informative inferences about performance outside the computer environment. This process is referred to as *evidence accumulation* (Levy, 2013). In the current article, we take up both the process of evidence identification and the process of accumulation. We have developed and empirically tested a specific type of SBA, which we have called Multimedia-based Performance Assessment (MBPA) because it relies heavily on multimedia (video and photo material) and is used to assess KSAs that are currently being tested in a PBA. The MBPA is used to test the KSAs of confined space guards. A confined space guard (CSG) supervises operations that are carried out in confined spaces (e.g., a silo or a tank). Students carry out a one-day training program and then have to pass a multiple-choice knowledge test and a PBA. We have tried to convert a sample of the tasks in the PBA into a multimedia-based equivalent.

A confined space guard has an important role. By guarding operations inside a confined space, the CSG literally is responsible for the lives of the workers inside the confined space. Confined spaces are potentially extremely dangerous, and multiple serious, and sometimes deadly, accidents happen on a yearly basis in the Netherlands. It is therefore important to assess the skills of the CSG to the highest standards of reliability and validity. This is currently not possible with the performance-based assessment because of reasons that will be explained below. It may be possible to assess these skills more reliably and validly using MBPA, but it is important then to know, how to *score* interactive student behavior in the MBPA, and how to *accumulate* those scores in an overall score, using a psychometric model.

The MBPA is therefore used to illustrate a methodology for applying interactive task performance scores from an MBPA in a psychometric model —the Bayesian network. The methodology presented in this article enables us to address both challenges discussed above. The central question in the study is: Can we develop a modern scoring methodology for applying students' MBPA performance data in a psychometric model to make valid inferences about performance outside the virtual assessment context? More specifically, we try to answer the following questions in this study. How can a CSG student's interactive task performance in an MBPA assessment be scored in such a way that it accurately represents the student's KSAs? This question relates to the evidence identification challenge. How can these scores be applied in a psychometric model? This question relates to the evidence accumulation challenge. To illustrate the methodology, we have used the data from a sample of 57 students, who all completed the MBPA.

2. Theoretical background

Studies on PBA in the 1990s have shown that it is relatively susceptible to measurement error, as compared to more standardized forms of assessment because of generalizability and reliability issues (Brennan, 2000; Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991; Shavelson, Baxter, & Gao, 1993). Furthermore, PBA performance is usually scored on a rubric by a human rater. Both rubrics (Shepherd & Mullane, 2008) and human raters have been found to negatively influence the reliability of the assessment (Dekker & Sanders, 2008). In fact, to reach acceptable levels of reliability across raters, it is advisable to have students perform multiple assessments (Wass, McGibbon, & Van der Vleuten, 2001), something that is often not possible because of the logistical inefficiency and the costs of PBA.

The environment of a computer-based simulation can to some extent solve the measurement issues associated with PBA. For example, a computer environment can be standardized so that the simulation always reacts in the same fashion to input (e.g., mouse strokes, answers or interactions) from the student. A good example is the use of standardized patients in computer-based simulation

tasks in the United States Medical Licensing Examination developed by the National Board of Medical Examiners (Margolis & Clauser, 2006). In addition, because it is possible to use a standardized rating scheme for students' performance in the simulation, the overall reliability can be increased, especially compared to PBAs (Wainess, Koenig, & Kerr, 2011). Finally, using MBPA, the representativeness of the KSAs in the assessment can be increased (as compared to PBA). This means that, compared to the PBA, it is possible to present a multitude of tasks and scenarios in the MBPA, which enables a stronger representation of the domain (Baker et al., 2008). Overall, there are some strong arguments to suggest that SBA, and in our case MBPA, can be used in an assessment program to make more reliable and valid statements about student KSAs.

One of the biggest challenges in using innovative computer-based simulations in an assessment context is scoring the performance of students. In other words, we have to identify, within all aspects of performance, which (combinations of) pieces can be defined as evidence of proficiency, and thus produce *observable variables* that are relevant to characterizing student performance (Mislevy, Steinberg, & Almond, 2003; Rupp, DiCerbo, 2012, Rupp, Nugent, 2012). This is called the evidence identification challenge. For example, in a multiple-choice question the scoring rule and corresponding observable variable (OV) can be very simple (1 for a correct answer, and 0 for an incorrect answer), whereas translating student performance in a complex and interactive computer environment into a meaningful OV is much more difficult.

A factor that makes things even more complex is the so-called *change state* of the assessment (Mislevy et al., 2014). The complexity of actions is high and dependencies across observations are often caused by the constantly changing state of the game (Rupp, Gushta, Mislevy, & Shaffer, 2010). That is, based on what a player has done before, the possible actions that can be taken in a future situation are changed. This makes it possible to build universal scoring schemes, thereby making it difficult to minimize measurement error. In addition, Rupp et al. (2010) argue that multiple layers of human judgment are involved in defining the meaningful OVs, whether or not the scoring is automated, and that the data from the simulation may be relatively distal to the desired interpretations. To summarize, scoring students' interactive task performance in an MBPA is not straightforward.

After evidence of student KSAs in the performance data have been identified and translated into OVs, the next challenge is to weigh and aggregate them into some sort of final score. This is the evidence accumulation challenge (Levy, 2013). The accumulation of evidence consists of synthesizing all OVs for the facilitation of the desired inferences about student KSAs. A psychometric model is then used to model the OVs as random variables that are dependent on the KSAs. One can say that the assumptions in the psychometric model are used to translate all pieces of data that were characterized as meaningful for a student's overall evaluation into an overall score. The psychometric model that is most frequently discussed with respect to SBA is the Bayesian network (BN) (Levy, 2014; Levy & Mislevy, 2004; Mislevy, Almond, Yan, & Steinberg, 2000; West et al., 2010). BNs (Pearl, 1988) provide a graphical structure in which conditional probability relationships between a (large) number of random variables are represented. Through probabilistic (Bayesian) inference algorithms, it is possible to make probabilistic statements about the state of certain latent variables in the network, given the state of other OVs. BNs have been around for quite some time and they have been applied in many fields (Neapolitan, 2003). For example, they have been applied in medicine, for medical decision making (Lucas, 2001); in artificial intelligence, for learning systems (Korb & Nicholson, 2010); and in ecology, for environmental modeling (Aguilera, Fernández, Fernández, Rumí, & Salmerón, 2011).

Shute, Ventura, Bauer, and Zapata-Rivera (2009) show how the application of a BN to the data of a serious game can be used to yield information about student characteristics in an educational assessment context. Shute et al. were interested in students' *creative problem solving* (CPS) ability and measured this ability through a quest in a commercial video game. Students could take multiple actions in the game to solve the quest, and all these actions were rated and resulted in *novelty* and *efficiency* scores for each student. Using Bayesian modeling software, these scores were then entered into the BN. A final judgment about CPS could be then be made via the conditional probabilities between the manifest variables (i.e., students' scores) and the latent variables (i.e., creativity and problem solving). This small example shows how researchers have used student performance in a virtual environment to measure cognitive ability that also exists in a context outside the virtual environment.

In the current article, we take up the evidence identification and evidence accumulation challenges discussed above. We present an MBPA, with which we aim to measure students' KSAs for being CSGs. Using different types of multimedia and interactive tasks, we virtually simulate a real-world environment in which the CSG students can fulfill assignments that CSGs perform in their vocation. The MBPA is used to illustrate a methodological structure for scoring the interactive task performance of students in the MBPA, based on the raw log file data that a representative sample of students produced. Secondly, using a modern psychometric model, the BN, we use the scores of student performance in the interactive tasks to synthesize performance into an overall CSG score. Next, we will first discuss the methodology of our research. The multimedia-based performance assessment will be discussed in detail, including screenshots. The data collection, on basis of 57 students' response data vectors and six expert ratings, will be explained. Furthermore, in the first part of the results section we will demonstrate how expert ratings can be used to making a scoring scheme for interactive actions inside the virtual assessment, and in the second part of the results section, we will show how these scores can be recoded into *conditional probabilities* and then used in a psychometric model – the Bayesian Network. Using a graphical structure of both observable and latent variables, this psychometric model can be used to reason under uncertainty. That is, we can make inferential statements about students' proficiency on several latent KSAs (i.e., communication, vigilance, procedural knowledge), on basis of their interactive performance inside the virtual assessment (the MBPA).

3. Method

3.1. Participants

The participants in the empirical study included 57 CSG students (1 female and 56 male). Participants ranged in age from 18 to 62 years, with a mean age of 43 years ($\sigma = 11.55$). Of the 57 participants, 41 had Dutch ethnicity and 41 students had only participated in education up to high school or lower vocational education. They were asked to participate in the MBPA after they had completed training. Participation was voluntary and students did not receive payment. All participants were recruited at two training locations. The 57 participants are assumed to be a representative sample of the general population for two reasons. First, all participants had just finished the CSG course. Secondly, the sample in the experiment reflects the general student population with respect to age, ethnicity and education.

Furthermore, we also asked six experts to rate all actions in the MBPA on both essence and difficulty, and we asked them to estimate the probability that a minimally competent student (a borderline student) would successfully complete the action. Three

experts were confined space guard trainers at two training locations in the Netherlands. The three other experts were members of the committee that define the curriculum for confined space guard training, and the performance-based assessment objectives.

3.2. Materials

3.2.1. Multimedia-based performance assessment (MBPA)

The MBPA is a flexible and interactive assessment that is accessible through the internet. The MBPA has been build according to a framework for designing and developing MBPA (De Klerk, Veldkamp, & Eggen, *submitted for publication*), which is presented in Fig. 1. In the MBPA, the authentic work setting and real-life equipment are simulated by images and video fragments. The different settings in the MBPA have been chosen and designed in collaboration with subject matter experts (SMEs). In addition, the design has been guided by the scenarios of the existing PBA. The MBPA can be seen as a virtual version of the real-life PBA.

