MULTIMEDIA-BASED PERFORMANCE ASSESSMENT IN DUTCH VOCATIONAL EDUCATION

Sebastiaan de Klerk

Graduation Committee

Chairman	prof. dr. Th.A.J. Toonen
Promotors	prof. dr. ir. T.J.H.M. Eggen
	prof. dr. ir. B.P. Veldkamp
Members	prof. dr. M. Ph. Born
	prof. dr. C.A.W. Glas
	prof. dr. A.W. Lazonder
	prof. dr. A.F.M. Nieuwenhuis
	prof. dr. J.M. Pieters

De Klerk, Sebastiaan
Multimedia-based performance assessment in Dutch vocational education
PhD Thesis University of Twente, Enschede, the Netherlands. – Met een samenvatting in het Nederlands.
ISBN: 978-90-365-3997-5
doi: 10.3990/1.9789036539975
Printed by Ipskamp Drukkers, Enschede
Cover designed by Lianda Beterams

This research was supported by eX:plain.

MULTIMEDIA-BASED PERFORMANCE ASSESSMENT IN DUTCH VOCATIONAL EDUCATION

DISSERTATION

to obtain

the degree of doctor at the University of Twente,

on the authority of the rector magnificus,

prof. dr. H. Brinksma,

on account of the decision of the graduation committee,

to be publicly defended

on Friday, January 15th, 2016 at 12:45

by

Sebastiaan de Klerk

born on July 17th, 1986

in Haarlem, the Netherlands

This dissertation has been approved by the promotors:

prof. dr. ir. T.J.H.M. Eggen

prof. dr. ir. B.P. Veldkamp

Contents

Abbreviations

AERA	American Educational Research Association
APA	American Psychological Association
ATP	Association for Test Publishers
BN	Bayesian network
CAF	Conceptual assessment framework
CAT	Computer adaptive test(ing)
CBA	Computer-based assessment
CBT	Computer-based test(ing)
CPT	Conditional probability table
CSG	Confined space guard
CTT	Classical test theory
DAG	Directed acyclic graph
DCM	Diagnostic classification model
ECD	Evidence-centered design
EDM	Educational data mining
GBA	Game-based assessment
GLB	Greatest lower bound
IRT	Item response theory
ITC	International Testing Committee
KSAs	Knowledge, skills and abilities
MBPA	Multimedia-based performance assessment
MC	Multiple choice
MIRT	Multidimensional item response theory
MPCE	Multimediaal praktijkgericht computerexamen
OMR	Optical mark recognition
OV	Observable variable
P&P	Paper-and-pencil
PBA	Performance-based assessment
RCEC	Research Center for Examinations and Certification
SBA	Simulation-based assessment
SME	Subject matter expert
SMV	Student model variable
TBA	Technology-based assessment
VET	Vocational education and training

List of Figures

List of Tables

Chapter 1. General Introduction[1]

Driven by the digital revolution, computer-based testing (CBT) in educational assessment has witnessed an explosive rise during the last few decades. Computer-based administration of tests is taking place to an increasing rate in education and many formerly Paper-and-Pencil (P&P) tests have been transformed to a computer-based test.

More recently, efforts have been made to further innovate CBTs using new item formats, response actions, and multimedia. These innovative CBTs are characterized by a high degree of interactivity between the test taker and the test. As these tests become more complex, new scoring methods and psychometrics are also needed to analyze the performance data test takers produce. One such stream of innovative CBT can be defined as simulation-based assessment (SBA).

In an SBA a test taker is generally confronted with assignments in an artificially re-created computer-based environment that reflects a real-world setting. The most well-known example probably is the flight simulator in which pilots' skills in flying and controlling an aircraft, which are highly complex, under various conditions are being trained and tested. Of course, SBAs can also be designed and developed for more standardized and less complex professions.

There is a need for simulation-based assessment in Dutch vocational education, to fill a void in the assessment programs of many qualifications. Currently, vocational education in the Netherlands relies heavily on performance-based assessment (PBA). PBA is an assessment method that is prone to measurement error, which results from several sources as demonstrated by Shavelson, Baxter, and Gao (1993). Adding SBA to the assessment program may diminish measurement error over the whole assessment program.

In this chapter, vocational education in the Netherlands will be introduced and the way assessment, via assessment programs, is taking place in many of the educational qualifications in vocational education will be discussed. Then, a

[1] This chapter incorporates discussions presented in De Klerk, S. (2012). An overview of computer-based testing. In T.J.H.M. Eggen & B.P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC* (pp. 137-150). Enschede: RCEC and in De Klerk, S., Van Dijk, P., & Van den Berg, L. (2015). Voordelen en uitdagingen voor toetsing in computersimulaties [Advantages and challenges of assessment in computersimulations]. *Examens*, *12* (1), 11-17.

broader introduction to innovative CBT, and the introduction of a specific type of SBA in Dutch vocational education, called multimedia-based performance assessment (MBPA), will be discussed.

1.1 Vocational Education and Training in the Netherlands

During vocational education and training (VET) students are being prepared for a career in a specific vocation. Vocations can range from being a tradesman in a particular craft, for example a carpenter, to holding a professional position, for example an assistant at a law firm. The educational programs in VET are called *qualifications* and these are constructed by a collaboration of educational institutions and the labor market. Together, they build so-called *qualification profiles* that indicate which core tasks and work processes a student has to master during their education to become certified.

Students can follow two pathways on four levels: a school-based and a work-based pathway ranging from the entry level to the middle-management or specialist level. The difference between both pathways is that the school-based path exists of full-time education alternated with internships, while the work-based path exists of four days of work and one day of schooling. Thus, both pathways emphasize learning in practice. The core tasks and work processes depicted in the qualification profile should be part of the work or the internship of the student so that these can be mastered in practice.

Most students do their vocational education at one of the seventy VET colleges in the Netherlands, but it is also possible to take part in VET in private or adult education. In general, students start with their vocational education at the age of sixteen and finish by the age of nineteen or twenty. When students finish their vocational education they should be ready to function as entry-employees in the profession they have been prepared for. As mentioned earlier, learning by practice is an important characteristic of Dutch VET and students have to be assessed accordingly. Therefore, an important role in the assessment programs of many qualifications is reserved for performance-based assessment. That is, student knowledge is mostly tested using P&P tests, while students' practical skills are usually subject of measurement in the PBA.

1.2 Assessment in Vocational Education and Training

The cornerstone of assessment in VET is that multiple assessment methods together form an assessment program that assesses all core tasks and work processes described in the qualification profile of the particular vocation. All types of assessments and a broad delineation of the content of the assessments are presented in an assessment program. Therefore, the assessment program also functions as a first step toward designing and developing a new assessment. The assessment program contains traditional tests, consisting of closed format items or essay questions, portfolio assignments, oral exams, and PBAs.

1.2.1 Performance-based Assessment

PBA may be the oldest form of assessment in the world. Already in 1115 BC, Chinese candidates for government positions were assessed in six fields in civil service examinations: music, archery, horsemanship, writing, arithmetic, and the rites and ceremonies of public and private life (Mellenbergh, 2011). During the middle ages and early modern time, performance-based assessment was especially used in the system of craft guilds. An apprentice was trained by his master in the well-kept secrets of his craft. For the apprentice, to become a master himself, he had to undergo an examination which was called a masterpiece. Usually the apprentice had to make a product, which was then evaluated by his master (connoisseurship evaluation) (Madaus & O'Dwyer, 1999). In the 1900s, the performance-based assessment came under pressure. In the interest of economy and time, students were given tests that they could do with the whole class at once, instead of individualized tests. Then, in 1914, Kelly introduced the multiple-choice item (Madaus & O'Dwyer, 1999). The introduction MC questions lead to a predominance of MC-tests in education until the current day. Especially the efficiency of MC tests, both in test construction and psychometrics, has been the reason for its long period of domination in educational assessment. However, since the beginning of the 1990s an increasing emphasis in educational assessment has been on PBA.

There were several reasons for a renewed interest in using PBAs. PBAs were believed to lead to more contextualized and richer teaching, thereby widening the narrowed down curriculum again. In fact, it was believed that PBA could function as a lever of educational reform. Complex performances, open-ended problems, hands-on tasks, learning by doing and new skills should be the

focus of education, not mere knowledge replication tested by filling out an answer sheet. At the moment of writing, more than 20 years later, educational reformers are still working on bringing these types of skills (now called 21st century skills) into the classroom. Therefore, the PBA is still a very important measurement tool at all levels of education.

This also holds for Dutch vocational education which changed its focus in the late 1990s and early 2000s from the traditional type of education, which emphasizes transmittance of knowledge from teacher to student, to a competence-based form of education, which emphasizes learner driven development of competencies, skills, and abilities in a contextualized environment. Accordingly, the PBA became the most important assessment method in the assessment program of most qualifications. During their training students frequently perform in PBA.

These PBAs take place either in a real job environment, for example during their internship, or in a simulated setting, for example at school. PBAs that are carried out during an internship sometimes take days or even weeks. In that period the student is observed on a regular basis by his supervisor, often a manager or practical instructor. First, the student gets a specified period of time to learn the job, then a period of assessment starts in which the student can demonstrate that the core tasks or work processes of the job have been mastered. In a simulated setting, the PBA usually takes less time. The student fulfills one or more assignments that resemble real-world tasks or processes. It is not uncommon that actors take part in these assessments and that physical elements from the real-world environment are present. In general, the performance of the student is observed and rated by one or more raters.

Although PBA has some unique characteristics compared to traditional tests and seems like an excellent tool for assessing student competency in a contextualized environment, there have also been questions regarding its validity. For example, Shavelson, Baxter, and Gao (1993) have demonstrated, using generalizability theory (Brennan, 1983), that a combination of the rater being used, the occasion of the assessment and the task being presented are a major source of measurement error in PBA. A solution for the measurement error that results from the use of PBA may be to computerize performance assessments. Computers are objective scorers, present standardized assessments, and provide the opportunity to present a multitude of tasks that would be possible

in a real-world setting. The possibility of building computerized performance-based assessments or, as we call it, multimedia-based performance assessments, has been made possible by another important and growing phenomenon in educational testing: computer-based testing or CBT.

1.3 An Overview of Innovative Computer-based Testing

Although CBT has been introduced several decades ago, its widespread implementation in Dutch vocational education is still underway. The first CBTs were mainly computer transformed P&P tests. However, under the influence of the rapidly progressing digital revolution, nowadays CBT is much more. Innovative item types, the inclusion of multimedia, computerized adaptive testing, and the use of simulations and serious games as assessment instruments are all ongoing innovations in CBT (Parshall, Spray, Kalohn, & Davies, 2002).

The availability and utilization of personal computers has been growing explosively since the 1980s, and will continue to do so in the coming decades. The educational system has not been oblivious to the explosive rise of PCs and technology in general. For example, the development of high-speed scanners, or Optical Mark Recognition (OMR), halfway through the 1930s of the 20th century introduced the possibility of automatically scoring multiple-choice tests. More recently, during the late 1970s, the first computer-delivered multiple-choice tests emerged, and computer-based testing (CBT) was born. Further improvements and cost reductions in technology made the application of large-scale, high-stake CBTs during the 1990s possible. Present advances in technology continue to drive innovations in CBT, and new CBTs are being designed on a regular basis by a whole range of educational institutions. Nowadays, test developers can incorporate multimedia elements into their CBTs, and they can develop innovative item types, all under the continuing influence of technology improvements.

Because innovations in CBT continue to emerge in many different forms, a dichotomous categorization of CBTs as innovative versus non-innovative is not possible. More specifically, however, innovation in CBTs may be seen as a continuum along several categories. For instance, some CBTs may be highly innovative (scoring innovativeness in multiple categories), while other CBTs are less innovative (scoring innovativeness in only one category). Inclusion of media (e.g., video, animation, or pictures), test format (e.g., adaptive), item format

(e.g., drag- and drop-, matrix-, or ranking and sequencing questions), and construct measurement (e.g., skills or competencies) are all attributes upon which the innovativeness of a CBT can be determined. In general, using PCs or technology to develop creative ways of assessing test takers or to measure constructs that were previously impossible to measure is the most important category for innovations in computerized testing. Below, we will discuss eight categories of innovation in computer-based testing. The eight categories of CBT innovation are: item format, response action, media inclusion, level of interactivity, and scoring method, measurement of change, dynamic assessment, and modern psychometric models. The first five categories are already discussed in Parshall et al. (2002). The reason that we can add three categories is that CBT is evolving fast, which results in new possibilities in only 13 years. Each of the eight innovation categories will be discussed below.

1.3.1 Eight Categories of Innovation in Computer-based Testing

The first category is the item format, and this category makes reference to the response possibilities of the test taker. The multiple-choice item format probably is the most well-known item type, and can also be used in paper-and-pencil tests. Multiple-choice items fall into the category of so-called *selected response* formats. The characterizing feature of these formats is that the test taker is required to select one or multiple answers from a list of alternatives. In contrast, *constructed response* formats require test takers to formulate their own answers, rather than select an answer from a list of alternatives (Drasgow & Mattern, 2006). A fill-in-the-blank item type is an example of constructed response format, but essay questions and short answers are also constructed response items. All of the selected- and constructed-response item types can be administered by computer and, even more importantly, a growing amount of innovative item types are uniquely being designed for CBTs.

Scalise and Gifford (2006) present a categorization or taxonomy of innovative item types for technology platforms. The researchers have identified seven different item formats, and 28 corresponding item examples (four per category) after a profound literature search, and reported these item examples in their paper. Most of the 28 item types are deliverable via a PC; however, there are a substantial number of item types that have specific advantages when computerized. For example, categorization, matching, ranking and sequencing,

and hot-spot items are item types that are most efficiently administered by computer, compared to paper-and-pencil administration. Innovations in item format demonstrate that innovation is actually twofold. On the one hand, we can create new item types to measure constructs differently (improved measurement). On the other hand, we can also create new item types to measure completely different constructs that were difficult to measure before. This will also hold for the other categories of innovation, as will become clear in the following sections.

The second innovation category is response action, and this category represents the physical action(s) a test taker has to perform in order to answer a question. The most common response action is of course filling in an answer sheet of a multiple-choice test in a paper-and-pencil test, or mouse clicking in a CBT. However, computerized testing software and computer hardware offer some interesting features for response actions. For example, test takers can also report their answers by typing on the keyboard, or speak them into a microphone (possibly integrated with voice recognition software). These types of response actions can hardly be called innovative nowadays, because they have been available for quite some time now. However, they show the constant progress in educational testing, influenced by the technological revolution.

Response actions in CBTs of skill assessment have been studied for the last two decades, with researchers looking for possibilities to assess skill in a way such that the response action corresponds with the actual skill under investigation. For example, joysticks, light pens, touch screens, and trackballs were used by the test takers as tools for the response actions. This resulted in another stream of innovations in assessment. The current innovations in assessment show that a whole new movement of response actions is emerging. Researchers are trying to unite response action and skill assessment, for example, through virtual environments, serious gaming, camera movement recognition, simulation software, and other innovative technologies that require test takers to physically perform a range of actions (e.g., a flight simulator). Van Gelooven and Veldkamp (2006) developed a virtual reality assessment for road inspectors. Because traffic density keeps on rising, road inspectors have taken over some tasks that used to be the duty of the traffic police, for instance, signaling to drivers, towing cars, and helping to fill in insurance documents after accidents. The test takers (road inspectors) are confronted with a virtual reality projected

on a white screen. The director starts a specific case, and test takers can walk through the virtual environment with a joystick. During the assessment, all sorts of situations or problems develop, and the test takers are required to carry out actions with their joystick in the virtual environment. This example shows how assessments can be designed with innovative use of the response actions (controlling a joystick) a test taker has to perform.

The third category is media inclusion, and indicates to what extent innovative CBTs incorporate (multi)media elements. Addition of media elements to CBTs can enhance the tests' coverage of the content area and may require test takers to use specific (cognitive) skills. Also, the tests validity may be improved by using multimedia. Furthermore, reading skills become less influential during testing. Media that are regularly found in CBTs are, among others, video, graphics, sound, and animations. The simplest form is providing a picture with an item stem, as is sometimes the case in paper-and-pencil tests. Ackerman, Evans, Park, Tamassia, and Turner (1999) have developed such a test of dermatological disorders that provides test takers with a picture of the skin disorder. Following presentation of the picture, the test taker is asked to select the disorder from a list on the right side of his screen. The assessment remains rather "static"; however, it would be a more complex assessment form if test takers had to manipulate the picture provided with the item, for example, by turning it around or fitting it into another picture. Still more difficult are items in which test takers have to assemble a whole structure with provided figures or icons, for example, when they have to construct a model and the variables are provided.