A central feature in the MBPA is its so-called toolbox. Students can collect equipment and documentation in their toolbox (a box at the bottom right of the screen). Equipment is collected by clicking on the image. A student can then open the item (e.g., a work permit) by clicking on it in the toolbox. Some items, or combinations of items, open up assignments which students have to complete. For example, collecting a work permit and information from the operator gives students the c to check the work permit for

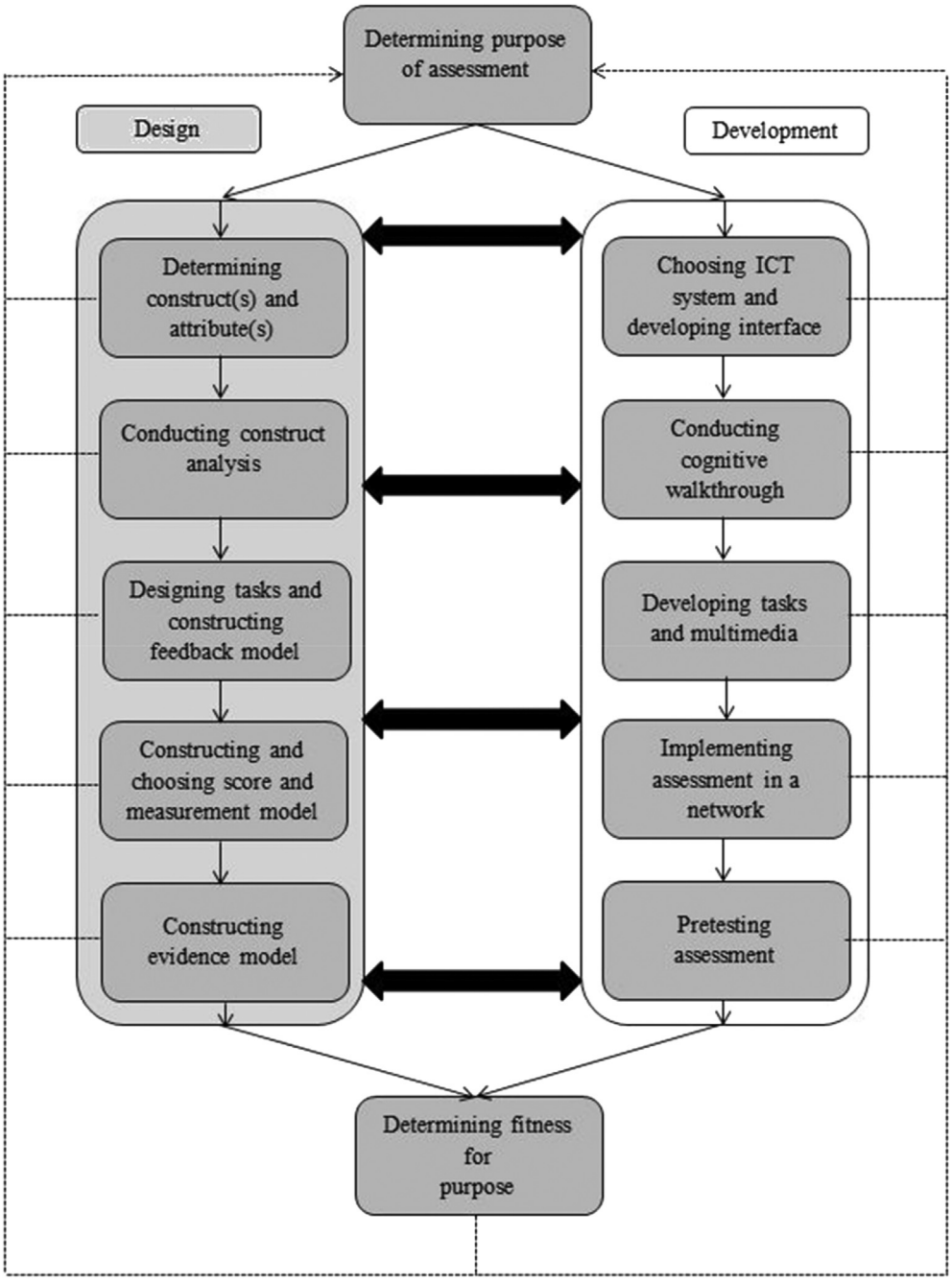

Fig. 1. A framework for the design and development of MBPA.

possible errors. The MBPA consists of four different settings. First, there is a practice setting in which the students can get familiar with the different functionalities of the assessment. Secondly, there is an office setting in which information has to be collected and processed. Thirdly, there is an outside setting at the confined space in which the students have to supervise operations that are carried out inside the confined space. Finally, there is a setting in which students have to react to a plant alarm.

The MBPA started with a welcoming screen and operating instructions. Use of the buttons was explained with enlarged pictures of the buttons. The goal of the assessment was also explained. Students could take as much time as they needed to read the instructions and get familiar with the different buttons in the assessment. They could then progress to the practice setting. In the practice setting, students were requested to carry out a number of small actions in order to get a feel for participating in the assessment.

Students were shown an image of a factory cafeteria in which they could click on several objects, which would open an assignment. The goal was to find money (i.e., click on a wallet), get coffee (i.e., click on the coffee machine), and then go to the operator's office (i.e., click on the door to leave the room). After the practice setting, students were told that the assessment proper would now begin.

In the next setting, the students were shown an image of an office. Again, several elements in the image were active and clicking on them could open an assignment. Students could click on the work permit, the operator, the walkie-talkie, the exit sign, and a worker. Clicking on one of these elements would add an icon to the toolbox on the right-hand side of the screen. The general rationale was that students should start by getting their work permit and walkie-talkie by clicking on these objects. They should then click on the operator to collect information about both items. Students were to go back to these items subsequently to process the information that the operator had given them. For example, the operator tells a student to review the work permit to ensure that it is free of errors. Students then need to click on the walkie-talkie to put it on the right channel for communication, as instructed by the operator. The worker, who is part of the image and can be clicked upon, can be regarded as a distractor because students do not have to be in contact with him during this process. Finally, after students have completed this procedure and have finished all assignments in this setting, they should click on the exit sign above the operator to leave the room. If a student clicks on the exit sign, a question pops up asking for confirmation that he or she really wants to leave this setting and go on to the next setting.

The third setting is an outside setting at the confined space. A student can click on two objects in this setting: the confined space itself and the wind direction indicator (shown in red and white, behind the confined space). If the student clicks on the wind direction indicator, then an assignment is presented in which the student has to answer which meeting point should be used, considering the wind direction, in the case of a gas alarm. Again, if the student clicks on the confined space, a question pops up asking the student if he or she wants to start the next assignment. In this assignment, students are shown two video fragments in which two workers enter into the confined space to carry out cleaning operations. The students can click on a red stop button at the top right corner of the screen when they see incorrect behavior on the part of the workers. As the workers make three mistakes in each fragment (e.g., removing their helmets, calling in the confined space, or taking illegal electrical equipment inside the confined space), students can obtain a maximum of six correct answers. The stop button disappears after students have clicked on it three times.

In the fourth and final setting, students are confronted with an emergency situation. They are told that there is a plant alarm and that they have to use the icons in the toolbox to take the correct and most efficient actions to get everyone safely to the meeting point. For example, workers must be moved out of the confined space before closing it with a “do not access” sign. The assessment automatically ends when students click on the meeting point icon and indicate that they want to go there. Two screenshot of the MBPA are shown in Fig. 2.

3.2.2. Student 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?” (N = 5)); 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?” (N = 3)); 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” (N = 7)). Reliability for computer experience and usability were $\alpha = .64$ and $\alpha = .75$, respectively. The questionnaire was based on a translated version of the System Usability Scale (Bangor, Kortum, & Miller, 2008) and a questionnaire on the use of Internet and the computer at home, developed by Cito (Cito, 2014). Thus, students' computer use and the usability of the MBPA can be classified as subscales of the questionnaire. A translated version of the questionnaire can be found in Appendix A.

3.2.3. 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 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. A number of 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 a number of 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 had to then 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.

The performance-based assessment rater used a rubric made up of four criteria ($\alpha = .84$) that focused on students' proficiency in communication, proactive behavior, environmental awareness, and procedural efficiency. Raters were asked to rate 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 on the rubric. A

Fig. 2. Interface of the MBPA. The toolbox with a few tools is shown at the right side of the screen. In the first screenshot, the stop button, at the top right is shown for the intervention questions. In the second screenshot the office setting can be seen. Below is a bar in which instructions for the student are Shown.

translated version of the rubric can be found in Appendix B.

3.2.4. Expert rating scheme

As said, experts have been requested to rate all actions in the MBPA on the essence of the action and the difficulty of the action. These ratings have been used to develop a scoring model. Specifically, for different settings in the MBPA, the raters were asked to rank order all actions in that setting from most essential to least essential and from most difficult to least difficult. Furthermore, they were also asked to rate, on a 5-point likert scale ranging from highly improbable to highly probable, the probability that a borderline student (a minimally competent student) would successfully perform each of the actions in the MBPA. This can be regarded as the Angoff method for standard setting (Downing & Haladyna, 2006). The expert rating scheme can be found in Appendix C.

3.3. Procedure and data collection

The MBPA was administered in the computer room of each training location. Students had just completed the training program when they were requested to do the MBPA. They were told that the performance on the MBPA would not be used for an overall pass/fail decision. The students were seated behind a laptop. All assessments were administered under the supervision of the first author. Students logged in with a personal login and password on the MBPA website. No time limit was imposed on students either for separate assignments or for the assessment as a whole. Student questions were answered by the supervisor, but only if the questions were related to the assessments' functioning. After students had completed the MBPA, they left the computer room to carry out their PBA in the reconstructed work setting. All student response data that is analyzed in the results section was collected during the administration of the 57 MBPAs on the training locations, after students had performed their PBA.