Audio is most often used in foreign language tests, and usually requires test takers to put on headphones. However, other fields have also used audio in (computerized) testing. For example, the assessment of car mechanics sometimes relies upon sound. Test takers have to listen to recorded car engines and indicate which cars have engine problems. In addition, medical personnel are presented with stethoscope sounds during assessment, and they are asked which sounds are unusual. Another innovative application of sound in assessment is to present questions in sound for people who are dyslexic or visually-impaired.

Video and animations are other media elements that may be incorporated into CBTs. These media elements are highly dynamic, and are highly congruent

with authentic situations that test takers will face outside of the assessment situation. Several researchers have carried out case studies in which assessment included video. Schoech (2001) presents a video-based assessment of child protection supervisor skills. His assessment is innovative because it incorporates video in the assessment, but it is not highly interactive. The test takers watch a video, and then answer (multiple-choice) questions about the video that they have just watched. Drasgow, Olson-Buchanan, and Moberg (1999) present a case study of the development of an interactive video assessment (IVA) of conflict resolution skills. Because they introduce an innovative idea for making a CBT relatively interactive, their study is described below, in the section about the level of interactivity (the fourth innovation category) of a CBT.

Interactivity, the fourth category of innovation, indicates the amount of interaction between test taker and test. As such, paper-and-pencil tests have no interaction at all. All test takers are presented with the same set of items, and those do not change during the administration of the test. In contrast, CBTs may also be highly interactive because of an adaptive element. Computerized adaptive tests (CATs) compute which item should be presented to a test taker based upon the answers given to all previous items. In that way, the CAT is tailored to the proficiency level of the test taker (Eggen, 2008, 2011). CATs are now widely used in assessment (both psychological and educational), but were initially a huge innovation made possible by the explosive growth of PCs and technology, and the introduction of Item Response Theory (IRT).

Another form of interactivity, also based on the concept of adaptive testing, is the incorporation of a two- or multistep branching function, possibly accompanied by video. Drasgow et al. (1999) present such a case study of an innovative form of a CBT. The CBT is structured upon two or more branches, and the answer(s) of the test taker form the route that is followed through the branches. The IVA of conflict resolution skills presented by Drasgow et al. required test takers to first watch a video of work conflict. Test takers then had to answer a multiple-choice question about the video. Following their answers, and depending upon their answers, a second video was started, and the cycle was completed once more. In essence, the more branches you create, the higher the assessment scores on interactivity, because it is highly unlikely that two test takers will follow exactly the same path.

Developing assessments that score high in the interactivity category is rather difficult, especially compared to some of the other innovation categories. Test developers are required to develop enough content to fill the branches in the adaptive interactive assessment. Another difficulty is the scoring of interactive CBTs. As test takers proceed along the branches of the interactive assessment, it becomes more difficult to use objective scoring rules, because many factors play a role, including weighing the various components of the assessment, and the dependency among the responses of the test taker. However, innovation in the level of interactivity has the potential to open up a wide spectrum of previously immeasurable constructs that now become available for measurement.

The fifth innovation category is the scoring method. High-speed scanners were one of the first innovations in automatic scoring of paper-and-pencil multiple-choice tests. Automatic scoring possibilities have been developing rapidly, especially in the last two decades. Innovative items that score relatively low on interactivity and produce a dichotomous score are not too difficult to subject to automatic scoring. Other innovative CBTs, for example, complex performance-based CBTs, may require scoring on multiple dimensions, and are much more difficult to subject to automatic scoring. In performance assessment, the process that leads to the product is sometimes equal to or even more important than the product itself; however, it is a complicated task to design an automatic scoring procedure for process responses as well as product responses in complex performance-based CBTs. Consider, for example, the above-mentioned branching of CBTs that incorporate video as well. Response dependency can be an obstructive factor for the scoring of these types of CBTs. This means that test takers' responses on previous items may release hints or clues for subsequent items. An incorrect answer on an item, after a test taker has seen the first video in the IAV, releases another video that may give the test taker a hint to his mistake on the previous item. Another issue is the weighing of items in a multistep CBT. Do test takers score equal points for all items, or do they score fewer points for easier items that manifest themselves after a few incorrect answers by the test taker?

Automated scoring systems also demonstrate some key advantages for the grading process of test takers' responses. The number of graders can be reduced, or graders can be completely removed from the grading process, which

will also eliminate interrater disagreement in grading. Researchers have found that automated scoring systems produced scores that were not significantly different from the scores provided by human graders. Moreover, performance assessment of complex tasks is especially costly; molding these assessments into a CBT is extremely cost and time efficient. Thus, computers offer many innovative possibilities for scoring test takers' responses. For example, the use of text mining in assessment or classification is possible because of innovations in computer-based scoring methods. Text mining refers to extracting interesting and useful patterns or knowledge from text documents. This technique provides a solution to classification errors, because it reduces the effects of irregularities and ambiguities in text documents (He & Veldkamp, 2012). Yet another stream of innovation in scoring lies in test takers' behavior, and results in the scoring or logging of mouse movements, response times, speed-accuracy relationships, or eye-tracking.

The sixth innovation category, measurement of change, refers to the fact that modern assessment systems are based on large databases, and can continually measure the growth of students in particular domains of knowledge and skills. An example is the Math Garden (Straatemeier, 2014). In this online test (and game), students in primary education can log in with their personal id, and perform sums that differ in complexity. By solving increasingly complex sums, they can have their garden flourish. Behind the screen, their growth in math proficiency is logged and can be used, for example, for diagnostic purposes. These types of tests are part of item-based learning systems. The use of an integrated system for both learning and testing is a trend in educational assessment (Wauters, 2012). As these systems are often used to track (educational) progress over time, we subsume these innovations under the sixth innovation category.

The seventh innovation category is dynamic assessment. Computers allow test developers to design assessments that are dynamic rather than static. That is, based on what a student does during the CBT, the assessment changes accordingly. An example is game-based assessment (GBA). Games are dynamic (online) virtual environments that change on the basis of what the player has done during playing time. Integrating computer games with assessment is a long and difficult endeavor (Mislevy et al., 2014).

The eight and final innovation dimension is the modern psychometric model. As tests are increasingly being used for diagnostic purposes or in dy-

namic settings, for example, different models are needed. The Math Garden, for instance, uses innovative and efficient speed/accuracy models for estimating students' math proficiency (Maris & van der Maas, 2012), and dynamic Bayesian networks (DBN's), are often used in game-based assessment (Mislevy et al., 2014). There are, of course, more innovative psychometric methods to discuss, but the point to drive home is that psychometrics is one of the innovation categories in CBT.

The key point that flows forth from the eight innovation categories described above is twofold: not only do test developers become more capable of measuring constructs *better*, they also find themselves in a position to measure *new* constructs that were difficult to measure before.

As the above innovation categories have shown, CBTs offer a lot of opportunities, and the educational field can greatly benefit from these opportunities. However, every medal has two sides, and the other side of CBTs is that they are also subjected to several (potential) risks.

First, CBTs can be very costly and difficult to develop. CBTs can be especially costly when interactive features, multimedia, or gaming elements are being used in the CBT. Secondly, a validation study of a CBT takes a substantial amount of time. Maybe more than a traditional test, because usability and design features of the CBT should also be subject of inquiry. Thirdly, specific (IT) expertise is needed to minimize the possibility of test disclosure and problems with test security. Fourthly, because technology improves so quickly, and because it takes a substantial amount of time to build a CBT, the CBT may look outdated when finished. Or as Schoech (2001) subtly notes:

> Walk the fine line between current limitations and potentials. Exam development is often a multi-year process, yet technology changes rapidly in several years. Thus, a technology-based exam has the potential to look outdated when it is initially completed.

1.3.2 Research in Computer-based Testing

One stream of current CBT research focuses on improving the measurement of skills and performance abilities. Computers enable test developers to create high-fidelity computer simulations that incorporate innovations on all of the categories discussed above. Those types of CBTs are designed with the goal

of measuring skill and demonstrating performance. Additionally, they correspond to actual task performance to a great extent, which is defined as the authenticity of an assessment. Therefore, these CBTs rely more upon the concept of authenticity than multiple-choice tests do, for example. Integration of multimedia, constructed response item types, highly interactive designs, and new (automatic) scoring methods will lead to an assessment form that closely approximates performance assessment in its physical form. The research presented in this thesis is based on the design, development and analysis of this type of CBT, which we have called multimedia-based performance assessment (MBPA).

The research on SBA that has been reported in the scientific literature mainly focuses on cognitive performance tasks, usually in primary school (e.g., arithmetic skills). Some case studies exist that have tried to measure particular constructs in skill-based professions, for example, in the medical professions or ICT. Current research also focuses on measuring skill constructs in vocational professions that rely upon physical skills rather than cognitive or intellectual skills. The continuing technological revolution makes it possible for test developers to further innovate, create, and revolutionize CBTs. The coming decade will be very interesting for the educational measurement field, and there is a whole new range of SBAs (or CBTs) to look forward to.

1.4 Simulation-based Assessment

Simulation-based assessment is an overarching term for all simulation driven assessment forms in which a real-world setting or activity is imitated in a virtual setting (Levy, 2013). Using computer-based simulations for testing student proficiency can both expand and strengthen the domain of testing. Expand, because SBA can uncover particular knowledge, skills or abilities (KSAs) of students that were difficult, if not impossible, to measure with P&P tests and/or PBAs. In this respect, one can think about dangerous situations, for example handling a plane during a crash, or working with hazardous substances in a high-risk environment. These skills can be tested in a realistic virtual environment, but certainly not in a real-world environment. And strengthen, because the use of SBA, in combination with P&P tests and/or PBAs, may result in more valid inferences about specific KSAs of students. For example, in a PBA it is often possible to perform one or at the most a few tasks, simply be-

cause of time, logistics, and cost considerations. While it is possible to have students perform a multitude of tasks, in multiple settings, in the virtual environment, which increases the reliability of the assessment.

1.4.1 Advantages of Simulation-based Assessment

As said, a good example of expanding the domain of testing through SBA is that SBAs enable the possibility to assess students' skill in uncommon or dangerous situations. These are difficult skills to test in a PBA, but an SBA can, to a growing extent, realistically simulate these situations and innovative response actions can be used to capture student performance. A good example of strengthening the domain of testing through SBA is that the efficiency of SBA makes it possible to present a multitude of cases and assignments than possible in a P&P test or PBA. In that way, it is possible to gather more information about students' KSAs, which in turn influences the overall validity of the inferences that one intends to make.

Thus, compared to P&P tests and PBA, the SBA can heighten overall representativeness of the construct under measurement by increasing the amount of cases and tasks and by incorporating tasks that do not correspond with other ways of testing. The *Navy Damage Control Simulation*, designed and developed by CRESST, illustrates the first way of heightening the overall representativeness (Iseli, Koenig, Lee, & Wainess, 2010). In this simulation, navy personnel's KSAs on handling dangerous situations aboard a naval ship are tested. Using a virtual avatar, the test taker can move through different compartments on the ship. During the exploration of the ship they encounter different dangerous situations. For instance, the test taker may be confronted with a fire. The test taker can then use an interactive interface to indicate what type of fire it is, which fire extinguisher should be used and if there is a need for assistance or not. The *SimScientists* SBA developed by WestEd provides an illustration of heightening construct representation by incorporating tasks that are difficult to administer in P&P tests or PBAs (Quellmalz, Timms, Silberglitt, & Buckley, 2012). The simulation is intended for students' in K-8 and consists of all sorts of science assignments. For example, one of the tasks is that students have to build a food web by drawing arrows from each food source to the eater. Of course, this can also be done using pen and paper, but the big advantage of using interactive computer assignments is that the computer can log all actions

that the student has undertaken during building the food web. It may very well be that this results in more meaningful information about students' knowledge, thereby directly influence the validity of the inferences made on basis of the performance of the student.

Another advantage of SBA is that simulations offer the possibility not only to collect *product data* but also *process data*. Product data can best be defined as end product observable variables that test takers produce by completing a simulation. In the SimScientists example, described above, this would be the final configurations of the food web that has been build. Process data, on the other hand, are log files that can show in great detail *how* students have produced their product data. Mouse clicks, navigational behavior, reaction times or the use of tools in the simulation can, among others, all be part of the process data (Rupp et al., 2012). The process data, in the SimScientists example, could be composed of students' number of tries, or the strategy followed. Has the food web been build top down or bottom up? The process data may also serve a diagnostic purpose, for example by revealing specific types of errors in students' way of thinking. Thus, both product data and process data can serve a formative and a summative purpose. In addition, scoring and resulting data in an SBA is always fully standardized. This is another important data advantage of SBA, especially compared to the PBA which is often characterized by measurement error resulting from the use of raters to judge student performance.

Simulations have been successfully used as a part of E-learning programs for quite some time now (Clark & Mayer, 2011). A next step would be to not only use them for instruction but also deploy them as measurement instruments. SBA may have some strong advantages as compared to P&P tests and PBA. The first and foremost advantage, for students, is that it is more fun to do an SBA than to fill out an answer sheet. Researchers have connected the theory of flow (Csikszentmihalyi, 1991) to *playing* a computer-based simulation (Shute, 2011), even to such an extent that SBA enables so-called *stealth assessment*: students being assessed without even noticing it. At the least, it appears to speak for itself that playing, fun, and flow influence the intrinsic motivation of the student to perform well in the simulation and that test anxiety is less experienced.

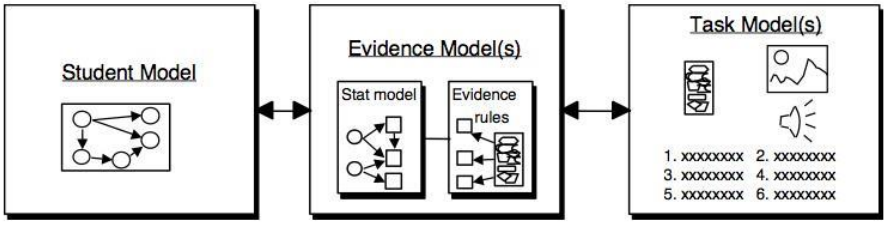
Finally, the SBA can, at different levels, be more efficient than P&P tests and PBAs. Performance evaluation can be real-time so that results and feed-

back can be communicated almost immediately to the student. In the long run, the SBA is cheaper and logistically more efficient than the PBA, presents a more contextualized environment than a traditional P&P test, and the controlled computer environment of an SBA offers the possibility to be flexible and efficient in proctoring and test security.

1.4.2 Developing Simulation-based Assessment: Evidence-centered design

Time and costs for developing simulations are still substantial, but with the ongoing progress of technological innovation and growing availability of technology worldwide, the possibility to develop a simulation as a measurement instrument grows for many educational institutions. However, there are still many challenges ahead for SBA. The big challenge lies in investigating the validity of simulations as measurement instruments. A very useful point of departure for research into the theory and practice of SBA is the evidence-centered design (ECD) framework (Mislevy, Almond, & Lukas, 2004). Therefore, before the challenges of SBA will be discussed in greater detail, the ECD framework, and in particular the conceptual assessment framework (CAF) within the ECD approach will be introduced (see Figure 1.1).

Figure 1.1
Conceptual Assessment Framework (CAF) within the ECD Framework



The ECD framework consists of various models and can be used as an approach to develop educational assessments following the rules of *evidentiary reasoning*. Evidentiary reasoning indicates that student performance on tasks within the assessment can be seen as a representation of KSAs that are being measured. In other words, evidence is collected to link student performance on assessment tasks to what is being measured. In ECD terms, these are called the

evidence model, the *task model*, and the *student model*. These models are all part of the CAF which is one of the models within the ECD framework. For the current discussion, the focus will be on the models within the CAF. We will first discuss the student model (i.e., what we want to measure), then the task model (i.e., how we want to measure what we want to measure), and finally the evidence model (i.e., how can we empirically and psychometrically link the student model and the task model). We think that this order describes best how the CAF functions at the conceptual level.

The student model articulates what construction of knowledge, skills, and abilities is being measured by the assessment. The student model may consist of one or more (latent) variables and can become complex if a combination of variables (i.e., KSAs) are required for successful completion of a combination of tasks in the assessment. A relatively straightforward student model, for example, would exist of a single proficiency variable that is represented by all items in the test. This is often the case for a knowledge-based test or a cognitive measure. Performance-based assessments, on the other hand, are instruments that are often used to measure multiple KSAs in one assessment and students have to use a combination of KSAs to successfully complete the assessment's tasks. The variables in the student model are, logically, called student model variables (SMVs).

The task model describes what type of situations – tasks – are needed in the assessment to collect information – evidence – about the student model variables. Although *presentation material* and *work products* are part of the task model, it is not simply a list of all tasks that are part of the assessment. Rather, it is a set of conditions or specifications that creates a family of possible tasks in the assessment. Assessment tasks are actually made from the task model by following the specifications for the tasks, the presentation material described and the work products needed. A PBA typically consists of more families of tasks than an MC test.