The six experts, whom were asked to rank order the actions in the MBPA, also performed the MBPA to get to know the assessment. They could perform the MBPA at home or at work. After that, they were sent a rank ordering form, via e-mail, on which they could rank all actions in the different settings of the MBPA from most essential to least essential, and from most difficult to least difficult. Finally, as said, for each action in the MPBA, they estimated the probability that a minimally competent student would successfully execute the action.

4. Results

Because of the technical nature of the Results section, we will first discuss how the variables from the method section synthesize into a *methodology* for scoring interactive task performance and accumulating these scores in a psychometric model. As said, two challenges for MBPA have been defined in literature. First, the evidence identification challenge. That is, how can we (best) score interactive task performance in a virtual assessment? And secondly, the evidence accumulation challenge. That is, how can we then combine these scores in a psychometric model? In the first section of the Results section, we will therefore discuss all 22 actions in the MBPA, and how these have been rated on both essence and difficulty by the six experts. These ratings are then transformed into a *score* for all 22 actions in the MBPA. In the second section of the Results section, we will then translate these scores into *conditional probabilities*, which express the relationship between 22 observable variables (has the student performed the action, yes or no), and several latent variables that relate to students knowledge, skills, and abilities. These processes together *form* the methodology, which is illustrated through the 57 response data vectors produced by students performing in the MBPA.

4.1. Scoring interactive task performance in the multimedia-based performance assessment—evidence identification challenge

As mentioned earlier, the first challenge was to score students' performance in the interactive tasks of the MBPA. This is the process of *evidence identification*; namely, to define what can be considered as evidence of the CSG KSAs of students within the flexible and interactive environment of the MBPA. For example, we can choose to look only at the number of correct actions that the student has performed, which is called *product data*. But we can also look at efficiency (i.e., the number of actions needed), time, or order, which is called *process data*. The process of evidence identification ultimately results in a set of OVs that can be used in a psychometric model (*evidence accumulation*).

Before getting to the actual point of scoring student performance, we first discuss all the actions or combinations of actions that a student could take during the three settings in the MBPA. As stated earlier, the MBPA was designed in consultation with SMEs. Therefore, in this case, the evidence identification for the MBPA is a direct consequence of the design decisions that were made earlier. This means that the MBPA has been designed to disclose the extent to which students can correctly perform the CSG actions that are tested in the PBA. The (correct) actions that a student can perform in the MBPA are listed below.

In the office setting, there are eight correct actions that a student can complete before advancing to the next setting:

A.1 Collect the work permit (click on the work permit shown on the table).
B.1 Ask for information about the work permit (after action A1, click on the operator, and then ask the operator for an explanation, through a multiple selection question).
C.1 Find an unregistered finish time on the work permit (after actions A1 and B1, click on the work permit again, to answer, the question whether there are no mistakes in the work permit, through a yes/no question).
D.1 Find that a signature is missing from the work permit (after actions A1, B1, and C1, again answer the question whether there are no mistakes in the work permit, through a yes/no question).
E.1 Collect the walkie-talkie (click on the walkie-talkie on the table).
F.1 Ask for the channel for communication with the operator (after action E1, click on the operator, and then ask the operator for the channel for communication, through a multiple selection question).
G.1 Set the walkie-talkie to the correct channel (after actions E1 and F1, click on the walkie-talkie again, to select the correct channel for communication, through a multiple-choice question).
H.1 Ask for further documentation (click on the operator, and then ask the operator for further documentation, through a multiple selection question).

In the outside setting, there are also eight correct actions that a student can complete before advancing to the next setting:

A.2 Check the wind direction (click on the wind direction indicator).
B.2 Select the correct meeting point, considering the wind direction, in the case of a gas alarm (after A2, indicate the correct meeting point, through a multiple-choice question).
C.2 Stop work when worker removes signaling cord (press stop button when seeing this behavior in a video fragment).
D.2 Stop work when worker removes helmet in confined space (press stop button when seeing this behavior in a video fragment).
E.2 Stop work when worker brings electrical equipment into the confined space (press stop button when seeing this behavior in a video fragment).
F.2 Stop work when worker removes gas meter from confined space (press stop button when seeing this behavior in a video fragment).
G.2 Stop work when worker removes safety gloves when working inside the confined space (press stop button when seeing this behavior in a video fragment).
H.2 Stop work when worker is using cellphone inside the confined space (press stop button when seeing this behavior in a video fragment).

Finally, in the alarm setting, there are six actions that a student can complete correctly, before ending the assessment:

A.3 Warn workers inside the confined space in the case of an alarm (in the toolbox, click on the signaling cord icon or the workers icon).
B.3 Sign off workers from the person registration list (PRL) when they leave the confined space in the case of an alarm (in the toolbox, click on the PRL icon).
C.3 Secure tools around the confined space in the case of an alarm (in the toolbox, click on the tools icon).
D.3 Attach the “do not enter” sign to the confined space in the case of an alarm (in the toolbox, click on the no access icon).
E.3 Contact the operator through the walkie-talkie in the case of an alarm (in the toolbox, click on the walkie-talkie icon or the operator icon).
F.3 Take workers to the meeting point in the case of an alarm (in the toolbox, click on the meeting point icon).

As can be seen, some actions are *nested* within other actions. That is, you can only (correctly) perform such an action if the higher order action has already been correctly performed. This is what has already been referred to as the *change state* of the assessment (Mislevy et al., 2014). This makes it more difficult to build a suitable scoring scheme for the MBPA. Rupp et al. (2010) have argued that multiple layers of human judgment are involved in defining the OVs. We agree with this statement. Accordingly, to identify the scores with their accompanying actions, we consulted six CSG SMEs to get ratings on both the difficulty and essence of all (combinations of) actions in each setting. Three raters were CSG training instructors at the training locations of our study. The other three raters were members of a commission of experts who define the training curriculum and the content of the PBA. We asked them to rank the actions per setting from least essential to most essential, and from least difficult to most difficult. In their ratings, the experts addressed the fact that some actions are nested (i.e., they can only be performed if other actions have already been performed), which may make them more difficult for students. This resulted in 264 ratings (22 actions $\times$ 2 ratings $\times$ 6 experts). To check whether their ratings were useful, we first calculated interrater reliability, interrater agreement, and Cronbach's alpha. For these indices, see Table 1. We used the intraclass correlation coefficient (ICC), which can be used, in contrast to Cohen's Kappa, to calculate interrater reliability and agreement where more than two raters are used (Shrout & Fleiss, 1979). Specifically, we used the two-way random effects model because the raters and actions in the MBPA are a sample from a larger population of raters and actions.

The interrater reliability ICCs and the alphas show that the reliability of the ratings can be considered fair (Cicchetti & Sparrow, 1981; Fleiss, 1981). The interrater agreement ICCs show that the raters' absolute agreement is slightly higher than their shared reliability. These indices give enough support to build a scoring scheme on the basis of the experts' ratings. Because the experts were compelled to *rank* each action, instead of *rate* each action separately, the variance in ratings becomes higher; this diminishes the interrater indices.

The average ratings for both essence and difficulty for all the actions in the MBPA are shown in Table 2. The lower the ranking (i.e., the higher the number in the third and fourth column), the less essential or difficult the action was considered by the experts. For example, a rating of 1, would mean that the action is most difficult or most essential, whereas a rating of 8 would mean that the action is least difficult or least essential.

These average ratings were then used to calculate total scores for each of the actions in the MBPA. In Table 2, two total scores are presented. In essence, we tested two models. In the first model, C^{SCORE1}, the most difficult and least essential actions are considered to be most informative because we assume that the best students correctly perform both the most difficult and the least essential actions in the MBPA. The C^{SCORE1} is therefore calculated by first recoding the essence ratings, so that a high number corresponds with high essence and a low number with low essence (and, of course, also corresponds with the difficulty ratings). The essence and difficulty ratings are then added. For example, for the first action (A1), the essence rating is recoded into a 3 (9 minus 6; 9 because the ratings start at 1), and then added to the difficulty rating (7), which makes 10. The lower the number in the fifth column of Table 2, the more informative the action is considered to be for student KSAs. On the other hand, if the intended purpose of the MBPA is to accredit students, which it is, and not to identify the best students, a second model, C^{SCORE2}, in which the most difficult and most essential actions in the MBPA are considered to be most informative, may be more suitable. That means that for C^{SCORE2} the original ratings can be added, but we need to emphasize that a high score then corresponds with low informative value.

We theorize that the interaction between difficulty and essence is most informative regarding student CSG KSAs. This theory is based on Ebel's standard setting procedure (Ebel & Frisbie, 1991). For standard setting purposes, Ebel suggested that raters judge test items on three difficulty levels (easy, medium, and hard) and four relevance categories (essential, important, acceptable, and questionable). The rationale is that a borderline student has a higher probability of not only answering the easier items correctly, but also the most essential items (especially the items which are both easy and essential). We have adopted this method to an extent by keeping difficulty constant in both models and only changing the influence of essence. In the second part of the results section we will further investigate which model is most informative regarding student KSAs.

To determine the cutoff score, which will also be discussed in the second part of the results section, we asked the raters to estimate the probability that a minimally competent student (also called a borderline student) would successfully complete each of the actions in the different settings of the MBPA (e.g., “What is the probability that a minimally competent student would set the walkie-talkie to the correct channel?”). Raters could then rate each action on a 5-point Likert scale ranging from highly improbable to highly probable. In Table 3, we show interrater reliability, interrater agreement, and Cronbach's alpha. This method for determining a cutoff score is a slightly adjusted version of the Angoff method for standard settings (Angoff, 1971; Cizek, 2006).

Overall, the reliability and agreement indices indicate poor to

Table 1
Interrater reliability and interrater agreement (ICC's) and Cronbach's alpha for essence and difficulty of the actions in the three settings.

Setting	ESS/DIF	Interrater Reliability	Interrater Agreement	α
Office	ESS	0.31	0.34	0.73
	DIF	0.38	0.42	0.79
Outside	ESS	0.34	0.37	0.75
	DIF	0.20	0.22	0.60
Alarm	ESS	0.35	0.39	0.76
	DIF	0.36	0.41	0.77

Table 2
Experts' average ratings on essence and difficulty for the actions in the MBPA.

Setting	Action	$M^{ESS}(\sigma)$	$M^{DIF}(\sigma)$	C^{SCORE1}	C^{SCORE2}
Office	A1	6.00 (3.16)	7.00 (1.27)	10.00	13.00
	B1	7.00 (0.63)	4.17 (2.48)	6.17	11.17
	C1	2.50 (1.64)	3.00 (2.28)	9.50	5.50
	D1	4.00 (2.53)	2.50 (1.05)	7.50	6.50
	E1	3.83 (2.31)	6.83 (0.98)	12.00	10.66
	F1	4.00 (0.89)	4.17 (2.04)	9.17	8.17
	G1	2.83 (0.98)	5.17 (2.79)	11.34	8.00
	H1	5.83 (1.47)	3.17 (2.04)	6.34	9.00
Outside	A2	6.17 (2.79)	3.67 (3.01)	6.50	9.84
	B2	6.83 (1.47)	2.17 (2.40)	4.34	9.00
Outside	C2	5.17 (2.04)	3.67 (0.81)	7.50	8.84
	D2	2.33 (1.51)	5.00 (1.55)	11.67	7.33
	E2	4.67 (1.51)	4.67 (1.86)	9.00	9.34
	F2	5.00 (2.10)	5.67 (2.25)	9.67	10.67
	G2	3.17 (1.72)	4.33 (1.97)	10.16	7.50
	H2	2.67 (1.37)	6.83 (1.84)	13.16	9.50
Alarm	A3	6.00 (0.00)	5.33 (1.21)	6.33	11.33
	B3	2.67 (1.63)	3.33 (0.82)	7.66	6.00
	C3	2.83 (1.47)	1.50 (1.23)	5.67	4.33
	D3	3.17 (0.75)	3.17 (1.47)	7.00	6.34
	E3	2.83 (1.72)	3.33 (1.97)	7.50	6.16
	F3	3.50 (1.76)	4.33 (1.21)	7.83	7.83

Table 3
Interrater reliability and interrater agreement (ICC's) and Cronbach's alpha for Ratings on the Probability that a Minimally Competent Student Would Successfully Complete the Action.

Setting	Interrater Reliability	Interrater Agreement	α
Office	0.26	0.24	0.68
Outside	0.29	0.31	0.71
Alarm	0.26	0.26	0.67

fair agreement among the experts (Cicchetti & Sparrow, 1981; Fleiss, 1981). In Table 4, the mean ratings are presented. As can be expected, the rank order of the actions in Table 4 resembles the rank orderings in Table 2. However, the orderings are not identical. The reason is that the probability ratings in Table 4 take into account the fact that borderline students have a rather high probability of correctly performing difficult or unessential actions, when, for example, there is a lot of emphasis on that particular action during training.

4.2. Application of a Bayesian Network (BN) on student scores—evidence accumulation challenge

The ultimate goal of the MBPA is to make an informed and valid decision, based on student performance in the MBPA, about a student's overall KSAs to work as a minimally competent CSG. We therefore need to synthesize and aggregate student scores (i.e., the correct actions they took during the MBPA) to reach an overall judgment about their level of proficiency. The process has now changed from evidence identification to evidence accumulation. In general, a psychometric model is used to weigh and aggregate all pieces of evidence into a final score (Mislevy et al., 2014; Rupp, Nugent, & Nelson, 2012). The BN presents a structure of reasoning on which a psychometric model is imposed (Levy & Mislevy, 2004). A BN is composed of one or more OVs that inform the state of one or more latent variables, which are also called student model variables (SMVs). The BN is therefore helpful for handling uncertainty by using probabilistic inferences to update and improve the belief values regarding the latent variables. As Shute (2011) formulates: “The inductive and deductive reasoning capabilities of Bayesian nets support 'what-if' scenarios by activating and observing evidence that describes a particular case or situation, and then propagating that information through the network using the internal probability distributions that govern the behavior of the Bayes net. (p. 511)”.

Bayesian Networks have been used in assessment research and practice for quite some time now. In educational measurement, we face two problems when we have to reason with evidence: uncertainty and complexity (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015). Uncertainty, because a correct answer or a correct execution of an action in the assessment does not one on one relate to an exact level of proficiency. For example, less proficient students may sometimes correctly perform difficult tasks. We therefore have to administer a representative number of tasks or questions, and then reason from uncertainty to what extent the student has shown his or her proficiency (Almond et al., 2015). Furthermore, a substantial number of observable and latent variables (the constructs to be measured) are often underlying assessment, which makes it complex to model responses and overall scores (Mislevy et al., 2014). Fortunately, the characteristics of the Bayesian network make it an excellent tool to reason under uncertainty, through specified *conditional probability tables*, and it can be used to model complex hierarchical structures in which multiple observable variables are dependent on multiple latent variables. After updating a Bayesian network with a vector of student responses, the probability that a student is sufficiently proficient in ore more constructs is defined. In that way, statements about students' knowledge, skills, and abilities can be made.

In this case, the actions that a student has performed during the MBPA are defined as the OVs. As stated, the OVs inform the state of SMVs. As a result, in collaboration with the SMEs, we defined several SMVs as latent variables that are informed by the OVs. There are three lower-level SMVs -*information communication*, *vigilance*, and *following procedures*. There is one upper-level SMV, which we have defined as *overall confined space guard proficiency*. The OVs and SMVs are presented in a simple second-order measurement model, which means that the lower-level SMVs are modeled as dependent on a second-order latent variable, which comprises the upper-level SMVs. It is a factorially simple model because each of the OVs is only dependent on one lower-level SMV. The graphical structure of the measurement model is depicted in Fig. 3.

As can be seen, the model consists of one upper-level variable, θ_o, which is the overall proficiency of students as a CSG. This latent variable is then translated into three lower-level SMVs, θ_c, θ_v, and θ_p. The subscripts of the latent SMVs are abbreviations of overall proficiency, communication, vigilance, and procedures. The OVs can be found in the lowest level of the model. To simplify the model, we have only represented 9 OVs in Fig. 3, whereas there are, of course, more in the MBPA (22 in total). The OVs (A1–H1) in the office setting of the MBPA are dependent on the first SMV—communication. The OVs (A2–H2) in the second setting (outside) are dependent on the second SMV—vigilance. And the OVs (A3–F3) in the third setting (alarm) are dependent on the third SMV—procedures. This structure is the same structure that is represented in the BN.

We used the GeNIe software package to specify and estimate this model in a Bayes net (Drudzel, 2005). We first built the structure by specifying the score variables (i.e., correct or incorrect performance of a specific action) as OVs and the lower- and upper-level SMVs as latent variables. We then specified the conditional dependencies between the OVs and the lower-level SMVs, and between the lower-level SMVs and the upper-level SMV. These

Table 4
Experts' average probability ratings that a minimally competent student would successfully complete the action (from 1-highly improbable to 5-highly probable).

Setting	Action	$M^{PROB}(\sigma)$
Office	A1	4.50 (0.84)
	B1	3.17 (0.75)
	C1	2.83 (1.17)
	D1	3.17 (1.17)
	E1	4.17 (0.98)
	F1	2.83 (0.75)
	G1	3.17 (1.17)
	H1	2.67 (1.51)
	Avg. office	3.31 (0.62)
Outside	A2	2.67 (1.51)
	B2	3.50 (0.84)
	C2	2.83 (1.17)
	D2	2.83 (0.75)
	E2	2.83 (0.75)
	F2	2.67 (0.82)
	G2	2.67 (1.37)
	H2	3.33 (1.51)
	Avg. outside	2.92 (0.30)
Alarm	A3	4.33 (1.21)
	B3	2.67 (1.03)
	C3	2.67 (1.37)
Alarm	D3	2.17 (0.75)
	E3	3.00 (1.27)
	F3	3.50 (0.55)
	Avg. alarm	3.06 (0.70)

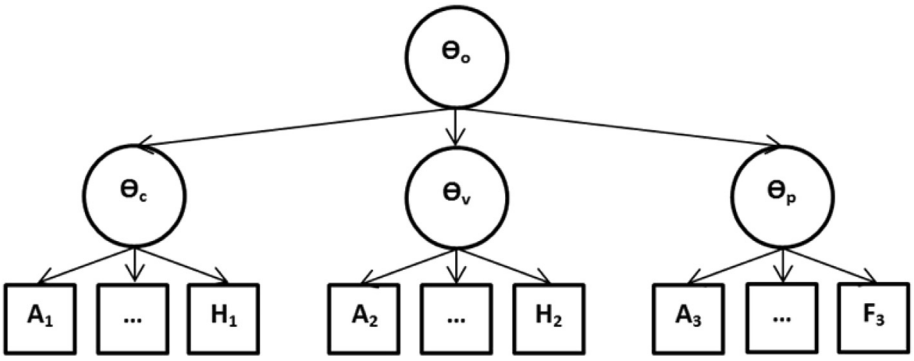

Fig. 3. Graphical representation of the simple second-order measurement model used for the MBPA in which θ_c corresponds with the office setting, θ_v corresponds with the outside setting, θ_p corresponds with the alarm setting and θ_o corresponds with the overall MBPA.

conditional dependencies are represented by the arrows in the network, which are called *arcs*.