The evidence model connects the student model and the task model by providing two different yet linked processes, *evidence identification rules* and the *psychometric* or *measurement model* (also called *evidence accumulation*). The evidence identification rules explain which work products produced by students during the completion of the tasks in the assessment can be defined as observable variables (OVs) that provide evidence about the SMVs. The psychometric

model then explains how evidence resulting from the OVs is accumulated and translated into meaningful (quantitative) statements about the SMVs. The relationship between SMVs and OVs

1.4.3 Challenges for Simulation-based Assessment

Within the ECD framework, a number of challenges for SBA can be formulated. The first challenge that researchers face is to find out which constructs can be measured in an SBA. That is, which types of SMVs can possibly be measured using a computer-based simulation? The first steps to answering this question have already been made. For example, a construct that has already been measured in a virtual environment is creativity. Shute, Bauer, Ventura, and Zapata Rivera (2009) operationalized creativity by use of a commercial video game called *Oblivion*. Creativity may be considered a personality trait, which is rather stable over time (Eysenck, 1983). Other initiatives focus on measuring more fluent SMVs, for example cognitive abilities that develop over time. The *Math Garden* is an example of such an initiative (Klinkenberg, Straatemeier, & van der Maas, 2011). In this game, children work on developing their arithmetic ability by solving questions related to arithmetic. Correct answers enable them to grow a simulated garden within the game. Still other SBAs have a stronger focus on more practical constructs, or skill. The above discussed *Navy Damage Control Simulation* developed by CRESST, for instance, focuses on measuring actual behavior or the skill to handle correctly under different emergencies (Iseli, Koenig, Lee, & Wainess, 2010). The examples above delineate an evolving field, wherein researchers are trying to measure new and more complex SMVs through SBA.

The second challenge for using a computer-based simulation as a measurement instrument lies in determining to what extent it is possible to build justifiable task models composed of scorable and objective items, tasks or *actions* that yield valid inferences about SMVs. As mentioned before, in contrast to the traditional CBT, SBAs do not necessarily include the item – response format, quite the contrary, they are built on a string of actions and decisions made by the student. But which student actions or decisions made during the SBA can be regarded as OV that can be used in a measurement model? In addition, performing a simulation for a while regularly results in a log file that fills many pages. Essentially all log file entries may provide valuable information,

but it is difficult to decide beforehand which information is needed. Therefore, educational data mining (EDM) techniques are often applied to performance data produced by students in an SBA. Using EDM, clusters of actions can be grouped, specific student strategies can be identified, and student performance can be predicted. In fact, EDM's exploratory character might reveal important OVs that can be used in the more confirmatory psychometric models to make final judgments about students' KSAs.

A third challenge therefore is to build suiting measurement models that can capture the complex and versatile nature of SBAs. This challenge strongly relates to the third and final model within the CAF: the evidence model. Theory and data are united in the evidence model through two separate yet connected parts: evidence identification and evidence accumulation. The theoretical relationship between the SMVs and the OVs are formalized in the evidence model on basis of the data.

1.5 Multimedia-based Performance Assessment

In this thesis, research on a specific type of SBA is presented. The multimedia-based performance assessment (MBPA) is an SBA in which real-world activities and contexts from performance assessments are simulated in a computer-based environment by using multimedia. In that sense, it approximates the game like feeling, but it cannot be considered a video game because it does not provide an open space in which a student can freely wander around with a virtual character. Yet, there is a high amount of interactivity between student and computer and the format of the items and responses in the MBPA can range from the traditional item-response format to navigational path analysis. In the following chapters of this thesis, MBPA will be discussed in great detail.

1.6 Outline Thesis

This thesis largely covers three areas of research into MBPA: 1) literature study; 2) development of MBPA; 3) practical, psychometrical and empirical investigation into the use of MBPA in Dutch vocational education.

1.6.1 Literature Study

Chapter 2 looks at the MBPA in more detail. Several forms of performance-based assessment in Dutch vocational education are discussed in great

detail with examples. In particular, the discussion focuses on several measurement and practical concerns concerned with the PBA and an argument for the use of MBPA, to overcome these concerns, is presented. MBPA is discussed in depth and a rationale for effective use of MBPA in VET is made. The chapter ends with a pilot example of an MBPA and a structured planning of future developments of the pilot. The main question to be answered in the second chapter is: *"Why should we use MBPA?"*.

Chapter 3 presents a systematic literature review on the psychometric analysis of performance data of SBA and provides an example of a psychometric model used for analyzing students' performance in an MBPA. The purpose of this study was to map all initiatives on the use of SBA in educational measurement and to investigate the effectiveness of different psychometric methods for analyzing the performance data of SBA. The systematic review was carried out following the method described by Petticrew and Roberts (2006). In total, we found 31 articles that satisfied our criteria. In all these papers, an SBA was presented including a discussion on the psychometric analysis of the performance data of students. We end this chapter by providing an example of an MBPA, including a modern psychometric model, called the Bayesian Network, for the analysis of the performance data produced by performing the MBPA. The main question to be answered in the third chapter is: *"Which psychometric models can be used to analyze the performance data of MBPA?"*.

1.6.2 A Developmental Framework for Multimedia-based Performance Assessment

Chapter 4 presents a framework for designing and developing MBPA. In this chapter, the emphasis is on the application of the framework in vocational education, yet the framework can also be applied in other (educational) settings. Assessment development should be a structured, careful and iterative process in which multiple specialists from different fields collaborate. Because of the complex nature of this process it is highly recommendable to use a framework or set of guidelines. A developmental framework for building MBPAs was not available in the literature yet. Therefore, in this chapter, a framework consisting of two general stages, which in turn consist of a total of thirteen steps for the design and development of MBPA is presented and validated. The framework was constructed on basis of a literature syntheses and consultation of assess-

ment experts. For validation, other experts were then asked to read the paper and closely study the framework so that they could be questioned in a semi-structured interview regarding the framework. The main question to be answered in the fourth chapter is: *"How do we build an MBPA?"*.

1.6.3 Psychometric and Empirical Investigation into Multimedia-based Performance Assessment

Chapter 5 discusses the design, development, and psychometric functioning of an MBPA for a vocation called *confined space guard* (CSG). We have designed and developed the MBPA ourselves, according to the framework presented in the previous chapter. To become a confined space guard, students follow a vocational training and are subsequently assessed on their knowledge using a MC test, and their skills using a PBA. A CSG supervises operations that are carried out in a confined space. In this chapter, we first discuss why there is a need for an MBPA to assess a CSG's KSAs. Then, the design and development according to the framework discussed in chapter 4 is explained. Thirdly, an empirical study including a sample of real students is used to evaluate the psychometric functioning of the MBPA. In the empirical experiment, the students' MBPA score is compared to students' PBA score. Furthermore, we report test and item characteristics, investigate the role of computer experience, MBPA usability and students' background characteristics, and study the underlying structure of the MBPA. The main question to be answered in the fifth chapter is: *"What is the relationship between scores on an MBPA and scores on a PBA, which aim to measure the same constructs?"*.

Chapter 6 then explores the use of MBPA even deeper by presenting another MBPA for assessing the confined space guard's KSAs. Whereas the MBPA in chapter 5 is relatively structured and linear, the MBPA discussed in chapter 6 provides more of an open space for students in which they can click on items on the screen and open different assignments. That is, the first MBPA alternates between multimedia and questions in the same way for all test takers while the second MBPA gives test takers the opportunity to select different objects (using an interactive interface) to carry out assignments. In essence, the structured MBPA may be seen as an extension of the earlier mentioned innovative CBTs, whereas the interactive MBPA really belongs to the SBA category. In this chapter, a study is presented in which a sample of students performs in

the *interactive* MBPA and their PBA. The psychometric properties of the MBPA will be discussed, a model for scoring interactive student behavior is presented and an empirical comparison is made between the MBPA and the PBA in an experimental setting. The main question to be answered in the sixth chapter is: *"How can we score complex and interactive behavior in an experimental MBPA, and then apply those scores in a psychometric model?"*.

Finally, in the epilogue (chapter 7), the research presented in this thesis is discussed in a broader context. Furthermore, some general remarks and chapter specific remarks will be made, and future directions will be discussed. And one final question will be answered: *"When is it efficient and effective to use MBPA?"*.

Together, the chapters in this dissertation answer the general research question, which is defined as: "*Can we develop and use a multimedia-based performance assessment for which students' performance in the MBPA provides valid inferences regarding their knowledge, skills, and abilities?"*.

Table 1.1 presents a schematic outline of this thesis. The chapters in this thesis can be read in any order and can all be read independently.

Table 1.1
Outline of the Chapters in this Thesis

Chapter	Research topic	Method	Research objectives
1	General Introduction	Literature study	● Historical perspective of research ● Introducing terminology ● RQ's, hypotheses, framework of research
2	MBPA versus PBA and a rationale for MBPA	Literature study and pilot study	● Studying and mapping (measurement) concerns PBA ● Building an argument for the use of MBPA
3	Review of literature on MBPA and Bayesian network analysis	Systematic review of the literature and psychometric analysis	● Mapping and reviewing all literature ● Building a psychometric model for MBPA analysis

Table 1.1 (continued)
Outline of the Chapters in this Thesis

Chapter	Research topic	Method	Research objectives
4	Design and development of MBPA	Literature study, focus group meetings and expert interviews	● Building and validating a comprehensive framework for MBPA design and development
5	Building and testing a structured MBPA	Empirical experiment and psychometric analysis	● Building a full MBPA ● Practical testing of an MBPA ● Empirical comparison with PBA ● Psychometric properties of MBPA
6	Building and testing an experimental MBPA	Empirical experiment and psychometric analysis	● Building an experimental MBPA ● Scoring interactive student behavior ● Psychometric properties of MBPA
7	Epilogue	Literature study	● When to use MBPA? ● Overview of research ● Future directions

References

Ackerman, T.A., Evans, J., Park, K.S., Tamassia, C., & Turner, R. (1999). Computer assessment using visual stimuli: A test of dermatological skin disorders. In F. Drasgow & J.B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 137-150). Mahwah, NJ: Erlbaum.

Brennan, R.L. (1983). *Elements of generalizability*. Iowa City, IA: American College Testing Program.

Burstein, J., Braden-Harder, L., Chodrow, M., Hua, S., Kaplan, B., Kukich, K., et al. (1998). *Computer analysis of essay content for automated score prediction* (ETS Research Report RR-98-15). Princeton, NJ: Educational Testing Service.

Clark, R.C., & Mayer, R.E. (2011). *E-learning and the science of instruction*. San Francisco: Pfeiffer.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R.K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and Advances* (pp. 59-75). Chichester: Wiley.

Drasgow, F., Olson-Buchanan, J.B., & Moberg, P.J. (1999). Development of an interactive video assessment: Trials and tribulations. In F. Drasgow & J.B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 177-196). Mahwah, NJ: Erlbaum.

Eggen, T.J.H.M. (2008). Adaptive testing and item banking. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 215-234). Göttingen: Hogrefe.

Eggen, T.J.H.M. (2011, October). *What is the purpose of CAT?* Presidential address at the 2011 Meeting of the International Association for Computerized Adaptive Testing, Monterey, CA.

Gelooven, D. van, & Veldkamp, B. (2006). Beroepsbekwaamheid van weginspecteurs: een virtual reality toets. In E. Roelofs & G. Straetmans (Eds.), *Assessment in actie: Competentiebeoordeling in opleiding en beroep* (pp. 93-122). Arnhem, the Netherlands: Cito.

He, Q., Veldkamp, B.P., & Westerhof, G.J. (2012). Classifying unstructured textual data using the Product Score Model: an alternative text mining algorithm. In T.J.H.M. Eggen & B.P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC* (pp. 47-62). Enschede, Netherlands: RCEC.

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

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

Madaus, G., & O'Dwyer, L. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688-695.

Maris, G., & van der Maas, H.L.J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615-633.

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

Mellenbergh, G.J. (2011). *A conceptual introduction to psychometrics*. Den Haag, the Netherlands: Eleven International Publishing.

Mislevy, R.J., Almond, R.G, & Lukas, J. (2004). A brief introduction to evidence-centered design. *CSE Technical Report.* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/r632.pdf

Mislevy, R. J., Oranje, A., Bauer, M., von Davier, A. A., Hao, J., Corrigan, S., et al. (2014). Psychometric considerations in game-based assessment. New York, NY: Institute of Play.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer.

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

Quellmalz, E.S., Timms, M.J., & Buckley, B. (2010). The promise of simulation-based science assessment: the Calipers project. *International Journal on Learning in Technology, 5*(3), 243-263.

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

Rupp, A.A., Levy, R., DiCerbo, K., Sweet, S.J., Crawford, A.V., Calico, T., Benson, M., Fay, D., Kunze, K.L., Mislevy, R.J., & Behrens, J.T. (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* (1), 49-110.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved [March 20, 2012] from http://www.jtla.org.

Schoech, D. (2001). Using video clips as test questions: The development and use of a multimedia exam. *Journal of Technology in Human Services, 18*(3-4), 117-131.

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

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.

Shute, V.J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias and J.D. Fletcher (Eds.), *Computer Games and Instruction* (pp. 503-523).Charlotte, NC: Information Age Publishing.

Straatemeier, M. (2014). *Math Garden: A new educational and scientific instrument*. Doctoral dissertation, UvA, the Netherlands.

Wauters, K. (2012). *Adaptive item sequencing in item-based learning environments*. Doctoral dissertation, KU Leuven, Belgium

Chapter 2. A Blending of Computer-based Assessment and Performance-based Assessment: Multimedia-based Performance Assessment. The Introduction of a New Method of Assessment in Dutch Vocational Education[2]

Abstract

Innovation in technology drives innovation in assessment (Conole & Warburton, 2005; Bartram, 2006; Drasgow, Luecht, & Bennett, 2006; Mayrath, Clarke-Midura, & Robinson, 2012). Since the introduction of computer-based assessment (CBA), a few decades ago, many formerly paper-and-pencil tests have transformed in a computer-based equivalent. CBAs are becoming more complex, including multimedia and simulative elements and even immersive virtual environments. In Vocational Education and Training (VET), test developers may seize the opportunity provided by technology to create a multimedia-based equivalent of performance-based assessment (PBA), from here on defined as multimedia-based performance assessment (MBPA). MBPA in vocational education is an assessment method that incorporates multimedia (e.g., video, illustrations, graphs, virtual reality) for the purpose of simulating the work environment of the student and for creating tasks and assignments in the assessment. Furthermore, MBPA is characterized by a higher amount of interactivity between the student and the assessment than traditional computer-based tests. The focal constructs measured by MBPA are the same as are currently assessed by performance-based assessments. Compared to automated delivery of item-based tests, MBPA realizes the full power of ICT. In the present Chapter we will therefore discuss the current status of MBPA, including examples of our own research on MBPA. We provide an argument for the use of MBPA in vocational education too.

[2] This chapter is a minor revision of De Klerk, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2014). A blending of computer-based assessment and performance-based assessment: Multimedia-based performance assessment (MBPA). The introduction of a new method of assessment in Dutch Vocational Education (VET). *Cadmo, 22*(1), 39-56.

2.1 Introduction

Technological advancements continue to drive innovation in CBA. Presently, test developers already incorporate multimedia elements into their CBAs (de Klerk, 2012). Educational institutions are designing technology-based assessments and items in a growing rate (Bennett, 2002). These assessments simulate a highly contextualized environment and students are confronted with tasks that they could encounter in real life (Issenberg, Gordon, Gordon, Safford, & Hart, 2001; Ziv, Small, & Wolpe, 2000). The general rationale behind innovative assessments is that they provide more meaningful observations about student skills than traditional multiple-choice tests or performance-based assessment (PBA). For example, research has shown that immersive environments are capable of capturing observations that are not possible to capture in a conventional classroom setting (Clarke, 2009; Ketelhut, Dede, Clarke, Nelson & Bowman, 2008). In scientific literature, multimedia and even immersive virtual reality possibilities are being discussed in relation to assessment (Susi, Johanneson, & Backlund, 2007; Sliney, & Murphy, 2011; Clarke-Midura & Dede, 2010). Thus, technological progress provides opportunities for the design and development of innovative and interactive technology-based assessments. However, technological advancement in assessment is ahead of research and psychometrics.

Therefore, in this Chapter, we present a research project that tries to fill the void between the opportunities that technology provides for assessment and the foundation of these opportunities in theory and research. We provide an argument for technology-based assessments, we discuss the current status of technology in assessment, and we present our current research on an innovative method of assessment in Dutch vocational education, called multimedia-based performance assessment (MBPA). MBPA in vocational education is an assessment method that incorporates multimedia (e.g., video, illustrations, graphs, virtual reality) for the purpose of simulating the work environment of the student and for creating tasks and assignments in the assessment. Furthermore, it is characterized by a higher amount of interactivity between the student and the assessment than in traditional computer-based tests. Finally, the focal constructs under measurement in MBPA are the same as are currently assessed using performance-based assessment. The introduction of MBPA and technology in assessment in general is not just about doing the same things differently.

It is primarily about introducing a measurement instrument to the vocational educational field that provides new and improved possibilities for measuring students' skills.

The vocational education and training (VET) sector may be strongly diversified in Europe, or even worldwide, for that matter. However, the core elements of VET have universal applications over countries. Through the use of MBPA we try to translate these elements in a computer-based assessment. First, one of the core elements in VET is that it prepares students for vocations that for the largest part emphasize manual/practical skills or procedural knowledge. Therefore, the tasks in the MBPA should be designed around these constructs. Secondly, although countries differ in extent, for some part VET is always carried out in the original vocational setting during an apprenticeship or internship. The MBPA should reflect the vocational context as if students were working in a real setting. For instance through video material or even virtual reality elements. Finally, VET is always concerned with getting students to a proficiency level with which they can act as entry employees in their vocation on the labor market. Thus, the MBPA should distinguish between students that can act as entry employees (mastery) and students that have not reached the right level of proficiency yet (non-mastery). The main goal for assessment methods in VET, therefore, is to reach high levels of predictive validity.

Unfortunately, the current assessment methods in vocational education are not sufficient for validly measuring students' skills and competencies. For example, one of the most common assessment methods in vocational education, PBA, is prone to several sources of measurement error. In PBA, the generalizability of scores may be impaired by several factors; due to task and rater error (Dunbar, Koretz, & Hoover, 1991), due to administration occasion error (Cronbach, Linn, Brennan, & Haertel, 1997) or due to assessment method error (Shavelson, Baxter, Gao, 1993). Furthermore, it is difficult to standardize the assessment setting and PBAs are time consuming, expensive, and logistically challenging (Lane & Stone, 2006).

Technology offers interesting opportunities to create assessments that are capable of improved measurement of the same constructs that are now measured with a PBA. Research should point out whether innovative assessments can fulfil this promise or not. Therefore, in the current Chapter, we present an argument that, based on the current status of technology in assessment, justifies

a solid investigation into the use of MBPA in vocational education. The presented argument is a first effort to bring technological advancements in assessment and research together. To further specify our argument we also present a pilot example of our first attempt to create an MBPA for a vocation that is now assessed with a PBA. Next, we discuss the background context of our research project, and will then turn the discussion to our argument for the use of MBPA in vocational education.

2.1.1 Assessment in Vocational Education and Training: Challenges and Concerns

Student learning in vocational education generally revolves around acquiring complex and integrated knowledge, skill, and attitude constructs which are often referred to as 'competency' (Klieme, Hartig, Rauch, 2008). Of course, competency is not directly observable in students (Grégoire, 1997), thus to make statements about student competencies we have to rely on indirect reasoning from evidence that we collect in an assessment setting. In the assessment setting the student is confronted with assignments or tasks that require responses or behaviors. Subject matter experts (SMEs) and assessment experts together develop a model that reflects the degree to which the student has mastered a competency based on the performance of the student in the assessment setting. Student learning in vocational education is becoming more and more focused on mastery of competency. Id est, vocational education has shifted from a traditional testing culture to an assessment culture (Baartman, 2008). In practice, this resulted in an increasing emphasis on a comprehensive alternative form of assessment, generally referred to as 'authentic assessment', 'alternative assessment', or 'performance(-based) assessment' (Linn, Baker, & Dunbar, 1991; Marzano, Pickering, & McTighe, 1993; Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson, Baxter, & Pine, 1992; Sweet & Zimmermann, 1992). Although multiple-choice tests and other assessment methods are still used, performance-based assessment (PBA) is now the most pervasive assessment method in vocational education (Segers, 2004; Baartman, 2008; Dierick & Dochy, 2001; Van Dijk, 2010).

Traditional paper-and-pencil tests are not capable of capturing the complex and integrated constructs assessed in vocational education but are still used when the acquisition of knowledge is tested. Additionally, PBA has high face

validity when competency is the focal construct of assessment. For example, compared to paper-and-pencil tests most people would choose PBA to be the prevailing instrument to measure vocational competencies. Although PBA seems to be a promising method of assessment in vocational education it imposes some serious challenges and concerns upon its developers and users. Before the challenges and concerns of PBA can be discussed it should be noted that there exist multiple types of PBA. Roelofs and Straetmans (2006) discuss three types of PBAs: hands-on, simulation-based, and hands-off.

Hands-on PBAs are assessments that take place during students' work placement. The evidence of competency is collected during the observation of students performing a vocation in real life. Simulation-based PBAs are assessments that take place in more or less standardized and reconstructed settings (e.g., in school). Simulations simplify, manipulate or remove parts of the natural job environment and students are cognizant of the fact that the situation is not a real-world setting. Hands-off PBAs are paper-based assessments in which students are confronted with hypothetical vocational situations. Subsequently, the students are asked how they would or should react in these situations. In the current Chapter we will only refer to the first two types because those are used in vocational education.

Both types of PBA have their pros and cons. For example, hands-on PBAs are generally considered to be very authentic, which would increase their validity. However, they are also very difficult to standardize and sometimes students are not allowed to carry out specific tasks because of risk. Simulation-based assessments, on the contrary, can be more standardized measures of competency but are less authentic. In practice, hands-on as well as simulation-based PBAs are prone to several measurement issues and practical concerns. In Table 2.1, some (but not all) characteristics of hands-on and simulation-based PBAs are presented. We consider these characteristics to be the most important characteristics of PBA and the most important for the current discussion.

PBAs are generally characterized by flexible open ended tasks. Students are required to construct original responses, which results in a unique assessment for every student, and generally more than one correct outcome. Above that, the assessment setting is called authentic and meaningful because it resembles a real-life setting and students can better associate with the tasks compared to traditional measurement (Gulikers, Bastiaens, & Kirschner, 2004).

This leads to the first, and one of the most important measurement concerns of PBA: the interaction between standardized measurement on the one hand, and an authentic assessment setting on the other hand. As can been seen in Table 2.1, the typical hands-on PBA takes place in an authentic assessment setting, but lacks standardized tasks. Possible concerns in standardization could be detected using generalizability theory. Generalizability theory allows for the estimation of multiple sources of error in measurement (Brennan, 1983, 2000, 2001; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991).

Table 2.1
Types of Performance-based Assessment and Corresponding Characteristics

	Type	
Characteristic	Hands-on	Simulation
Standardization	-	±
Authenticity	+	±
Rater induced error	+	+
Representativeness	-	-
Feasibility	-	±
Reliability	-	±

Note. + = PBA type scores high on particular feature, ± = PBA type scores neither high nor low on particular feature, - = PBA type scores low on particular feature. Table 2.1 is a rough delineation of corresponding characteristics for the types of PBA. The table is based on a synthesis of this paper's literature.

In generalizability theory, variance components can be estimated for each facet of the assessment and the interactions between facets (Lane & Stone, 2006). Facets are, for example, the tasks, raters, and occasion. These variance components indicate to what extent the facets in the assessment cause measurement error. Furthermore, generalizability theory provides coefficients that are used to examine how well the assessment scores generalize to the larger construct domain. Poor standardization in the tasks, raters or occasion may translate into low generalizability coefficients and construct irrelevant variance. This is exactly what has been found in empirical research on measurement error

in PBA. Measurement error originates due to the selected tasks in the assessment (Baxter, Shavelson, Herman, Brown, & Valdez, 1993; Gao, Shavelson, & Baxtor, 1994), due to the use of raters in the assessment (Breland, 1983; Dunbar, Koretz, & Hoover, 1991; van der Vleuten & Swanson, 1990), due to the occasion (Cronbach, Linn, Brennan, & Haertel, 1997), or a combination of those (Shavelson, Ruiz-Primo, & Wiley, 1999).

The second concern related to PBA in vocational education is the use of raters. Several studies have shown that rater induced error and inter-rater reliability varies considerably between different performance-based assessments (Clauser, Clyman, & Swanson, 1999). For example, Dunbar, Koretz, and Hoover (1991) found reliability levels ranging from .33 to .91, and van der Vleuten and Swanson (1990) reported reliability coefficients in the range of .50 to .93. The dispersion of these numbers shows that rater induced error is a significant factor concern. Rater induced error has already been exemplified in an excellent manner by Edgeworth in 1888 (cited by Bejar, Williamson, & Mislevy, 2006): ... *let a number of equally competent critics independently assign a mark to the (work). ... even supposing that the examiners have agreed beforehand as to ... the scale of excellence to be adopted .. there will occur a certain divergence between the verdicts of competent examiners. (p.2).*

Moreover, many rater effects that influence scores have been described in literature (Eckes, 2005; Dekker & Sanders, 2008). These effects unintentionally affect the scores of students and cause construct irrelevant variance. Hence, the variance that is created in the ratings of students through the rater effects threatens the validity of the assessment (Messick, 1989, 1995; Weir, 2005). The most well-known example of a rater effect is the halo effect (Thorndike, 1920). The halo effect is a cognitive bias that influences rater's judgment of particular behavior of students based on the overall impression of the student to be judged. The training of raters and raising their conscience on these effects is found to reduce the effects' influence (c.f. Wolfe & McVay, 2010). However, rater effects are inevitable, simply because the raters are human.

The third concern is the representativeness of PBA. Fitzpatrick and Morrison discuss comprehensiveness and fidelity as two aspects of the representativeness of PBA (1971, p.240): *comprehensiveness, or the range of different aspects of the situation that are simulated, and fidelity, the degree to which each aspect approximates a fair representation of that aspect in the criterion.* That is, within each assessment domain

there is a whole range of possible tasks to include in the assessment. However, which tasks are included in the final assessment may influence students' performance. Ideally, no matter which tasks selected, they should all result in the same statements about students' mastery of skill (Straetmans & Van Diggele, 2001). In reality, students usually perform a limited amount of tasks in a PBA, because it is costly and logistically challenging to have students perform many tasks in many situations. This results in insufficient representativeness of the PBA. Poor representativeness combined with limited standardization and rater induced error are the main sources of low reliability of PBAs.

The fourth and final concern, feasibility, is not a measurement concern, but may result in diminished measurement quality. In general, PBA is considered an inefficient type of assessment. Efficiency in assessment can be described along several dimensions. The most important are time, costs, and logistics. PBAs are often time consuming, costly, and logistically challenging to design and develop. To retain some efficiency vocational schools may cut back on resources used for the assessment (i.e., money or time) or on technical aspects of the assessments (i.e., psychometric evaluation), thereby reducing the validity and overall quality of the assessment (Shavelson, Baxtor, & Gao, 1993; Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Ruiz-Primo, & Wiley, 1999; Webb, Schlackman, & Sugrue, 2000; Haertel, Lash, Javitz, & Quellmalz, 2006).

In summary, the concerns described above indicate that it is very difficult to design and develop PBAs that provide valid results about student competency. This has led Kane (1992) to a firm conclusion about the development of PBA: *basically you can't win.* At the least, it requires very careful and thoughtful development, extensive rater training, and detailed psychometric analysis to be comfortable about its functioning. These things are, unfortunately, often not possible when budget and time are considered. Technological advancement in assessment may provide the potential to realize assessments that do provide valid and reliable statements about student mastery of competency combined with practical and cost efficient development and administration.

2.1.2 A Rationale for Multimedia-based Performance Assessment

Multimedia-based Performance Assessment is grounded in the realization of the full power ICT provides. The general rationale behind MBPA is: (1) that

it enables improved measurement of competency compared to PBA because MBPA might be less prone to several of the measurement concerns discussed above, and (2) that it is a more efficient assessment method than PBA.

MBPA has higher standardization than PBA and still retains authenticity (Clarke, 2009; Schoech, 2001; Bakx, Sijtsma, Van Der Sanden, & Taconis, 2002). Furthermore, MBPAs provide the possibility to improve the representativeness of the assessment. For example, students may be presented with a multitude of situations and tasks in the MBPA compared to a PBA. To have students perform multiple tasks in multiple situations enhances the reliability of the instrument. MBPA provides the possibility to confront students with highly contextualized critical situations. Thereby, the two aspects of the representativeness of the measure, namely comprehensiveness and fidelity, are improved considerably using MBPA.

Another advantage of MBPA is that raters are ruled out of scoring. As mentioned earlier, raters are one of the most important sources of measurement error in PBA. MBPA enables test developers and administrators to automatically score student performance. Rater effects, and subsequently rater induced measurement error, can thereby be diminished.

Finally, compared to PBA, MBPA has higher efficiency. The assessment is administered virtually rather than physically. That is, physical situations where students, raters, and possible actors have to meet to perform the PBA are not needed in MBPA. Furthermore, vocational schools can simultaneously administer the MBPA in large groups. Compared to PBA, which is individually administered, MBPA is more efficient on this aspect.

To summarize, the rationale behind MBPA is not just to replicate what PBA is also capable of, but to add new features to the measurement spectrum in vocational education. Or, to quote Thornburg (1999): *The key idea to keep in mind is that the true power of educational technology comes not from replicating things that can be done in other ways, but when it is used to do things that couldn't be done without it. (p. 7).* We have hypothesized that MBPA might perform better on standardization, representativeness, reliability, rater effects, and efficiency than PBA.

2.1.3 Research on Innovations in Assessment

Research on innovations in computer-based assessment covers a wide variety of topics, from using technology in testing to detect potential fraud (Wol-

lack & Fremer, 2013) on the one end to highly interactive virtual reality assessments on the other end (Clarke-Midura & Dede, 2010). The most recent innovation that is now widely implemented and is most relevant for the current discussion is the introduction of so-called innovative (or alternative) item types. Scalise and Gifford (2006) present an overview of different innovative item types that have emerged and become available for test developers. Innovative item types sometimes incorporate multimedia (e.g., graphs or illustrations) and research has shown that these items provide test developers with the opportunity to test specific aspects of student learning (e.g., application of knowledge, inquiry) that are impossible to test with traditional multiple-choice tests (Scalise & Gifford, 2006).

Current technological progress creates unparalleled possibilities for assessment and announces a whole new era of assessments to look forward to. For example, researchers are now starting to introduce immersive virtual environments into the educational measurement field (Clarke, 2009). Originally stemming from e-learning applications (see for example Monahan, McArdle, & Bertolotto, 2008), these virtual environments simulate a more or less real-world environment which often requires highly interactive operating. Students can perform a wide variety of tasks and objectives within the virtual environment and the computer automatically records, logs, and scores student behavior.

Research on the use of multimedia in assessment is done in the field of organizational psychology as well. For example, Oostrom, Born, Serlie, and Van der Molen (2011) have done research on the use of a multimedia situational test in which the test items are presented as video clips. Furthermore, the same group of researchers (2010) has also experimented with an innovative open-ended multimedia test in which the test takers responses are recorded with a webcam. Finally, serious gaming and assessment is another multimedia influenced topic of research. For example, researchers and practitioners have designed and tested virtual manager games to assess test takers competencies (Chang, Lee, Ng, & Moon, 2003).

It is important to note that these innovations do not solely result from innovative technology. Technology enables researchers and test developers to design innovative CBAs, but technology does not determine the success of assessment innovations (Williamson, Bejar, & Mislevy, 2006). It is the combination of technology and structured design, grounded in everything we know

about assessment that results in solid and coherent assessments. Williamson et al. subtly remark that technological advances have even outpaced general assessment methods for the interpretation of scores that result from innovative CBAs. Of course, the challenge lies not in developing richly designed, highly contextualized and magnificently looking multimedia-based assessments, however the challenge lies in having them function psychometrically as well.

2.1.4 Introducing Multimedia-based Performance Assessment in Vocational Education and Training

To illustrate the added value of MBPA in VET we now present a pilot example of an MBPA that we are designing and developing. The MBPA is used to assess students' skills and abilities for a specific vocational education: *safety guard for confined spaces*. Currently, students that pass a PBA become certified safety guards. A safety guard for confined spaces ensures that confined space work is carried out responsibly and safely by workers. For example, by doing job safety analyses and maintaining an adequate communications system. Our goal is to develop an MBPA that is capable of capturing students' mastery of skills in a way that is more valid and reliable than the current PBA. The first research question therefore is:

RQ1. Based on psychometric and empirical comparison; to what extent does the MBPA perform better than the current PBA?

The design and development will take place according to the developmental framework for MBPA in VET that is presented in Chapter 4. During the research project we will mainly focus on studying the measurement properties (validity and reliability) and the efficiency of MBPA. The second research question that we try to answer in this research project therefore is:

RQ2. Using the framework referred to above; can we construct an MBPA that provides valid and reliable inferences about students' mastery of skills to be certified as a safety guard confined spaces?