After the model was built in GeNIe, we had to define the *conditional probability tables* (CPTs). The CPTs are specified for each node, given its *parents*, in the network. The node that influences another node through an arc is called a parent. For example, θ_c is a parent node of A1, and A1 is a *child* node of θ_c. As θ_o has no parent node in our network, an unconditional probability table is used. In the CPT, the conditional probabilities of a joint distribution of two or more variables are defined. In this case, the conditional probabilities are between the OVs and the lower-level SMVs, and in turn between the lower-level SMVs and the upper-level SMV. In fact, the CPTs define how the marginal probabilities of the different states of the parent node change as information about the children nodes is added to the BN. That is, students can perform action A1 correctly or incorrectly. In the first case, the state of A1 is changed to 1; in the latter case, the state of A1 is changed to 0. Because action A1 has a joint probability distribution with θ_c, the state of θ_c will change on the basis of what the student has done in the MBPA, and subsequently also θ_o of course. All the lower-level SMVs have two states: sufficient and insufficient. When the value (0 or 1) of the OVs has been entered in the Bayes net, the value of sufficient/insufficient changes accordingly, which then indicates whether a student is more or less proficient in that particular attribute. The conditional probabilities define how strong the relationship is between the variables. For example, if correct performance of action A1 may be more indicative of sufficient ability in θ_c than action B1, then the CPT of A1 and θ_c, will look different to the CPT of B1 and θ_c. The CPTs can be defined from data (e.g., IRT parameters from a pretest) or from expert input. For our Bayes net, we used the experts' evaluation of the actions in the MBPA to define the CPTs.

As noted earlier, based on a theory provided by Ebel (Ebel & Frisbie, 1991), we investigated two models. In the first model, we identified the most difficult and least essential actions in the MBPA as most informative of student KSAs. In contrast, in the second model, the most difficult and most essential actions in the MBPA are seen as the most informative of the KSAs. This is translated in the CPTs of the nodes in two Bayes nets: one for each interaction (i.e., most difficult × least essential and most difficult × most essential). The ratings of the experts on difficulty and essence can be seen in Table 2. The combined scores are reflected in the last columns. Of course, these scores cannot be entered as conditional probabilities in the joint distributions of OVs and SMVs in the BN. The scores therefore have to be converted into probabilities. We therefore calculated the z-scores, for each action and per setting, to standardize the values with a mean of 0 and a standard deviation of 1. Then, for each z-score, the one-sided percentile was calculated, which resulted in a percentage for each action. This percentage will later be expressed in the conditional probability tables (CPTs) of the Bayes net. These indices can be found in Table 5 for Model 1 (most difficult × least essential), and in Table 6 for Model 2 (most difficult × most essential).

Finally, we were able to record the conditional probabilities in the CPTs for the joint probability distribution of each OV and SMV. The base state of each OV in the Bayes net is a 50/50 distribution, which means that 0 (incorrect) or 1 (correct) both provide the same amount of evidence regarding proficiency in the SMV. We used the experts' input to change this distribution. For example, for action A1 and G1, using the percentiles in Table 5, we calculated the CPTs, which can be seen in Table 7.

The difference between both models is demonstrated in Table 7. The experts' ratings indicated that action A1 (i.e., collecting the work permit by clicking on the work permit in the image) is not very essential, and very easy. Furthermore, the experts have rated action G1 (i.e., setting the walkie-talkie on the right channel after actions E1 and F1) as considerably more essential and a bit more difficult than action A1. The difference can be seen when we look at the probabilities in Model 1 and Model 2. In Model 1, correctly performing action A1 in the MBPA considerably increases the probability that a student is in the sufficient category for θ_c, whereas incorrect performance considerably decreases the

Table 5
Model 1 C^{SCORE1}, Z-score, and One-sided Percentile for each Action in the MBPA.

Setting	Action	C^{SCORE1}	*z-score*	Percentile
Office	A1	10.00	−0.49048	31.19
	B1	6.17	1.3927	91.82
	C1	9.50	−0.24462	40.34
	D1	7.50	0.73879	77.00
	E1	12.00	−1.47389	7.03
	F1	9.17	−0.08236	46.72
	G1	11.34	−1.14936	12.52
	H1	6.34	1.30916	90.48
Outside	A2	6.50	0.94118	82.67
	B2	4.34	1.75437	96.03
	C2	7.50	0.56471	71.39
	D2	11.67	−1.00518	15.74
	E2	9.00	0	50
	F2	9.67	−0.25224	40.04
	G2	10.16	−0.43671	33.12
	H2	13.16	−1.56613	5.87
Alarm	A3	6.33	0.09444	53.76
	B3	7.66	−0.65173	25.73
	C3	5.67	2.14783	98.41
	D3	7.00	−0.28145	38.92
	E3	7.50	−0.56197	28.71
	F3	7.83	−0.74711	22.75

Table 6
Model 2 C^{SCORE2}, Z-score, and One-sided Percentile for each Action in the MBPA.

Setting	Action	C^{SCORE2}	*z-score*	Percentile
Office	A1	13.00	−1.60177	5.46
	B1	11.17	−0.86896	19.24
	C1	5.50	1.40155	91.95
	D1	6.50	1.00111	84.16
	E1	10.66	−0.66473	25.31
	F1	8.17	0.33237	63.02
	G1	8.00	0.40044	65.56
	H1	9.00	0.00000	50.00
Outside	A2	9.84	0.74172	22.92
	B2	9.00	0.00221	50.09
	C2	8.84	0.14392	55.72
	D2	7.33	1.48123	93.07
	E2	9.34	−0.29890	38.25
	F2	10.67	−1.47680	6.99
	G2	7.50	1.33067	90.84
	H2	9.50	−0.44060	32.98
Alarm	A3	11.33	−1.80775	3.53
	B3	6.00	0.41664	66.15
	C3	4.33	1.11359	86.73
	D3	6.34	0.27475	60.82
	E3	6.16	0.34987	63.68
	F3	7.83	−0.34708	36.43

Table 7
Conditional probability tables for action A1, G1 and SMV θ_c in model 1 and model 2.

Model 1	θ_c	
Action A1	Sufficient	Insufficient
Zero	0.3440	0.6560
One	0.6560	0.3440
Action G1	Sufficient	Insufficient
Zero	0.4374	0.5626
One	0.5626	0.4374
Model 2	θ_c	
Action A1	Sufficient	Insufficient
Zero	0.4727	0.5273
One	0.5273	0.4727
Action G1	Sufficient	Insufficient
Zero	0.1722	0.8278
One	0.8278	0.1722

probability that the student is in the sufficient category for θ_c. That is, action A1 is a rather informative OV regarding student proficiency in information communication (SMV). Action G1, for example, is less informative, correctly performing this action only slightly increases the probability of being in the sufficient category, whereas incorrect performance on slightly decreases the probability of being in the sufficient category for θ_c. Because the most essential actions have a stronger influence in Model 2, correctly performing A1 increases the probability of being in the sufficient category less than in Model 1 (because action A1 is not very essential). Yet, correctly performing more essential actions, like G1, have more influence on a students' probability of belonging to the sufficient category in Model 2 than in Model 1.

These probabilities have been calculated by multiplying the base state with the percentile score for each action (e.g., for action A1 50*1.0546). We have also produced an influence graph (see Figs. 4 and 5) in GeNIe, which shows how strongly the OVs influence the SMVs, and graphically demonstrates the difference between Model 1 and Model 2. The thickness of the lines in the figures indicates how strong the relationship between the nodes is. In Fig. 4, for example, the line between θ_c and A1 is thicker than in Fig. 5. This shows that correct performance of action A1 in Model 1 influences the joint probability distribution of θ_c and underlying OVs more strongly than in Model 2.

After all CPTs for both networks have been completed, it is possible to enter student data as evidence in the network. Of course, the student data are the OVs, which is a vector of 22 zeros and ones. The data for each student has to be entered individually. After evidence has been entered, the network can be updated, which means that all conditional probability distributions between the variables are calculated and the states of the SMVs are updated. You can see in Fig. 5 that the OV question marks (see Fig. 3) have been changed with gray grounding symbols, which means that beliefs have been updated (i.e., a students' response pattern has been entered). We used the default algorithm in GeNIE, which is the clustering algorithm (also called the junction tree algorithm). The clustering algorithm was first proposed by Lauritzen and Spiegelhalter (1988) and later improved by Jensen, Lauritzen, and Olesen (1990) and Dawid (1992). This algorithm works in two phases. First, the directed graph is compiled into a junction tree. Secondly, the probabilities are updated in the junction tree.

The clustering algorithm is the most basic and most widely-used algorithm for BNs (Almond et al., 2015). The result is the probability that a student, based on his or her performance in the MBPA, belongs to the sufficient or insufficient category for that particular SMV.