Finally, based on the results of the current research project and gained experience the third research question is:

RQ3. Is MBPA both an efficient and effective assessment method in VET?

2.1.5 Multimedia-based Performance Assessment: An Example

The MBPA we are designing and developing is a simulation of the real job of a safety guard confined spaces and is based upon real work processes. Because of the simulative element we can quickly change the tasks or situations that the safety guard is virtually confronted with. The first pilot version of the MBPA has already been created. Using video clips in which a real safety guard and context is presented we provide students with a virtual environment in which they can already perform several tasks (see Figure 2.1).

Figure 2.1
Confined Space Guard in Authentic Work Environment

In the assessment, students follow the safety guard and a worker performing their tasks (guarding and cleaning a confined space) and students are required to intervene when they observe incorrect behavior of the worker performing in and around the confined space and of the safety guard self. For example, safety guards have to determine the optimal escape route in case of a hazard or a factory alarm. One aspect of an optimal escape route is the wind, and when students observe the safety guard determining the escape route incorrectly they can intervene by pressing the stop button, see Figure 2.2. Also, students have the opportunity to study the work permit during the assessment as

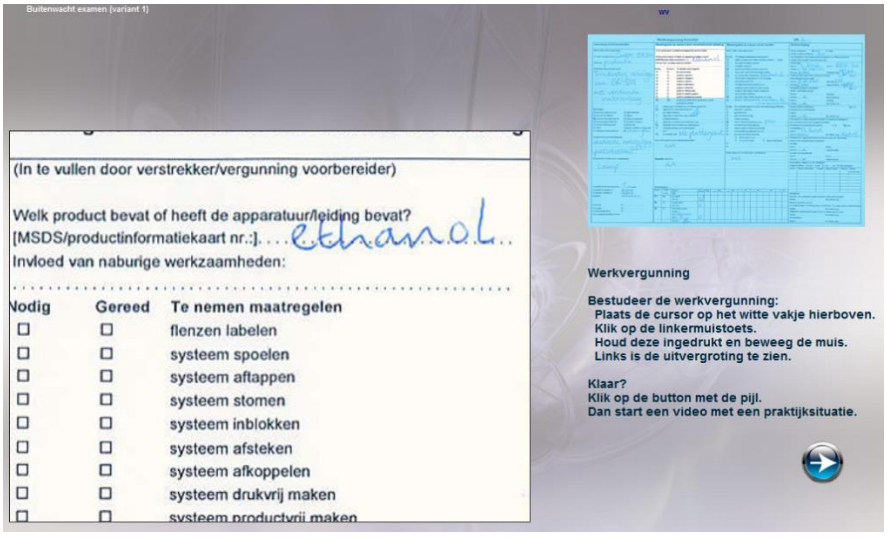
can be seen in Figure 2.3. For example, to see which gasses or substances have been in the confined space or to identify possible hazards.

Figure 2.2

Confined Space Guard Determines Optimal Escape Route

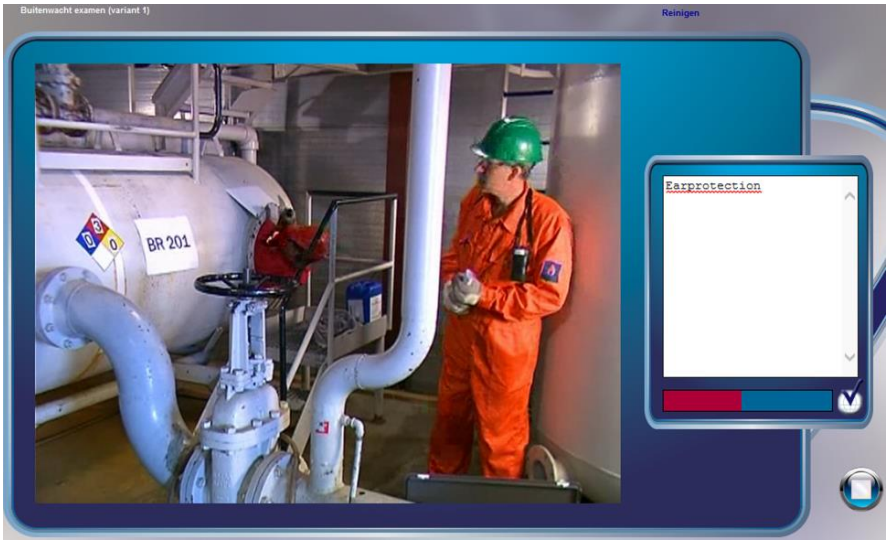
Figure 2.3

Students Can Open the Work Permit During the Assessment



Students are introduced to the assessment via a video that explains what they are going to see and what their options and tools are in the assessment. In the pilot version we log and score all interventions made during the assessment. When students intervene, a new window pops up in which they can type the incorrect behavior they have observed (see Figure 2.4). Key terms are possible to score (e.g., "ear protection" in the example) as well as sentences, but the results provided by the pilot version still need to be scanned by a rater. Thus, students have to observe and decide about erratic behavior displayed by either the worker or the safety guard and intervene when they do see that happening. All interventions and student reactions are recorded and provide an observation about students' mastery of skills to work as a safety guard.

Figure 2.4
Intervention Made by Student



2.2 Method

With co-funding from the Foundation Cooperation for Safety (SSVV) we are currently working on the design and development of an expanded version of the MBPA pilot version for safety guards confined spaces. The new assessment incorporates more multimedia elements and a higher amount of interactivity between student and assessment. Above that, the MBPA should be fully

independent of raters; all actions of students have to be logged and scored automatically. Of course, the point of departure for implementing technological improvements in an MBPA should always be improved measurement of students' mastery of skills. That is, with these new technologies we try to produce better observations about student learning than is possible with traditional measurement methods (e.g., PBA). For example, the PBA is limited in the amount of situations and tasks a student can perform but in the MBPA we can confront students with a multitude of tasks and situations In addition, we can provide students with tools (e.g., a work permit, communication set, and measurement instruments) and continually update their status and the information in the MBPA as they progress through the MBPA. Furthermore, we are able to log and save everything that the student does in the virtual environment for later (psychometric) analysis.

To answer the research questions posed above, we plan to conduct several studies based upon the MBPA for safety guards confined spaces that we are developing. First, we will be studying the psychometric functioning of the assessment using the versatile data that the MBPA produces. One of the challenges in using MBPA in certification settings lays not so much in the design and development of MBPA, but in the psychometric analysis of the data that it produces. For example, which responses or variables in the assessment provide evidence about students' mastery of skills? We will fit different psychometric models to the data (e.g., Bayesian networks), and we will determine the MBPA's reliability.

Secondly, we want to perform a comprehensive validity study based on Kane's (1992) argument-based approach to validation. Validity is probably the most central concept in assessment (Messick, 1989). The argument-based approach to validation emphasizes the evaluation of the plausibility of the various assumptions and inferences involved in interpreting assessment observations as a reflection of students' mastery of skills. Of course, psychometric functioning and reliability are part of the assessments' overall validity but we also want to include an empirical comparison of the MBPA and PBA as a validity argument. In an empirical study, students will either first do the PBA and then the MBPA or the other way around. We will then analyze results from both assessments and report our findings in a later chapter of this thesis.

2.3 Discussion and Conclusion

Technology provides unparalleled opportunities in assessment. Now, assessment developers, practitioners and researchers may seize the opportunity to develop and study a new type of assessment in vocational education: multimedia-based performance assessment. This is a valuable endeavor because technological possibilities are ahead of psychometrics. Furthermore, current assessment methods in VET may not be sufficient to validly and reliably measure every aspect of students' mastery of skills. We have shown that time and resources make it difficult to design and develop PBAs that provide valid and reliable inferences. MBPA might provide a solution to this problem.

In contrast to PBA, MBPA is fully standardized which reduces measurement error in measuring students' skills. Using MBPA it is also possible to present more situations and tasks than in PBA, which implicates improved representativeness of the MBPA compared to the PBA. Improved representativeness also results from the possibility to incorporate high risk tasks and infrequent tasks in an MBPA. Furthermore, one of the major causes of construct irrelevant variance in PBA, raters, can be ruled out of the scoring process. Together, this may result in more valid and reliable MBPA scores compared to PBA scores. Finally, the feasibility of MBPA may be higher than PBA because there is no need for printed materials and personnel (actors, raters, etc.). MBPA provides the opportunity of large scale and remote administration too.

The main goal of the research project presented in this Chapter is to investigate the overall validity, reliability and feasibility of MBPA. We are trying to answer the question whether MBPA is both an effective and efficient method of assessment in VET. Currently, we look at MBPA as a promising assessment method for the assessment programs of most qualifications in VET. However, we also expect that in a first stage MBPA will complement rather than replace the traditional measurement methods in VET. Research should first point out how to use, psychometrically speaking, the data that MBPA produces. Furthermore, validation studies are needed to determine the practical value of MBPA in an educational setting. Thus, although MBPA is full of promise, there is still a lot of work to be done before actual large-scale implementation can take place.

References

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

Bakx, A.W.E.A., Sijstma, K., Van Der Sanden, J.M.M., & Taconis, R. (2002). Development and evaluation of a student-centerd multimedia self-assessment instrument for social-communicative competence. *Instructional Science, 30*, 335-359.

Bartram, D. (2006). Testing on the internet: issues, challenges, opportunities in the field of occupational assessment. In D. Bartram and R.K. Hambleton (Eds.), *Computer-based Testing and the Internet* (pp. 13-37). Chichester: Wiley.

Baxter, G.P., Shavelson, R.J., Herman, S.J., Brown, K.A., & Valadez, J.R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education, 24*, 190-216.

Bejar, I.I., Williamson, D.M., & Mislevy, R.J. (2006). Human scoring. In D.M. Williamson, R.J. Mislevey & I.I Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49-81). Mahwah, NJ: Lawrence Erlbaum.

Bennett, R.E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram and R.K. Hambleton (Eds.), *Compouter-based Testing and the Internet* (pp. 201-217). Chichester: Wiley.

Breland, H. (1983). *The direct assessment of writing skill: A measurement review* (College Board Report No. 83-6). New York: College Entrance Examination Board.

Brennan, R.L. (1983). *Elements of generalizability*. Iowa City, IA: American College Testing Program.

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

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Chang, J., Lee, M., Ng, K., & Moon, K. (2003). Business Simulation Games: The Hong Kong Experience. *Simulation & Gaming*, *34*, 367-376.

Clarke, J. (2009). *Studying the potential of virtual performance assessment for measuring student achievement in science.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.

Retrieved September 5, 2012, from http://virtualassessment.org/publications/aera_2009_clarke.pdf

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

Clauser, B.E., Clyman, S.G., & Swanson, D.B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement, 36*(1), 29-45.

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *Research in Learning Technology, 13(1)*, 17-31.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: John Wiley.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

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

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M (2014). *A framework for designing and developing multimedia-based performance assessment.* 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.

Dierick, S., & Dochy, F.J.R.C. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*, 307-329.

Drasgow, F., Luecht, R.M., & Bennett, R.E. (2006). Technology and testing. In R.L. Brennan (Ed.), *Educational Measurement* (pp. 471-530). Westport, CT: Praeger.

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.

Eckes, T. (2005). Examining rater effects in TestDaf writing and speaking performance assessments: A many-facet Rasch analysis. *Language assessment quarterly, 2*(3), 197-221.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599-635.

Fitzpatrick, R., & Morrison, E.J. (1971). Performance and product evaluation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 237-270). Washington DC: American Council on Education.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science. Promises and problems. *Applied Measurement in Education, 7*, 323-334.

Grégoire, J. (1997). Diagnostic assessment of learning disabilities. From assessment of performance to assessment of competence. *European Journal of Psychological Assessment, 13*(1), 10-20.

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

Haertel, E.H., Lash, A., Javitz, H., & Quellmalz, E. (2006). An instructional sensitivity study of science inquiry items from three large-scale science examinations. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Issenberg, S.B., Gordon, M.S., Gordon, D.L., Safford, R.E, & Hart, I.R. (2001). Simulation and new learning technologies. *Medical Teacher, 23*(1), 16-23.

Kane, M.T. (1992). The assessment of professional competence. *Evaluation & the health professions, 15*(2), 163.

Ketelhut, D. J., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (2007). Studying situated learning in a multi-user virtual environment. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 37–58). Mahwah, NJ: Lawrence Erlbaum.

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

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

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

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

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

Monahan, T., McArdle, G., & Bertolotto, M. (2008). Virtual reality for collaborative e-learning. *Computers & Education, 50*, 1339-1353.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology, 19*(5), 532-550.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10*(2), 78.

Roelofs, E.C., & Straetmans, G.J.J.M. (Eds.) (2006). *Assessment in actie* [Assessment in action]. Arnhem: Cito.

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

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved [March 20, 2012] from http://www.jtla.org.

Schoech, D. (2001). Using video clips as test questions: The development and use of a multimedia exam. *Journal of Technology in Human Services, 18*(3-4), 117-131.

Segers, M. (2004). Assessment en leren als twee-eenheid: onderzoek naar de impact van assessment op leren [the dyad of assessment and learning: a study of the impact of assessment on learning]. *Tijdschrift voor Hoger Onderwijs, 22*, 188-220.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory*. Newbury Park, CA: Sage.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

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

Shavelson, R.J., Ruiz-Primo, M.A., & Wiley, E. (1999). Note on sources of sample variability in science performance assessments. *Journal of Educational Measurement, 36*(1), 56-69.

Sliney, A., & Murphy, D. (2011). Using Serious Games for Assessment. In M. Ma, A. Oikonomou & L. C. Jain (Eds.), *Serious Games and Edutainment Applications* (pp. 225-243). London: Springer.

Straetmans, G.J.J.M., & van Diggele, J.B.H. (2001). *Anders opleiden, anders toetsen* [Different instruction, different assessment]. BVE-brochurereeks: Perspectief op Assessment, deel 1 [BVE-brochure series: Perspective on Assessment, part 1]. Arnhem: Cito.

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games – an overview [Technological Report]. Retrieved from http://www.his.se/PageFiles/10481/HS-IKI-TR-07-001.pdf

Sweet, D., & Zimmermann, J. (1992). Performance Assessment. *Education Research Consumer Guide*(2), 2-5.

Thornburg, D.D. (1999). *Technology in K-12 education: Envisioning a new future*. Retrieved October 16, 2012, from http://www.edtech.ku.edu/resources/portfolio/examples/nets/Miller/www.air.org/forum/Thornburg.pdf.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29.

Van Dijk, P. (2010). *Examinering in de beroepspraktijk* [Assessment in vocational practice]. Amersfoort: ECABO.

Van der Vleuten, C.P.M., & Swanson, D.B. (1990). Assessment of clinical skills with standardized patients: The state of the art. *Teaching and Learning in Medicine, 2*, 58-76.

Webb, N.M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13*(3), 277-301.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. Houndmills, England: Palgrave Macmillan.

Williamson, D.M., Bejar, I.I., & Mislevy, R.J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D.M. Williamson, R.J. Mislevey & I.I Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 1-13). Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E.W., & McVay, A. (2010). *Rater effects as a function of rater training context*. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/6435A0AF-0C12-46F7-812E-908CBB7ADDFF/0/RaterEffects_101510.pdf

Wollack, J.A., & Fremer, J.J. (Eds.) (2013). *Handbook of test security*. New York, NY: Routledge.

Ziv, A., Small, S.D., & Wolpe, P.R. (2000). Patient safety and simulation-based medical education. *Medical Teacher, 22*(5), 489-495.

Chapter 3. Psychometric Analysis of the Performance Data of Simulation-based Assessment: A Systematic Review and a Bayesian Network Example[3]

Abstract

Researchers have shown in multiple studies that simulations and games can be effective and powerful tools for learning and instruction (cf., Mitchell & Savill-Smith, 2004; Kirriemuir & McFarlane, 2004). Most of these studies deploy a traditional pretest-posttest design in which students usually do a paper-based test (pretest) then play the simulation or game and subsequently do a second paper-based test (posttest). Pretest-posttest designs treat the game as a *black box* in which something occurs that influences subsequent performance on the posttest (Buckley, Gobert, Horwitz, & O'Dwyer, 2010). Less research has been done in which game play product data or process data itself are used as indicators of student proficiency in some area. However, the last decade researchers have started focusing on what is happening inside the black box to an increasing extent and the literature on the topic is growing. To our knowledge, no systematic reviews have been published that investigate the psychometric analysis of performance data of simulation-based assessment (SBA) and game-based assessment (GBA). Therefore, in Part I of this chapter, a systematic review on the psychometric analysis of the performance data of SBA is presented. The main question addressed in this review is: 'What psychometric strategies or models for treating and analyzing performance data from simulations and games are documented in scientific literature?'. Then, in Part II of this chapter, the findings of our review are further illustrated by presenting an empirical example of the – according to our review – most applied psychometric model for the analysis of the performance data of SBA, which is the Bayesian network. Both the results from Part I and Part II assist future research into the use of simulations and games as assessment instruments.

[3] This chapter is a minor revision of De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2014). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education, 85*, 23-34.

3.1 General Introduction

The use of computer simulations and games as assessment instruments (from here on referred to as *simulation-based assessment* (SBA)) has increased in popularity in the preceding years. The general rationale is that SBA has some advantages over traditional paper-and-pencil (P&P) tests and performance-based assessments (PBA) and that it can both expand and strengthen the domain of assessment (Clarke-Midura & Dede, 2010; De Klerk, Eggen, & Veldkamp, 2014). First, from the student's point of view, doing an SBA is more fun and entertaining than doing a paper based test. The storyline driven approach of SBA tends to induce *flow* (Csikszentmihalyi, 1990), which is a psychological state in which people lose perception of time and space. Effectively, students are immersed in the SBA when they experience flow. This may also mean that students are highly motivated and dedicated to completing tasks and attaining goals in the simulation while not being preoccupied by test anxiety (Shute, Masduki, Donmez, Dennen, Kim, Jeong, & Wang, 2010). On the other hand, other students may find it difficult to immerse themselves in a virtual environment, or may get confused with the construct-irrelevant aspects of the simulation (e.g., the interface or specific colors). If so, the use of SBA might have serious implications for some students, especially in high-stakes testing situations. Getting students accustomed to simulations, for example during schooling, is often suggested to overcome these possible negative effects of the use of SBA.

Secondly, SBA provides the possibility to place more emphasis on the application of knowledge in highly contextualized environments rather than the replication of knowledge as is usually the case in P&P tests. For example, through the design and use of interactive tasks in an SBA, WestEd researchers were able to improve measurement of the *conducting inquiry* science practice in middle school (Quellmalz, Davenport, Timms, DeBoer, Jordan, Huang, & Buckley, 2013). Other researchers have even started to investigate the possibility to use SBA for very practical professions, for instance medical and security professions (Mislevy, Steinberg, Almond, Russell, Breyer, & Johnson, 2001; Iseli, Koenig, Lee, & Wainess, 2010). With technological possibilities improving on a steady pace, quite possibly the assessment of practical/manual skills or at the least procedural/strategic skills through SBA will become more of a common practice in the future. Again, this specific advantage of SBA might also

have a negative counterpart. For learning or formative assessment purposes, a student can develop knowledge, skills, and abilities (KSAs) within a specific, contextualized virtual environment, which means that the KSAs are grounded in deep, specific experiences associated with the environment(s) presented in the simulation. Yet, in a high-stakes testing situation, the use of a contextualized environment induces low generalizability of the students' performance. In fact, an SBA for a summative assessment purpose might best be composed of different modules, based on different contextualized environments and tasks.

Thirdly, SBA offers the possibility to capture student's product data as well as their process data. Product data can be regarded as the final work products that students produce during the SBA, while process data are log file entries that indicate *how* student's produced their work products (Rupp, Nugent, & Nelson, 2012). Process data can be very useful for a formative or diagnostic purpose but they can also serve as a source of evidence for a summative purpose. The amount of process data can become very large as the time spent in the simulation increases. Students interacting with an SBA for some time may produce many pages of process data, which may be interesting to analyze for measurement purposes. Of course, not all process data is relevant for the statements that we want to make about a student's proficiency in the construct to be measured. Identifying the elements in the process data that are relevant for measurement and synthesizing and combining those elements with students' product data into a coherent psychometric model reflects one of the major advantages and challenges of using simulations and games as assessment instruments.

3.1.1 Evidence-centered Design

Above, we have discussed some advantages of using SBA and possible negative washback of using SBA in high-stakes testing situations. Another challenge for using SBA lies in defining and specifying coherent and complete psychometric models that fit the data that students' performance in SBA's produces. A useful point of departure for this discussion is the *conceptual assessment framework* (CAF) layer within the *evidence-centered design* framework (ECD) (Mislevy, Almond, & Lukas, 2004). The CAF consists of three separate, though strongly related, models: the *student model*, the *activity model* and the *evidence model*.

The student model relates to what we want to measure, it specifies one or more constructs that we are interested in and want to make statements about regarding students' proficiency. In ECD terms these constructs are called Student Model Variables (SMVs) and they are latent, which means that we cannot directly observe them and have to make inferences about these variables based on the observable variables produced by the performance of students in the SBA (Mislevy et al., 2004). Student models can easily become highly complex in SBA as it is often the case that multiple constructs at the same time define the performance of students in the SBA.

The activity model relates to how an SBA's situations and tasks are designed in which we measure what we want to measure. The activity model consists of all the tasks that are part of the SBA. In SBA, tasks are commonly specified as objectives or goals that students have to achieve during their performance in the simulation. In that sense, tasks in SBA's are often different from traditional item – response question formats that are common practice in traditional tests. In traditional assessment tasks, the activity model variables and values are already known to the assessment developer before the test is presented to the student. For example, a computer-based test consisting of 50 multiple-choice with three alternatives that can all be scored dichotomously (0 = incorrect, 1 = correct), and in which student responses are recorded by mouse clicks.

In general, SBAs have some variables and values of the activity model that are known in advance for every student progressing through the assessment, while others are not. Elements that are known, for example, are the interface or a specific situational feature that is the same for every student. Yet, as students are progressing through the simulation, the simulation may in some cases evolve into different states for different students. In that case, the game condition variables may change, also between students, including the rules, possible actions and interactions that are possible at that specific moment in the SBA. Mislevy et al. (2014) call this the state machine of the SBA. These *dynamic* activity model variables make it more difficult to psychometrically model and interpret a students' performance, because the actual values in the dynamic activity model can only be known and operationalized in the perspective of the state machine.

Building on the activity model, multiple sources of data are recorded and collected during a student's performance in the SBA. For example, reaction times, mouse clicks, navigational paths, or successful completion of objectives. Some, but not all, of these data will function as observable variables (OV) that provide information about SMVs through a measurement model. Which pieces of data can be identified as OV and how these pieces accumulate into a coherent measurement model is specified in the evidence model. The evidence model relates to how we measure what we want to measure. Theory and data are united in the evidence model through two separate, though strongly related processes: *evidence identification* and *evidence accumulation* (Rupp et al., 2012). The supposed theoretical relationship between SMVs and OVs are formalized in the evidence model on basis of the data that are produced by students performing in the SBA. As mentioned earlier, simulations and games offer the opportunity to record both product and process data which are subsequently saved into a log file. Process data are usually click-stream data (mouse-clicks, navigation paths, use of tools, etc.) that indicates what action students performed during the SBA to produce product data. In the evidence identification part of the evidence model the product and process data that are relevant for the statements that we wish to make about students' SMVs are identified. There might be multiple steps involved before the specific elements of data in the log files can be considered to take the role of OV. For example, if a simulation also records spoken statements of a student, then these statements first have to be transcribed, analyzed and scored before these performance data elements can be used in a psychometric model. Thus, in some cases, there needs to be a data reduction process from raw log file data to manageable data that can serve as input to the measurement model. Only then, these specific elements of data can take the role of OV and accumulate into a measurement model. In the measurement model the OVs are being weighed and aggregated to transform them into a final score or profile regarding one or more SMVs.

Next, a systematic review on the psychometric analysis of the performance data of simulation-based assessment will be presented in the first part of this chapter. In the second part of this chapter, the results of the systematic review will be illustrated in depth by providing an example of a Bayesian network.

PART I – A Systematic Review on the Psychometric Analysis of the Performance Data of Simulation-based Assessments

3.2 Introduction

In this systematic review, we specifically focus on the psychometric analysis of the performance data of SBA (i.e., the evidence model), although the student model and activity model are also part of our analysis. Studies reporting on the psychometric analysis of SBA are being published to an increasing degree in the preceding years. As the body of literature on the topic is growing and researchers from multiple institutions and universities are increasingly focusing on measurement in simulations and games, now may be a good moment to evaluate the current position of the field, identify common practices, and provide directions for future research. That is exactly the purpose of this systematic review.

It is also interesting to remark that the scientific database consists of thousands of articles on the use of simulations and games, mostly for learning, but that so few actually use psychometric modeling on the data that performing the simulation produced. There might be a few reasons why this is the case. First, it is quite a challenge to psychometrically model the very versatile data of SBA, which might keep some researchers at a distance. Secondly, the methods and statistical techniques for the analysis of complex data sets, in combination with computing power and software, have only recently improved to such an extent that psychometric analysis is more accessible. Thirdly, simulations and games are often used for (semi-)commercial purposes in which the need for psychometric modeling is not always there, this also holds for many of the simulations used in personnel selection for example. Finally, and as mentioned before, the simulation is often seen and used as a black box. The educational gains of the simulation are then measured in a pretest-posttest design and researchers do not look into the actual log file data from the simulation.

3.3 Material and Methods

3.3.1 Procedure

The procedure for this systematic review is based on the method provided by Petticrew and Roberts (2006). The review process includes three steps; collecting literature, processing and analyzing the content of the literature and drawing conclusions based on the literature's analysis (see also Van der Kleij, Timmers, & Eggen, 2011). In the first step, the research question is defined, the databases and search terms are chosen, the literature search is carried out, the inclusion criteria are defined, and, eventually, the studies found in the literature search are selected using the inclusion criteria. As a first addition to the traditional search and selection process we have also performed the *snowballing* technique (Doust, Pietrzak, Sanders, & Glasziou, 2004). Snowballing indicates that we have thoroughly searched through all the references of the first set of literature to identify more relevant literature for our systematic review. As a second addition to the traditional search and selection process we have also sent all the references to the most cited first authors (N = 12). We asked them to scan through the references and to indicate if we had missed any key publications of the author's or their peers.

In the second step, the characteristics of the studies found are analyzed and processed in such a way that studies can be compared. In the third and final step, the conclusions of the different studies are brought together and synthesized using a qualitative method. This implies that a systematic summary of the literature will be given which results in a final conclusion regarding the best methods to treat the performance data of SBA. Additionally, after the review process, we will demonstrate an empirical application of a psychometric model – the BN – that, according to our review of the literature, seems very suitable for the analysis of performance data from SBA.

3.3.2 Databases and Search Terms

We have searched three scientific databases: Scopus, ERIC and Google Scholar[4]. These are the most commonly used databases for educational research. We also searched Google Web to find so-called *grey literature* which is literature that is not published in scientific journals but may contain interesting information. Especially in the new and evolving field of assessment in simula-

[4] The literature search was performed in February 2014.

tions and games, grey literature may provide added value to the systematic review as multiple research programs are currently running. We identified 20 search queries with which we expected to find all relevant literature. The terms within the search queries were only connected with the "AND" command and included terms related to the use of simulations (e.g., simulation, assessment, psychometric) games (e.g., game, educational assessment) and innovative technology in assessment (e.g., innovative, assessment, psychometric) and psychometrics (e.g., serious game, psychometric model). All search queries can be found in Table 3.1. For example, the first query would be entered as "serious game AND psychometric" in the 4 databases (Scopus, ERIC, Google Scholar, and Google Web). In this way, we performed a total of 80 searches – 20 search queries in 4 databases.

3.3.3 Inclusion Criteria

Clear inclusion criteria are needed to select the right studies from the bulk of studies that result from entering the search queries in the different databases. The first inclusion criterion was that the study was in English language. This criterion is because of practical reasons and accessibility. Because we only used English search queries we did not find any non-English studies. The second criterion was that the study has been published by a research organization or researcher from a research organization in a research report, scientific journal, book chapter, dissertation or conference proceedings. By using the second criterion we aimed to include as many publications as possible that are linked to research organizations (e.g., universities or private and public research institutes). Although it is often not clear if and how research reports and dissertations have been reviewed, we found that we had to include them because of the limited amount of studies that are published on this relatively new and evolving topic. The third criterion was that the study either presents a computer-based simulation or game in an assessment context or discusses the topic. The third inclusion criterion ensured that we included articles that discussed simulations and games that were presented on a computer rather than in real life as is often the case in performance-based assessment. The fourth criterion was that the main goal of the study is to qualitatively or quantitatively (empirically) discuss the psychometric analysis of the data that simulation or game performance produces. By using the fourth and final inclusion criteria studies that did not

discuss psychometric functioning of the simulations and games were excluded from further analysis.

Table 3.1
Search and selection results based on database searches and snowballing technique

	Scopus		ERIC		Google		Snowballing			
Search query	N_{hits}	$N_{relevant}$	N_{hits}	$N_{relevant}$	N_W	N_S	N_{Scopus}	N_{ERIC}	N_W	N_S
Serious game, psychometric	1	0	1	0	2	3	0	0	0	2
Serious game, psychometric model	1	0	61	4	0	2	0	0	0	1
Serious game, assessment	225	12	16	1	6	1	16	0	5	13
Serious game, measurement	53	1	4	0	1	0	0	1	1	0
Serious game, data-mining	13	0	2	1	7	0	0	0	7	0
Serious game, performance analysis	65	0	636	3	0	0	0	0	0	0
Serious game, evaluation, performance	50	1	4	0	0	0	0	0	0	0
Simulation, assessment, psychometric	149	2	63	3	1	0	11	1	1	0
Simulation game, assessment, psychometric	2	2	2	0	1	4	0	1	0	5
Simulation, assessment, data-analysis	1224	25	78	2	0	0	0	13	0	0
Technology-based assessment, psychometric	1	1	13	0	0	1	0	0	0	0
Technology-based assessment, measurement	84	2	6	1	0	0	0	0	0	0
Educational video game, data-analysis	5	1	19	1	2	0	0	3	1	0
Educational video game, performance data	19	1	11	1	2	0	0	1	1	0
Testing, technology, psychometric	99	2	123	1	0	0	0	0	0	0
Automated scoring, complex, tasks	21	3	1617	30	3	0	5	38	1	0
E-assessment, performance	51	1	61	1	0	0	0	0	0	0
Innovative, assessment, psychometric	62	2	19	0	0	0	4	0	0	0
Innovative, computer-based assessment, data-mining	0	0	0	0	1	0	0	0	0	0
Educational assessment, game	377	5	624	15	2	0	3	0	2	0
Total	2502	61	3410	64	28	11	39	58	19	21

N_{hits}=Number of hits from search query, $N_{relevant}$=Number of relevant hits after title-abstract scan, N_W=Number of relevant hits from Google web search (first 150 results), N_S=Number of relevant hits from Google scholar (first 150 results). N under snowballing is number of relevant articles from references scan of all previous articles.

3.3.4 Selection Process

All references were imported in Thomson Reuters EndNote X7 (2013). Using EndNote we were able to eliminate all duplicate studies that were found during the database searches. Furthermore, we found that some articles were untraceable and had to be deleted from the list of references. Next, using the inclusion criteria, the studies were successively screened based on their relevance for the current review. The first author carried out the selection process and started with applying the criteria on title and abstract of all papers. If it was not clear whether an article complied with a criterion or not in this round it was included for the next round. The articles that remained were then subjected to a thorough scan through the total paper. Again, the inclusion criteria were applied. After this scan the final set of articles that were used in this review remained.

3.4 Results

3.4.1 Search and Selection

The 20 search queries carried out for the literature search resulted in 2502 hits in Scopus and 3410 hits in ERIC. Because the searches in these two databases yielded many results already we only searched the first 150 results from the Google Scholar and Google Web search. For all databases together, this resulted in a total of 6212 hits (see also Table 3.1). All studies were subjected to the inclusion criteria described above.

In the first selection round the titles and abstracts of the studies found, were read and the inclusion criteria were applied. In this way were able to eliminate a lot of articles, respectively, 2441 (Scopus), 3346 (ERIC), 122 (Google Web), and 139 (Google Scholar). This way we had a set of 164 articles. We then performed the snowballing technique on these articles (title and abstract scan) which resulted in the addition of another 137 articles. In total we now had 301 articles in our EndNote database. Using EndNote it was relatively easy to remove duplicate studies as a second selection step. The total amount of articles that remained was 196 (second selection round). Unfortunately, we could not retrieve all articles which resulted in the elimination of another 26 studies as a third selection round. The 170 articles that remained were retrieved and all of these articles were completely scanned through. Applying our selection criteria again after the total article scan resulted in the elimination of another 122 publi-

cations (fourth selection round). The references of the 48 publications that remained were exported to a Microsoft Word file and were then sent to the most cited first authors in the list (N = 12). As mentioned earlier, we asked them to scan through the references and to indicate if we had missed any key publications of the author's or their peers in our literature search. A total of 8 (67%) first authors responded on our request and in that way we were able to add another 5 publications that our preliminary literature searches had not yield.