Next we used the experts' probability ratings for successful completion of the action by a borderline student (see Table 4) to determine whether or not a student is sufficient in a particular SMV. This can be regarded as a form of standard setting. In essence, the Bayes net only provides an estimate of the marginal probability that a student belongs to a particular category or not (e.g., the probability that student X is sufficiently proficient in information communication for CSGs). The next step, then, is to determine what level of probability is acceptable. The CSG assessment has a credentialing purpose so we are most interested in the cutoff point between insufficient and sufficient performance. In the office setting, the average expert rating for the probability that a minimally competent student would successfully complete the actions was 3.31 out of 5, or 66%. The probabilities were slightly lower in the outside setting and the alarm setting, 58% (2.91) and 62% (3.06), respectively. We assume that the most proficient students have the highest probability of belonging to the sufficient category (e.g., close to 1), whereas the least proficient students have the lowest probability of belonging to the sufficient category (e.g., close to 0). The probability that a minimally competent student belongs to the sufficient category for the information communication SMV is 0.66 (according to the experts' input). Therefore, the cutoff point is at 0.66. If the value of θ_c is 0.66 or higher, then we consider the student to be sufficient in that SMV ($\theta_v \geq 0.58$, $\theta_p \geq .62$). We do not use the difficulty and essence ratings to define the cutoff scores as these have already been translated in the CPTs. In addition, although the prior probability in the Bayes net for the SMVs is 0.5 (i.e., we do not know whether a student will be sufficient or not), the cutoff score is not calculated from 0.5 but from 0, because that is the lowest probability possible (i.e., when al actions are incorrectly performed or not performed at all).

Finally, updating the joint distributions between the lower-level SMVs and the upper-level SMV produces an overall evaluation of students' CSG proficiency based on their performance in the MBPA. Again, CPTs are defined for each of the lower-level SMVs and the upper-level SMV. These CPTs are filled with the average probability of the actions that share a distribution with that particular SMV (see Table 8). For example, for θ_c in the table below, the conditional probabilities for actions A1 to H1 are summed, and then divided by eight. All three lower-level SMVs have approximately the same shared distribution with the upper-level SMV. The arcs in Figs. 4 and 5 between the lower-level SMVs and the upper-level SMV are therefore equally thick.

As the whole network has now been defined, we can enter the

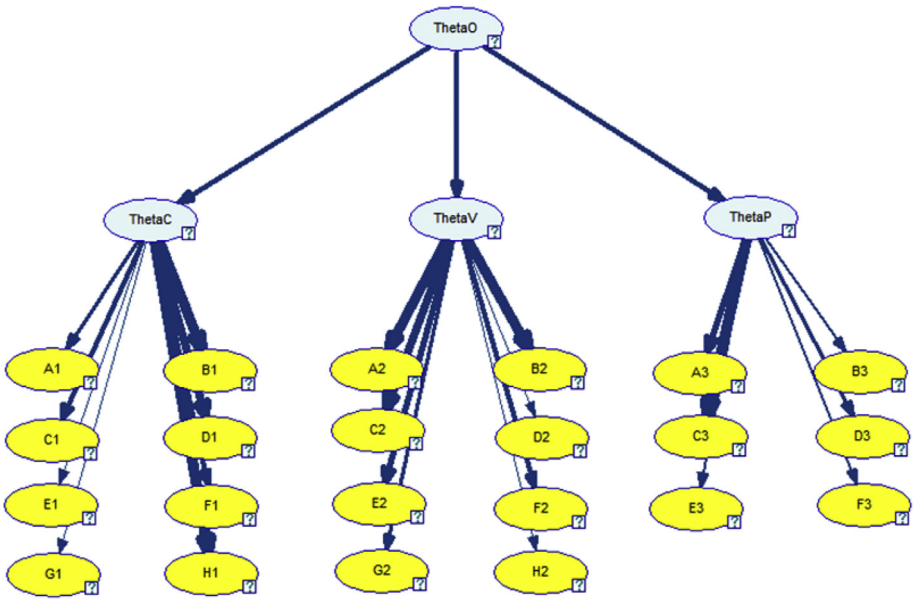

Fig. 4. Influence diagram of the model 1 Bayesian network for the confined space guard MBPA

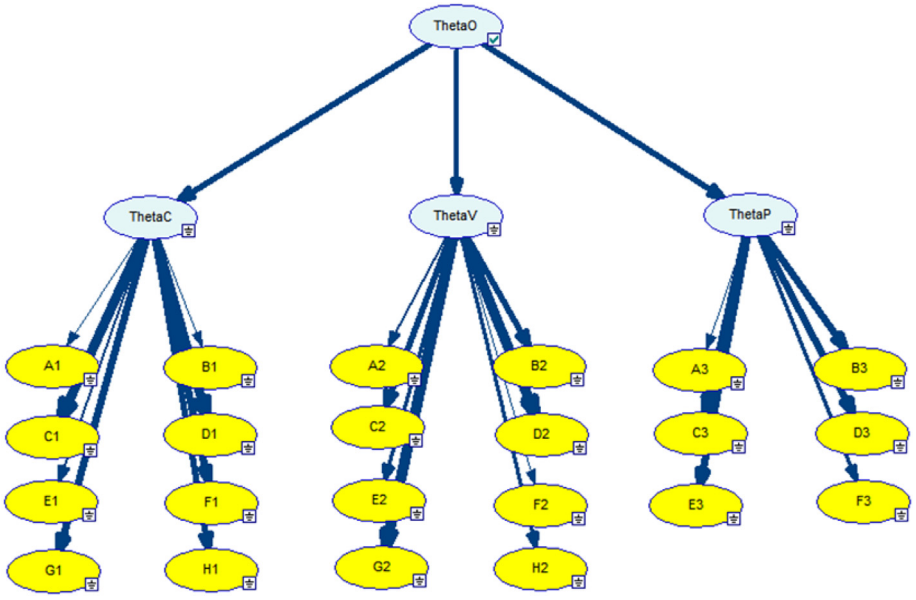

Fig. 5. Influence diagram of the model 2 Bayesian network for the confined space guard MBPA

students' responses as evidence. As stated, entering evidence has to be done individually for each student. After all evidence has been entered the network can be updated (using the junction tree algorithm discussed earlier). The results for the 57 students in our sample are presented in Table 9. The seven columns represent, respectively, the student number, the marginal probabilities that correspond to belonging to the sufficient category for the lower-level SMVs information communication, vigilance, procedural knowledge, the upper-level SMV, overall CSG proficiency, the raw sum score, and the score on the 12-point PBA rubric. The marginal probability in front of the forward slash corresponds with Model 1, whereas the probability behind the forward slash corresponds with Model 2. Note that, in general, the marginal probabilities are low, especially for the vigilance SMV. This indicates that the students did not perform well on the MBPA. Students made most mistakes in the actions in the second setting of the MBPA, which are the intervention questions (i.e., watching video fragments and pressing the stop button when incorrect worker behavior is observed).

The average probability for students to be in the sufficient category for each of the SMVs is higher in Model 1, as can be seen in the bottom line of Table 9. At the highest level, overall CSG proficiency, this results in four students belonging to the highest probability category (0.96) in Model 1, but not in Model 2 (0.97). These students belong to the sufficient category for all lower-level SMVSs. That is, their probability score is higher than the cutoff score, as defined earlier. On the other hand, there are also two students that belong in the highest probability category in Model 2, but not in Model 1. It is most likely that these students have performed well on the most essential actions, but less well on the least essential actions. There is a strong correlation between both

Table 8
Conditional probability tables for lower level SMVs θ_c, θ_v, θ_p and upper level SMV θ_o.

Model 1	θ_o	
θ_c	Sufficient	Insufficient
Sufficient	0.7482	0.2518
Insufficient	0.2518	0.7482
θ_v	Sufficient	Insufficient
Sufficient	0.7468	0.2532
Insufficient	0.2532	0.7468
θ_p	Sufficient	Insufficient
Sufficient	0.7233	0.2767
Insufficient	0.2767	0.7233
Model 2	θ_o	
θ_c	Sufficient	Insufficient
Sufficient	0.7530	0.2470
Insufficient	0.2470	0.7530
θ_v	Sufficient	Insufficient
Sufficient	0.7443	0.2557
Insufficient	0.2557	0.7443
θ_p	Sufficient	Insufficient
Sufficient	0.7646	0.2354
Insufficient	0.2354	0.7646

models, except for the CSG vigilance SMV (θ_c: $r(57) = 0.62$, $p < .01$, θ_v: $r(57) = 0.16$, $p > .05$, θ_p: $r(57) = 0.89$, $p < .01$, θ_o: $r(57) = 0.535$, $p < .01$). Notice that the marginal probabilities for θ_v in Table 9 strongly differ in some cases. It might be that the actions are equally difficult and essential in this setting, but the correlation diminishes because we forced raters to order all actions on difficulty and essence.

In the next step, we calculated the correlations between the marginal probabilities in Table 9 for both models and students' ratings on the computer experience and MBPA usability questionnaire. No significant correlations were found, which indicates that computer experience and the usability of the MBPA were not related to students' MBPA performance.

Finally, we investigated whether background characteristics (i.e., age, ethnicity, and education) are related to MBPA performance. Age was not significantly related to the marginal probabilities for Model 1 and 2. For ethnicity, we were especially interested in two groups: Dutch ethnicity versus all other ethnicities. We therefore created two groups (0 = Dutch (N = 41), 1 = other), and then calculated the point–biserial correlation between ethnicity and the thetas from Table 9. Again, we did not find any significant correlations. Finally, we created a low (N = 41) and high education group (0 = low, 1 = high). Students in the low education group continued education up to high school or lower vocational education. There were no significant correlations between education and MBPA performance. Overall, students' background characteristics are not related to their performance in the MBPA.

We looked at the relationship between θ_o (both models) on the one hand and students' sum score and PBA rubric score on the other to investigate which model is the best predictor for sufficient proficiency in CSG KSAs. The marginal probabilities of θ_o, for both Model 1 and Model 2, are strongly correlated with the sum score ($r(57) = 0.68$, $p = .00$ and $r(57) = 0.63$, $p = .00$, respectively). The correlation between θ_o for Model 1 and the sum score is a bit stronger, but there does not seem to be much difference. However, the correlation between θ_o for Model 2 and the PBA rubric score ($r(57) = -0.01$, $p > .05$) was significantly lower than the correlation between θ_o for Model 1 and the PBA rubric score ($r(57) = 0.225$, $p < .05$). Overall, these findings indicate that Model 1, in which the most difficult and least essential actions are considered to be most informative of students' KSAs, provides a better estimate of student proficiencies.