We now had a set of 53 articles, which we read and analyzed carefully. Based on this process we were able to exclude another 22 papers (fifth selection round). These publications were excluded for several reasons. First, several publications were too similar to distinguish them as two separate articles in the review. For example, the psychometric analysis of performance data of a simulation-based assessment was discussed that was already discussed in another publication. Secondly, for some publications we were not able to assess the quality of the analyses because the authors were too limited in their discussion of the analyses used. Therefore, the systematic review was performed on a total of 31 publications. In Table 3.1, the search and selection results based on the search queries in the databases and the snowballing technique can be seen. In Table 3.2, the results from the different rounds of selection can be seen.

Table 3.2
Results of Selection Rounds

Selection round	*Studies subjected*	*Studies excluded*	*Percentage excluded*	*Studies selected*
1	6212	6048	97%	164
2	301	105	35%	196
3	196	26	13%	170
4	170	122	71%	48
5	53	22	43%	31

Note. The studies selected and subjected between rounds 1 – 2, and 4 – 5 do not match because studies were actually added because of our two additional search strategies: snowballing and first author consult.

3.4.2 Content Analysis

The content analysis took place on basis of the 31 studies in Table 3.3. We used three perspectives to analyze the papers that our searches had yielded. This was done to make sound comparisons between the simulation-based assessments that were presented in the articles. The three categories are the student characteristics that were being measured with the simulation (the SMVs), the performance data on which the study focused (the OVs), and the psychometric models or statistical techniques used to analyze the performance data.

3.4.3 Student-model Variables

The first perspective is the SMV, i.e., 'what' the simulation intends to measure. Often, SMVs are latent constructs that we make inferences about based on students' observable performance in the tasks that are being administered in the SBA. Sometimes SMVs are practical skills (e.g., operating a machine) in which performance can be directly observed from students' physical actions. The largest proportion of the studies (N = 21) in our literature review focused on the former type of SMVs – latent (cognitive) constructs –, although some of the simulations presented in the articles (N = 10) also aimed at measuring a more practical skill oriented construct (see also Table 3.3).

Several studies focused on measuring mathematic skills through the analysis of performance data of math games like *Save Patch* (Buschang, 2012; Kerr, 2012; Kim, 2012; Levy, 2014) and *Math Garden* (Klinkenberg, 2011). Another set of studies analyzed student performance data in simulations intended to measure so-called 21st century skills as systems thinking (*SimCityEDU: Pollution Challenge!* (Mislevy, 2014); *Taiga Park* (Shute, 2010)), creative problem solving (*Oblivion* (Shute, 2009)), or causal reasoning (*World of Goo* (Shute, 2011). Currently, much research is also being done on the use of simulations as assessments of science, physics, and biology concepts in middle school and high school. Among other simulations, *SimScientists* (Quellmalz, 2012, 2013), *Virtual Performance Assessment* (Clarke-Midura, 2010), *ScienceAssistments* (Gobert, 2012), *Newton's Playground* (Shute, 2013), and *BioLogica* (Buckley, 2010) focus on assessing science concepts.

Table 3.3
Results of 31 articles for psychometric analysis of performance data

First author and publication year	*Name of simulation / game*	*OV data*	*NMVs*	*Analyses*
Braun (2006)	Architect Registration Examination	Product/process	Architectural competence	CTT/IRT
Buckley (2010)	BioLogica	Process	Genetics	EDM
Buschang (2012)	Save Patch	Process	Math	EDM
Clarke-Midura (2010)	Virtual Performance Assessment	Product	Science inquiry	GT
Gobert (2013)	Science Assistments (Inq-ITS)	Process	Science inquiry	EDM
Halverson (2012)	Progenitor X	Process	Stem cell manipulation	EDM
Iseli (2010)	Navy damage control	Product	Damage control	DBN
Kerr (2012)	Save Patch	Process	Math	CA
Kim (2012)	Save Patch	Product	Math	SA
Klinkenberg (2011)	Math Garden	Product/process	Math	S-A/IRT
Koenig (2010)	Navy damage control	Product/process	Damage control	DBN
Lamb (2014)	Mission Biotech	Product/process	Science inquiry	EFA/CFA/IRT/AMP/ANN
Levy (2004)	Net2ASS	Product/process	Computer network skills	BN
Levy (2013)	Packet Tracer	Product/process	Computer network skills	BN/DCM
Levy (2014)	Save Patch	Product/process	Math	DBN
Margolis (2006)	Medical Case Simulations	Product	Physician practice	GT
Mislevy (2001)	Dental Interactive	Product	Dental practice	BN
Mislevy (2014)	SimCityEDU: Pollution Challenge!	Product/process	Systems thinking	EDM/BN
Porayadas (2007)	Air Combat	Product/process	Air combat skills	DBN
Quellmalz (2010)	Mountain Rescue / Fish World	Product	Science inquiry	CTT/IRT
Quellmalz (2012)	SimScientists	Product/process	Science inquiry	IRT
Quellmalz (2013)	SimScientists	Product	Science inquiry	GT/CFA/MIRT
Rupp (2010)	Urban Science Simulation Corporation	Process	Urban planning	ENA
Rupp (2012)	Packet Tracer	Product/process	Computer network skills	EDM/BN/DCM
Shute (2009)	Oblivion	Product	Creative problem solving	BN
Shute (2010)	Taiga Park	Product/process	Systems-thinking	BN
Shute (2011)	World of Goo	Product	Causal reasoning	BN
Shute (2013)	Newton's Playground	Product/process	21st century skills	BN
Stevens (2006)	True Roots (TMMEX)	Product/process	Genetics	ANN
Vendlinski (2002)	Hazmat (TMMEX)	Product/process	Chemistry	MM
West (2010)	Packet Tracer	Product	Computer network skills	BN

AMP = Attribute Mastery Pattern, ANN = Artificial Neural Network, CA = Cluster Analysis, CFA = Confirmatory Factor Analysis, CTT = Classical Test Theory, (D)BN = (Dynamic) Bayesian Network, DCM = Diagnostic Classification Model, EDM = Educational Data Mining, EFA = Exploratory Factor Analysis, ENA = Epistemic Network Analysis, GT = Generalizability Theory, (M)IRT = Multidimensional Item Response Theory, SA = Survival Analysis, S-A = Speed-Accuracy

Another group of simulations is more oriented toward the assessment of practical skills. For example, several researchers have analyzed student performance in the *Packet Tracer* digital learning environment of the *Cisco Networking Academy* (e.g., West, 2010, Rupp, 2012; Levy, 2013). The SMV in this simulation can best be defined as being skilled in designing and building computer networks in an office setting. Yet other more practical skill oriented simulations focus on architectural competence (*Architect Registration Examination* (Braun, 2006)), flood and fire damage control aboard navy ships (*Navy damage control simulation* (Iseli, 2010; Koenig, 2010)), physician practice (*Medical Case Simulations* (Margolis, 2006)), dental practice (*DISC* (Mislevy, 2001)), or urban planning skills (*Urban Science Simulation Corporation* (Rupp, 2010)). Although these simulations have a focus on practical skills, the objection can be made that for a large part these simulations focus on the *procedural knowledge* aspect of the skill under measurement. In the *Navy damage control simulation*, for instance, the student can take on the role of firefighter. When the student encounters a fire in the simulated Navy ship he has to identify the type of fire and how to respond to that fire via a report interface in the simulation. Of course, students' input yields interesting information about their *knowledge* on different types of fires and *procedural knowledge* about how to act in the situation, but what a person would actually *do* in a real life situation remains uncertain.

3.4.4 Observable Variables

The second perspective is the type of data that are used as OVs, or as indicators of proficiency. Performing a simulation-based assessment can produce loads of data in short time because every 'action' from the student in the simulation can be logged. Every action or combinations of actions can either be regarded as process or product data. In general, students perform multiple actions during the simulation to arrive at a given (final) point in the simulation (the end of a 'level' for example) or to create a product in the simulated environment. The actions that contribute to the end state or the product are called process data, whereas the end state itself or the product are referred to as the product data. In general, performance in a simulation-based assessment yields more process data than product data, and the (statistical) techniques used to analyze both forms of data differ. Process and product data may also provide different kinds of information about students. For example, process data can be

very useful for diagnostic purposes while product data serves well as an indicator of overall proficiency.

Our review of the literature yielded 10 studies that focused on product data, 6 studies that focused on process data, and 15 studies that focused on a combination of both (Table 3.3). An example of the first category is a study performed by Quellmalz et al. (2013). Their SimScientists simulation was designed for measuring three science concepts in middle school science: identifying science principles, using science principles and conducting science inquiry. In three modalities (static, active, and interactive) students performed different science tasks in the simulation. The researchers used the final settings in the tasks or the answers students gave to traditional items during the simulation for further analyses, which can all be regarded as product data. In contrast, Kerr, Chung and Iseli (2011) focused on process data in a mathematics video game. For their analyses they did not use students' final solutions but captured process data as students were playing the simulation to identify the number of attempts and the types of errors made by the students. Rupp et al. (2012) present a study in which they use a combination of product and process data. The simulation they use for their analyses is the *Packet Tracer* digital learning environment of the *Cisco Networking Academy*. In this environment, students can learn to design, configure, and troubleshoot computer networks in an interactive interface. The product data are the final configurations that students 'hand in', while process data are detailed log file entries that indicate *how* students arrived at their final configurations.

Reviewing the studies, we would strongly advise to use a combination of both process data and product data for evaluating student performance. While the performance data indicate whether students attained the desired level of competence, the process data not only indicate if they performed the correct steps to reach the final product but also if the simulation functions as intended. Aberrant process data (e.g., long reaction times, repeated mouse clicks, search behavior) may not always indicate poor student performance but can also indicate that students experience difficulties with the interface of the simulation.

3.4.5 Performance Data Analysis

A third perspective on simulation-based assessments is how the performance data from the assessment is (psychometrically) analyzed using different

strategies and (statistical) techniques. How to treat performance data (both process data and product data) that simulation-based assessment yields is an emerging research field within (educational) assessment. In general, two different streams of research on the analysis of performance data can be identified. The first stream of research focuses on the use of several exploratory techniques or methods to find meaningful relationships between patterns in the data of simulation-based assessment and SMVs. Educational Data Mining (EDM) is the term used to classify these techniques (Rupp, Nugent, & Nelson, 2012). The second stream of research focuses on the application of confirmatory psychometric models on performance data for making (probabilistic) statements about SMVs. The use of an exploratory technique does not rule out a confirmatory model. In fact, these can be very complementary methods, and we would recommend using multiple exploratory and confirmatory techniques to handle the versatile, complex, interdependent and large quantities of performance data that results from student performance in an SBA (Mislevy, Behrens, DiCerbo, & Levy, 2012). The focus should be on trying to find meaningful and psychometrically sound indicators of students' proficiencies.

In Table 3.3, we demonstrate that 6 studies focused on process data, 10 studies focused on product data, and 15 students focused on a combination of product and process data. A good example of the first stream category is the study by Kerr, Chung, and Iseli (2011). They used a specific technique, called cluster analysis, to analyze the log data from an educational video game called *Save Patch*. Cluster analysis can be subsumed under the EDM techniques and aims at identifying patterns within the actions of students that reflect differences in their competence in a skill (Berkhin, 2006). *Save Patch*, a mathematical game, was designed to teach students the addition of fractions. Using cluster analysis the researchers tried to identify specific groups of actions that reflected either a solution strategy or an error pattern. The result was that 73.6% of all student attempts could be classified as either a solution strategy or an error pattern. Identifying the error patterns may be a very useful diagnostic tool for teaching students the right strategies both effectively and efficiently.

In contrast to the EDM techniques, which are mainly exploratory techniques used to identify indicators of performance in often very large pools of data, the confirmatory psychometric models are used to make (probabilistic) statements about student proficiency. Of course, the patterns discovered using

EDM might also serve as input for the confirmatory psychometric models. In addition to the more traditional psychometric models as Item Response Theory, Generalizability Theory, and Classical Test Theory, one of the most often used psychometric models in the context of simulation-based assessment is the Bayesian Network (BN). The BN is used to structure data in such a way that it can be modeled for probabilistic reasoning (Pearl, 1988). Our review shows that it is especially useful for simulation-based assessments because BNs are more applicable to model the complex structure of observations and student models in SBAs (Levy, 2013; Mislevy, Almond, Yan, & Steinberg, 2000).

In total, 14 studies in our systematic review used the Bayesian Network for the analysis of performance data of a simulation-based assessment. For example, Shute (2011) used a Bayes Net for modeling and assessing creative problem solving of students who were playing a commercial video game called *Oblivion*. In this game, students had to fulfil several quests with a character, for instance crossing a river filled with dangerous fish. The ways in which students could cross the river with their character were ranked by experts on their novelty and efficiency. Through probabilistic inferences in the BN the students' actual actions within the game changed the probabilistic statement of either being low or high on creative problem solving. For example, if a student performed a highly novel and efficient action in the game, then the probability that the student is high on creative problem solving would rise. Successive actions of the student in the game can continually 'feed' the BN whereby the beliefs about a student being high on creative problem solving is updated. The example just discussed reflects that Bayesian Networks are very suitable for handling the versatile data that are the result of performance in an interactive and open world simulation.

3.5 Discussion

The systematic review was performed using the method provided by Petticrew and Roberts (2006). In the first of three steps, we have used a multitude of techniques to search and identify literature in which simulations or games are related to (educational) assessment. This resulted in a complete overview of the current state of the literature on SBA. In the second step, all literature was analyzed, categorized and excluded from further analysis if irrelevant. In the third and final step, a deeper content analysis of the included literature in the review

took place. This analysis was being performed from an ECD perspective, and specifically the conceptual assessment framework within ECD. The first category is the type of observable variable(s), the second category is the type of student model variable(s), and the third category is the type of analyses performed using the produced data (evidence accumulation).

The most important conclusion from our review is that the BN is the most used psychometric modeling framework for analyzing the complex and versatile data that result from students' performance in the SBA. In the Bayes net, the evidentiary relationships are modeled in such a way that other psychometric models express or impose their structure on the network, like a MIRT model or a DCM. The BN, then, has some specific advantages that make it so suitable for SBA performance data. For example, joint measurement models for hierarchical (latent) variables can be estimated more easily (Rupp et al., 2012). These types of models are very common in SBA, because multiple, hierarchically structured latent variables influence the students' performance. Furthermore, it is relatively easy to update and improve the BN in real time as new evidence occurs (Shute, 2011). New evidence can be a specific action that a student did (or did not) perform in the SBA. Thirdly, the prior indicators in the BN can be determined based on student data (e.g., from a pilot a study) or elicited from experts. An indicator can be regarded as the probability with which a specific action of the student in the SBA informs some latent student model variable that you wish to make statements about. Finally, recent developments in BN's (and software) make it possible to design very flexible networks with very large collections of variables, which is often the case in educational assessment (Levy & Mislevy, 2004; Mislevy et al., 2014). To illustrate the BN in greater depth, we will present a practical example of a BN in Part II of this chapter.

PART II – A Bayesian Network Example

3.6 Introduction

As our review shows, the Bayes net is one of the most used modeling frameworks in simulation-based assessment. In this Part II section of the chapter, we will provide some background about what they are, how they are constructed and how they can be used. Furthermore, we will provide an example –

including a discussion of the construction – of a BN that is used to model data from an SBA that we have developed and are currently studying.

3.6.1 What are Bayesian Networks?

Bayesian networks (Pearl, 1988) provide a graphical structure in which conditional probability relationships between a large number of random variables are being 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 observed variables. 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), artificial intelligence, for learning systems (Korb & Nicholson, 2010), ecology, for environmental modeling (Aguilera, Fernández, Fernández, Rumí, & Salmerón, 2001), and many other fields among of course educational assessment (Mislevy, Almond, Yan, & Steinberg, 2000; Mislevy, et al., 2014).

The conditional probabilities in the network indicate how the BN 'behaves' as new information is propagated through the network. In an SBA context, for example, given that a student performs several actions within the simulation (observable), it can be probabilistically inferred how other variables' state changes. Of course, BNs can become very large as the number of variables in the network rises. The rationale is that the graphical structure of the network, depicted in a directed acyclic graph (a DAG), makes it possible to perform probabilistic inference among many variables in an acceptable amount of time. Furthermore, the DAG also provides an intuitive insight in how multiple variables are related to each other, and which have a direct influence on each other (Neapolitan, 2003).