Table 9
Students' (N = 57) Marginal Probabilities for Being Sufficient on the Lower Level SMVs and the Upper Level SMV Based on their Responses in the MBPA for Model 1 and Model 2, Students' Sum Scores (S), and Students' PBA Scores (P).

No.	θ_c	θ_v	θ_p	θ_o	S	P
1	1.00/0.98	0.07/0.01	0.00/0.00	0.28/0.24	9	12
2	1.00/0.42	0.81/0.32	0.01/0.81	0.77/0.27	13	4
3	0.00/0.00	0.00/0.02	0.06/0.20	0.04/0.03	6	5
4	0.00/0.00	0.00/0.01	0.01/0.01	0.04/0.03	5	12
5	0.00/0.00	0.13/0.06	0.02/0.01	0.04/0.03	8	2
6	1.00/0.36	0.98/0.00	1.00/0.76	0.96/0.27	14	12
7	1.00/1.00	0.99/0.80	1.00/0.92	0.96/0.97	15	9
8	1.00/1.00	0.98/0.01	0.99/0.84	0.96/0.77	15	12
9	0.62/0.71	0.06/0.56	0.99/0.88	0.23/0.77	13	5
10	1.00/0.72	0.94/0.14	0.00/0.00	0.77/0.24	9	11
11	0.00/0.01	0.00/0.00	0.01/0.18	0.04/0.03	7	10
12	0.16/0.00	0.97/0.00	1.00/1.00	0.72/0.27	12	9
13	1.00/1.00	0.00/0.00	0.11/0.39	0.28/0.24	14	8
14	0.00/0.00	0.00/0.01	0.06/0.20	0.04/0.03	6	9
15	0.00/0.00	0.83/0.00	1.00/0.75	0.72/0.27	9	7
16	1.00/1.00	0.00/0.06	0.00/0.03	0.28/0.29	11	10
17	1.00/0.01	0.95/0.08	0.07/0.01	0.23/0.03	11	11
18	0.00/0.00	0.00/0.00	0.99/0.75	0.23/0.27	6	10
19	1.00/1.00	0.36/1.00	1.00/1.00	0.72/0.97	17	6
20	0.00/0.00	0.00/0.00	1.00/1.00	0.23/0.27	9	4
No.	θ_c	θ_v	θ_p	θ_o	S	P
21	1.00/1.00	0.02/0.98	0.00/0.00	0.28/0.73	14	12
22	0.98/0.01	0.04/0.01	0.02/0.02	0.28/0.03	11	10
23	0.96/0.32	0.02/0.97	0.01/0.00	0.28/0.23	11	12
24	0.99/0.03	0.00/0.02	1.00/0.98	0.72/0.27	11	7
25	0.99/0.05	1.00/1.00	1.00/0.99	0.96/0.76	15	12
26	1.00/1.00	0.00/0.98	1.00/0.95	0.72/0.96	14	5
27	1.00/1.00	0.00/0.00	0.02/0.04	0.28/0.24	12	8
28	0.99/0.02	0.96/0.31	0.20/0.26	0.77/0.03	13	12
29	0.94/0.28	0.00/0.04	1.00/0.79	0.72/0.27	11	11
30	0.99/1.00	0.05/0.10	0.00/0.00	0.28/0.24	13	11
31	0.99/1.00	0.98/0.65	0.05/0.06	0.77/0.73	16	12
32	0.99/0.03	0.01/0.99	0.99/0.26	0.72/0.23	12	12
33	0.99/0.37	0.01/0.91	0.00/0.04	0.28/0.23	11	9
34	0.00/0.00	0.00/0.00	0.77/0.75	0.23/0.27	3	8
35	1.00/1.00	0.96/0.75	1.00/0.93	0.96/0.97	16	9
36	0.03/0.70	1.00/0.01	0.04/0.03	0.28/0.24	11	11
37	0.05/0.00	0.01/0.00	0.78/0.11	0.23/0.03	6	
38	1.00/1.00	0.00/0.00	0.04/0.03	0.28/0.24	12	8
39	1.00/1.00	0.00/0.00	0.00/0.00	0.28/0.24	10	12
40	1.00/0.21	1.00/0.02	0.00/0.00	0.77/0.03	12	12
No.	θ_c	θ_v	θ_p	θ_o	S	P
41	0.00/0.00	0.83/0.00	1.00/0.98	0.72/0.27	8	9
42	1.00/1.00	0.02/0.00	0.04/0.03	0.28/0.24	14	6
43	1.00/0.44	0.20/0.39	1.00/0.82	0.72/0.27	13	9
44	1.00/1.00	0.00/0.87	0.00/0.00	0.28/0.73	10	8
45	1.00/0.60	1.00/1.00	1.00/1.00	0.96/0.76	17	12
46	1.00/1.00	0.07/0.05	1.00/0.99	0.72/0.77	14	7
47	1.00/0.40	0.19/0.00	1.00/0.99	0.72/0.27	15	10
48	0.00/0.00	0.12/0.66	1.00/0.99	0.23/0.76	11	12
49	0.00/0.00	0.00/0.04	0.97/0.13	0.23/0.03	8	11
50	0.01/0.71	0.00/0.04	0.00/0.03	0.04/0.24	9	5
51	0.99/0.60	0.03/0.99	0.96/1.00	0.72/0.76	14	12
52	1.00/0.70	0.00/0.00	0.00/0.00	0.28/0.24	8	12
53	0.00/0.00	0.79/1.00	1.00/0.88	0.72/0.76	11	9
54	0.98/0.01	0.00/0.00	0.00/0.00	0.28/0.03	9	11
55	0.00/0.00	0.01/0.00	0.98/0.98	0.23/0.27	9	10
56	0.00/0.00	0.00/0.02	0.95/0.68	0.23/0.27	7	7
57	0.99/1.00	0.02/0.16	0.96/0.88	0.72/0.77	12	12
No.	θ_c	θ_v	θ_p	θ_o		
M1 μ(σ)	0.66 (0.46)	0.31 (0.42)	0.51 (0.48)	0.47 (0.30)	S(avg)	11.09 (3.20)
M2 μ(σ)	0.45 (0.43)	0.28 (0.39)	0.46 (0.43)	0.36 (0.30)	P(avg)	9.33 (2.61)

4.3. Explorative log file analysis

In this study, the primary observables are the correctly performed actions that are scored in relation to difficulty and essence,

which is the product data. In addition, other possible MBPA performance data elements, which can be found in the log files, may also provide evidence regarding student KSAs (the process data). These elements also need to be scored to make them useful as OVs in the psychometric model during the evidence accumulation process. We identified five other data elements that may be useful for making informed inferences regarding students' KSAs: the total number of actions that a student has performed, the ratio between total actions and correct actions, the total time spent on the assessment, the average time spent per action, and the order in which the actions were performed.

First, in the MBPA, clicking on a tool (e.g., the walkie-talkie) or interacting with a virtual character in the MBPA (e.g., the operator) is considered an action. Students were not instructed to perform as few actions as possible to get to the correct actions. The best students (i.e., the students with the highest number of correct actions) may intuitively perform fewer actions because they know how to get to the correct actions quickly. On the other hand, the best students may also be more explorative in the virtual environment, clicking (i.e., performing actions) on many objects in the MBPA. In that case, we would expect the total number of actions to positively correlate with the number of correct actions. This was the case, $r = 0.54$, n = 57, $p < .001$. Because the correlation is moderate, the total number of actions can be considered as an average indicator of overall performance. In addition, the total number of actions performed in the MBPA was only significantly correlated to overall performance for Model 2 ($r = 0.302$, n = 57, $p < .05$). A second indicator, which is related to the first, could be the ratio between the number of correct actions and the total number of actions itself. This ratio was significant at the 0.01 level for Model 1 ($r = 0.384$, n = 57), and also significant for Model 2 ($r = 0.265$, n = 57, $p < .05$). These findings indicate that the ratio between the number of correct actions and the overall number of actions may be a particularly interesting element of process data in Model 1.

A third indicator could be the total time spent on the MBPA. The total number of actions performed in the MBPA does not take time into account, especially because there was no time limit imposed on students. It may be that the best students score high on the number of correct actions in the least time. However, total time spent on the MBPA is only weakly correlated to the number of correct actions ($r = 0.22$, n = 57, $p > .05$) and is not correlated to overall MBPA performance in either model.

The fourth indicator, average time spent per action, is correlated to the number of correct actions, $r = 0.37$, n = 57, $p < .01$, and only significantly correlated to overall MBPA performance for Model 2 ($r = 0.279$, n = 57, $p < .05$). These findings indicate that the number of actions performed is a better indicator of overall performance than time spent on the MBPA.

Finally, the fifth indicator, the order of actions, is only important in the alarm setting. In the other settings, although some actions are nested, the order itself is not relevant. In the alarm setting, students were explicitly instructed to perform actions in the order they thought to be correct. Following the right procedure can save lives in an emergency setting. For example, when a plant alarm goes off, students first have to warn the workers inside the confined space by clicking on the sign rope in the MBPA. The correct order of actions has been defined through consultation with SMEs and there is only one correct order. We could then count, for each student, the number of correct actions in a row, starting with action A3 (warning the workers). If one mistake was made, other correct links in the order were not counted. This method of scoring the ordering task is theoretically the most defensible. In a high-risk environment where the order of actions is very important, all actions have to be carried out in the correct order. We then correlated the number of correct links with the overall performance score of students in Model 1 and Model 2, which surprisingly showed negative correlations, although these were not significant ($r = -0.189$ and $r = -0.175$, respectively).

Overall, we believe that some of the indicators discussed above may be useful for evidence accumulation. That means that they can be used as OVs in a psychometric model. However, some indicators do not relate to the overall performance of the student, which makes them less useful. The correlation of the indicator that most substantially relates to overall performance (the number of correct actions/total number of actions ratio) can be explained by the fact that the best students know to perform the correct actions without much explorative behavior in the MBPA (hence a higher number of total actions).

These exploratory findings indicate that it is essential to know why students behave in certain ways in the MBPA as this indicates the use of certain variables as evidence for their KSAs. This relates directly to the validity of the MBPA. A careful analysis of the distribution of actions (i.e., performing the same action multiple times or constantly performing different ones) and the behavioral aspects of performance in a simulation may be necessary to provide evidence for the overall validity of the MBPA. Such an analysis is beyond the scope of this article. The point to drive home here is that MBPAs and simulations in general provide the opportunity to collect more performance data than the traditional correct/incorrect differentiation, but that the use of these indicators should be KSA-driven. Although we do not use it here, the methodology provides the opportunity to use process data.

In summary, the results section (evidence identification and evidence accumulation) provided an illustration of how OVs can be created from complex and interactive actions in MBPAs (i.e., evidence identification) and how these OVs can be scored and synthesized in a psychometric model (i.e., evidence accumulation), for use as measures of several latent student model variables.

5. Discussion and conclusion

In this article, we presented an innovative MBPA for assessing the KSAs of CSGs in Dutch vocational education. Students were able to interact with multiple elements in the virtual environment of the MBPA to perform the actions of a CSG. We then discussed a methodology to score the complex and interactive behavior that students demonstrate in the assessment. The methodology is based on consultation with SMEs and their ratings of the actions in the MBPA. Based on their ratings, we constructed two models in which the interaction between the difficulty and essence of the actions define how informative they are regarding student KSAs. We also showed how the experts' ratings could be modified for use in a BN to make informed statements about students' proficiencies based on their interactive task performance. Finally, we used empirical data to illustrate the methodology applied and to investigate the qualities of both models.

The current study met two challenges: first, the evidence identification challenge, which refers to finding meaningful (combinations of) elements in the performance data of students; and secondly, the evidence accumulation challenge, which refers to synthesizing and aggregating the scores from evidence identification in a psychometric model (Rupp et al., 2010). By addressing these challenges, we showed how informative and valid inferences can be made about the proficiency of CSG students outside of the virtual MBPA environment. Previous research has focused strongly on using simulation-based assessments for a formative purpose, whereas our study focused on a summative purpose. Using the methodology applied in this article, it is possible to use interactive task performance data to make informed credentialing decisions in a vocational education context.

This research contributes to a broader stream of research on using innovative simulations in an assessment context. Relatively little of that research has focused on vocational education. Because vocational education has specific characteristics (e.g., a strong focus on the correct execution of specified procedures), we fill an important void in the current status of the field. In addition, the integrated approach that we followed in our study, addressing both evidence identification and accumulation in one study, in combination with the empirical illustration, provides a framework to design and carry out future research. Of course, there are also limitations to the use of Bayesian Networks. For example, because we do not have enough student data, we had to base the conditional probability tables on expert ratings. This has been done before (see Shute, 2011), however we believe that the Bayes net can be improved and updated after enough student data has been collected. Furthermore, the hierarchical structure in which both observable and latent variables are depicted remains to some extent subjective. Although this structure has been established by experts, we can only be sure about the structure when it is backed up by data. A specific question for future research therefore is to test the tenability of the methodology discussed in different (educational) settings. For example, a validity study could be used to investigate the extent to which a chosen model holds true.

The psychometric qualities of the model could also be further investigated. Empirical research on model diagnostics for BNs in SBAs is still rare (especially in a vocational setting) (Sinharay, 2006). For example, in the article, we have not discussed model diagnostics and goodness-of-fit statistics in great depth. Although these nowadays exist for Bayesian Networks (cf. Almond et al., 2015; Levy, Mislevy, & Sinharay, 2009; Sinharay, 2006), the techniques are not as straightforward as for example a chi-square test of goodness-of-fit. Furthermore, well-known and relatively easy applicable item response theory characteristics cannot be used as well, because of the nature of the variables in both models. BNs incorporate discrete and multivariate latent variables, whereas the variables in an item response model are continuous. Overall, a full and deep discussion (as would be needed for discussing model BN model diagnostics) is beyond the scope of this article. This is of course a very interesting research endeavor, and we hope to publish about it in a future publication.

To conclude, this article has presented a methodology for using students' interactive task performance scores in MBPAs to make valid inferences about KSAs outside the assessment context. Our methodology incorporates two important challenges for the psychometric evaluation of complex and interactive response patterns. Future research should focus on testing the tenability of the methodology. In a broader context, we have contributed to the theory and practice of (educational) assessment in virtual environments, which is an expanding field in research and practice. As the use of SBAs continues to grow, there is an increasing need for strong methodologies that can be used to analyze complex and versatile student performance data. This article provided such a methodology. Both the theoretical and practical field can build on our work to explore the possibilities for analyzing rich data about student KSAs.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.chb.2016.02.071.

References

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. *Environmental Modelling & Software, 26*(12), 1376–1388.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. New York: Springer.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2008). Design and validation of technology-based performance assessments. In J. M. Spector, M. D. Merrill, J. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 595–604). New York: Taylor & Francis Group.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*, 574–594.

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2007). *Problem solving in technology rich environments: A report from the NAEP technology-based assessment project, research and development series (NCES 2007–466)*. U.S. Department of Education, National Center for Educational Statistics. Washington, DC: U.S. Government Printing Office.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339–353.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency, 86*, 127–137.

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

Cizek, G. J. (2006). Standard setting. In T. Haladyna, & S. Downing (Eds.), *Handbook of test development* (pp. 225–259). Mahwah, NJ: Lawrence Erlbaum.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309–328.

Dawid, A. P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing, 2*, 25–36.

De Klerk, S. (2012). An overview of innovative computer-based testing. In Theo J. H. M. Eggen, & Bernard P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC. Ipskamp: Enschede* (pp. 151–161).

De Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Baysian network example. *Computers & Education, 85*, 23–34.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (submitted). A framework for designing and developing multimedia-based performance assessment in vocational education. Manuscript submitted for publication.

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

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associate Publishers.

Drudzel, M. (2005). *Genie 2.0*. Retrieved from https://dslpitt.org/genie/.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–304.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Fleiss, J. L. (1981). *Statistical methods for raters and proportions*. New York, NY: John Wiley & Sons.

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, Student Testing, Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R775.pdf.

Jensen, F. V., Lauritzen, S. L., & Olesen, K. G. (1990). Bayesian updating in recursive graphical models by local computations. *Computational Statistics Quarterly, 4*, 269–282.

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*(1).

Koenig, A. D., Lee, J. J., Iseli, M. R., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulation* (CRESST Report 771). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence*. Boca Raton, FL: CRC Press.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society (Series B), 50*, 157–224.

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

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

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4*, 333–369.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for

multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519–537.
Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance assessment: expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.
Lucas, P. (2001). Bayesian networks in medicine: a model-based approach to medical decision making. In *Proceedings of the EUNITE workshop on intelligent systems in patient care, Vienna* (pp. 73–97).
Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. Williamson, R. Mislevy, & I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123–167). Hillsdale, NJ: Lawrence Erlbaum Associates.
Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (2000). *Bayes nets in educational assessment: Where do the numbers come from?* (CSE Tech. Rep. No. 518). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Mislevy, R. J., Oranje, A., Bauer, M. I., Von Davier, A., Hao, J., Corrigan, S., et al. (2014). Psychometric considerations in game-based assessment. *GlassLab Report*. Retrieved from http://www.instituteofplay.org/work/projects/glasslab-research/.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.
Neapolitan, R. E. (2003). *Learning bayesian networks*. New York, NY: Prentice-Hall.
Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.
Prensky, M. (2001). Fun, play and games: what makes games engaging?. In *Digital game-based learning*. New York, NY, USA: McGraw-Hill.
Quellmalz, E., & Pellegrino, J. (2009). Technology and testing. *Science, 323*, 75–79.
Rupp, A. A., DiCerbo, K. E., Levy, R., Benson, M., Sweet, S., Crawford, A., et al. (2012). 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., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment, 8*(4). Retrieved from http://escholarship.bc.edu/jtla/vol8/4.
Rupp, A. A., Nugent, R., & Nelson, B. (2012). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining, 4*(1).
Schrader, P. G., & McCreery, M. (2012). Are all games the same? In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 11–28). New York, NY: Springer.
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 assessment. *Journal of College Teaching & Learning, 5*(9), 27–32.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.
Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias, & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–523). Charlotte, NC: Information Age Publishing.
Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.
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.
Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*, 1–33.
Squire, K. D. (2003). Video games in education. *International Journal of Intelligent Simulations and Gaming, 2*(1), 49–62.
Wainess, R., Koenig, A., & Kerr, D. (2011). *Aligning instruction and assessment with game and simulation design* (CRESST Report 780). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Wass, V., McGibbon, D., & Van der Vleuten, C. P. M. (2001). Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Medical Education, 35*, 326–330.
West, P., Wise Rutstein, D., Mislevy, R. J., Liu, J., Choi, Y., Levy, R., et al. (2010). *A Bayesian network approach to modeling learning progressions and task performance* (CRESST Report 776). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).