3.6.2 How are Bayesian Networks Constructed?

Effectively, building a Bayesian network is following a set of steps. However, in a real-world setting it is much more complex than following a sequence of steps, which is something to keep in mind (Korb & Nicholson, 2010). First, one or more variables of interest must be identified. In an educational assessment setting, this is usually already done for most part before the assessment was built. The SMV(s) that we want to measure is/are the latent variables of

interest. In traditional assessments, all questions can be used in the network to serve as observed indicators. This might be different in SBA, where specific behaviors, actions, decisions, or work products within the simulation can all serve as indicators, as is explained in the evidence identification part of the evidence model in ECD. It is also important to indicate the values or the states that the variables can take. Observable variables in educational testing are often dichotomous (where 0 stands for incorrect and 1 stands for correct), and latent variables can both be indexed dichotomously in two or more classes or be continuous. BNs, in contrast to multidimensional IRT models for example, describe the SMVs as discrete latent variables (Levy, 2013).

Then, the structure of the network should be laid out. That is, how the variables (both latent and observable) in the network influence each other. Because a BN is a DAG, the variables influence each other only in one direction. In a traditional, unidimensional measurement model the structure would look like the one depicted in Figure 3.1. There is one latent variable, θ, which 'causes' the answers on X_1, X_2, X_3, ..., X_j. In SBA, on the other hand, it is not uncommon that there are multiple latent variables, even in a multi-level structure, that have a factorially complex structure (Levy, 2013). That means that two or more latent variables are conditionally dependent on one or more of the same variables in the network. An example of such a multidimensional model is depicted in Figure 3.2. Variables in the network that are not graphically connected are considered to conditionally independent of each other.

Figure 3.1
Graphical Representation of a Unidimensional Measurement Model.

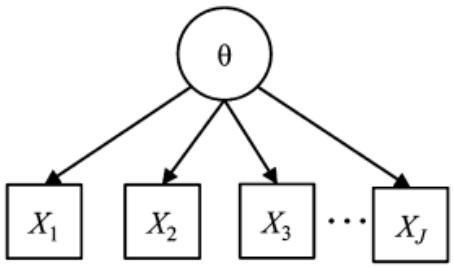
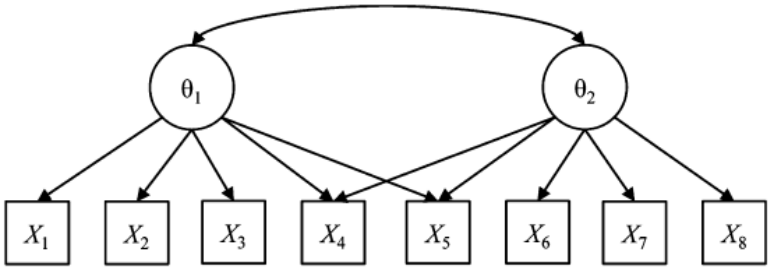


Figure 3.2

Graphical Representation of a Multidimensional Measurement Model with a Factorially Complex Structure



The last step in the sequence is defining the conditional probability tables (CPTs) for the different variables in the network. In educational assessment, probabilities may be learned from real student data or can be defined by experts (Shute, 2011). Each state (e.g., sufficient/ insufficient, low/average/high, etc.) of the latent variable has an associated probability and for each variable there first have to be looked at all possible combinations of the variable's parent nodes, which are the higher-order variables in the BN that have a direct influence on the variable of interest. The specific probabilities for each instantiation can then be defined on basis of the experts' input or empirical data. As mentioned earlier, this is a rather abstract delineation of the process of BN development. We refer to Pearl (1988), Neapolitan (2003), and Korb and Nicholson (2010), who all provide a far more detailed discussion of building BNs, learning BNs, and inference in BNs.

3.6.3 How Can Bayesian Networks be Used?

In educational assessment, the use of Bayes nets is not limited to SBA. In fact, BNs can be perfectly used in traditional, unidimensional assessments as well. However, as assessment developers, assessment users, and policy makers are demanding more from their assessments, for example for diagnostic purposes or for very specific decision making, multidimensional measurement models become more common practice. The characteristics of BNs described above indicate that they are very useful for modeling multidimensional measurement models. We refer to the references in Table 3.3 of our systematic re-

view to see many applications of where BNs are used in educational assessment. We will now discuss how we have built and used a BN for an SBA that we are currently researching.

3.7 Materials and Procedure

At this moment, we are studying a simulation-based assessment that aims at measuring constructs underlying an industrial security profession called *confined space guard* (CSG). A CSG supervises operations that are carried out in a confined space. A confined space is any space which by design has limited or restricted means of entry or exit, is large enough for a person to enter to perform tasks, has the potential for a significant hazard to be present and is not intended for continuous occupancy. Currently, anyone who wants to become a CSG has to do a course which concludes with both a paper-and-pencil test and performance-based assessment.

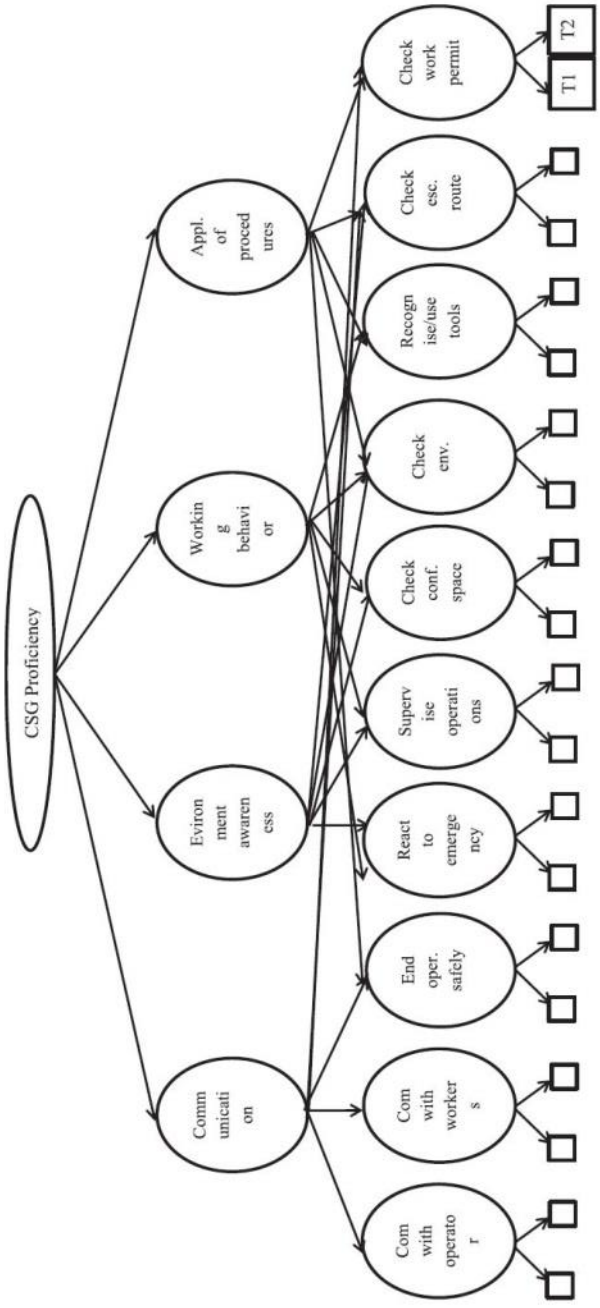
The SBA that we have developed aims at measuring the same constructs that underlie the tasks in the PBA. Using an interactive interface, multimedia and animations we let students go through the procedure as if they would go through in the PBA, the only difference is that they do this on a computer. During the simulation students have to perform tasks, answer questions, make decisions, observe video fragments, and examine documents and tools. For example, one of the tasks is that the student has to check if the work permit for the operations in the confined space is complete and correct. In the SBA, the work permit is first handed over by the operational supervisor in a video fragment, after which a scan of the work permit (with two intentional mistakes in it) is presented on the screen. The student is then asked to indicate if the work permit is correct, and if not what sections of the work permit need to be revised by the operational supervisor.

The SBA was built by a team comprised of subject matter experts, assessment experts, a web designer, a multimedia expert, and a programmer. we used the *final attainment levels* of the PBA to ensure that the content (tasks) in the SBA reflects the actual profession and to make tasks that can be scored. The variables in the Bayesian network reflect the hierarchical structure of the final attainment levels, in which the lowest level are the primary observable variables, and the highest level is a latent performance construct. For example, a primary observable variable can be an answer to a question or a specific action within

the SBA that is scored dichotomously (0=incorrect/not executed, 1=correct/executed). The upper level variable can be regarded as a latent construct called "CSG proficiency". Between the observable variables and the CSG proficiency latent variable are so-called higher order variables (also latent). In our case, we identified higher order variables (*communication, environment awareness, working behavior, and application of procedures*), and ten lower level variables (*communication with operator, communication with workers, end operations safely, react to emergency, supervise operations, check confined space, check environment of confined space, recognize/use tools, check escape route,* and *checking work permit*). The primary observables below directly inform the lower level variables, which in turn inform the higher order variables, which subsequently inform the upper level variable. The whole component-based competency model depicted in Figure 3.3 was built in accordance with subject matter experts.

Now, suppose that we want to create a Bayesian Network on basis of the competency model just explained. For simplicity, we will only focus on the tasks associated with the right hand side of the model – *checking the work permit*. The simplified model shows that two observable variables are connected to a person's ability in checking a work permit, and consequently in applying procedures, being aware of an environment and the overall proficiency in being a CSG, which are all latent variables. In the Bayes Net, the observable actions, that can be scored, inform the higher order latent variables through marginal probabilities. As mentioned earlier, in the assessment the student gets the assignment to check if the work permit is complete and correct. In addition, the student has to ensure that the environment of the confined space and the confined space itself are in accordance with the work permit. In the first task (T1), the student is presented with the work permit on the screen and can use several buttons to navigate through the work permit. The work permit contains two deliberate errors. Via an interactive interface the student can indicate if and where there are errors in the work permit. In the second task (T2), the student is confronted with a series of photos of the confined space and its environment. Two photos display dissimilarity between the work permit and the actual situation around the confined space. Using a multiple answer question type, the students can indicate if and which photos are not congruent with the work permit.

Figure 3.3
Graphical Representation of the Scoring Structure for the CSG Assessment



There are three actions that the student can undertake for T1 and T2, namely either find zero, one or two errors per task. The indicators for scoring the actions can be distilled from real student data, for example through a pilot test using the frequency of the different actions. Indicators can also be determined using experts' input in so-called conditional probability tables (CPTs), which we have done. Effectively, the indicators are conditional probabilities that indicate whether a student is sufficient in checking the work permit given his reaction(s) to the task(s). That is, the more errors a student finds, the higher the probability is that he or she is sufficient in checking a work permit. A student's action can be captured in real time as he or she is going through the SBA, and subsequently the accompanying indicators inform on student's latent competencies through a Bayesian Network using the right software.

3.8 Results

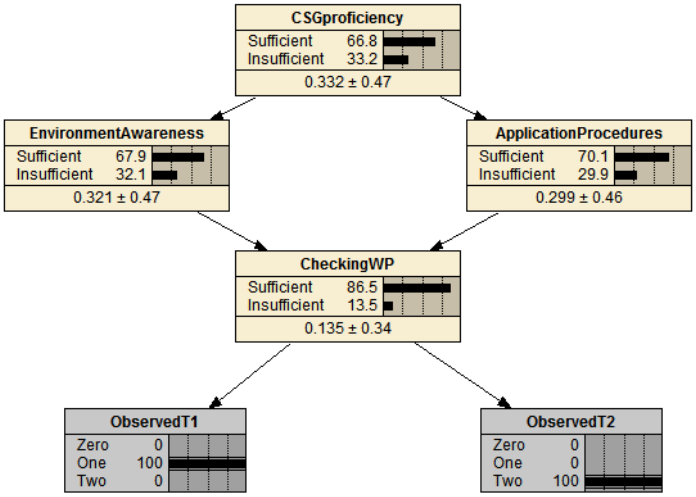
The resulting Bayesian network is shown in Figure 3.4. More specifically, a BN is displayed in which a student has responded to T1 and T2. In this particular case, the student has found one error in the first task, in which the assignment was to indicate whether there were errors in the work permit, and two errors in the second task in which students were being asked to find errors in a photo series of the confound space and its surroundings. As the student has found three of four errors in total, the probability that a sufficient level has been reached is pretty high (.865). In turn, this translates in rising probabilities for sufficient student ability levels for *environment awareness, application of procedures,* and the highest level latent variable *CSG proficiency*, which started at .5 for both sufficient and insufficient. The BN also shows that sufficiency in *checking the work permit* has more influence, through the conditional probabilities, on the application of procedures than on environment awareness.

3.9 Discussion

Of course, this example is a rather simple Bayesian network, and only covers a section of a larger competency model. Full BN's can quickly become very complex models with many latent and observable variables. It is important to note that the prior distributions in the model that we have specified were built with input from CSG subject matter experts. As the model becomes more complex it can be difficult to define the CPTs with expert opinion or to even

reach agreement on the BN structure. It may be a good idea then to use real student data to improve the model through refinement of these CPTs - a big advantage of the BN is that new evidence can be incorporated in the model. Overall, the Bayesian network seems to be the most promising tool for modeling the performance data from complex, multidimensional simulation-based assessments for psychometric analysis.

Figure 3.4
Bayesian Network



3.10 General Discussion and Conclusion

The systematic review of the psychometric analysis of performance data from simulation-based assessment, in Part I of this chapter, showed that the last decade has known an explosively growing body of literature on SBA. At the moment of our literature search, were able to identify more than 30 studies that focused on the psychometric analysis of data resulting from an SBA. Innovative psychometric and statistical methods have really opened up the *black box* of simulations used as measurement instruments. Using the ECD framework we were able to categorize each study according to three categories – what was being measured or the student model, how were they measured or the activity

model and observable variables, and how the performance data was treated or the evidence model. This analysis demonstrated that SBAs have been designed for a wide range of SMVs. Although the measurement of cognitive abilities (e.g., arithmetic ability) predominates in the studies found, there are also researchers that focused on the measurement of more practical skills (e.g., Iseli, 2010). Many studies focused on a combination of product and process data. Especially the process data can reveal interesting information on student model variables and is one of the defining characteristics of SBA. Furthermore, a combination of both exploratory data analysis methods (e.g., EDM) and confirmatory psychometric models is most suiting for the analysis of performance data of SBA. Another very important finding from our review was that that the Bayesian network is the most used psychometric modeling framework for analyzing the complex and versatile data that result from students' performance in the SBA. To further illustrate the use of the Bayesian networks in simulation-based assessment, we have introduced a deeper discussion about BNs in Part II of this chapter. In addition, we have also presented a practical example of a simplified BN for an SBA that we are currently researching.

To conclude this general discussion we would like to make five general recommendations based on our findings, both from Part I and Part II of this chapter, for researchers who are currently developing and researching an SBA. First, we would advise start the design and development of a new SBA by bringing together experts from multiple fields – psychometricians/test developers, educational experts, subject matter experts, programmers and multimedia experts. we have seen (and experienced) that multidisciplinary teams are the key to success in building valid SBAs. Secondly, the continuous innovations that are being made in SBA expand the range of observable variables. Observable variables are no longer answers to questions but can take any form of action within a virtual environment. Often to an extent that a student does not even notices that he or she is being assessed. It is advisable to be creative in determining the right OVs within the simulation. In contrast to traditional item – response question format tests, all product and process data that a student produces by performing in an SBA can be defined as OV. To that extent it is a good idea to make both a qualitative and a quantitative analysis of which data informs best on students' abilities. Thirdly, use multiple sources of data and combine these data to make valid inferences about students' proficiencies based on their per-

formance in the SBA. Fourthly, use multiple psychometric and statistical techniques to analyze the performance data from SBAs: combining both confirmatory and exploratory techniques may result in stronger interpretations about students. Fifthly, we would like to make a case for the Bayesian network that, because of its flexible properties, proves to be a very useful psychometric technique to model student performance from an often versatile and complex structure that is produced by the SBA. To conclude, the findings from our review and the practical example provided will assist in future research into the use of simulations and games as assessment instruments.

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.

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan, C. Nicholas & M. Teboulle (Eds.), *Grouping multidimensional data* (pp. 25-71). Berlin: Springer.

Braun, H., Bejar, I.I., & Williamson, D.M. (2006). Rule-based methods for automated scoring: Application in a licensing context. In D.M. Williamson, I.I. Bejar & R.J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83-122). Mahwah, NJ: Lawrence Erlbaum Associates.

Buckley, B., Gobert, J., Horwitz, P., & O'Dwyer, L. (2010). Looking Inside the Black Box: Assessments and Decision-making in BioLogica. *International Journal of Learning Technology, 5*(2), 166-190.

Buschang, R.E., Kerr, D., & Chung, G.K.W.K. (2012). *Examining feedback in an instructional video game using process data and error analysis* (CRESST Report 817). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal performance*. New York: Cambridge University Press.

De Klerk, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2014). A blending of computer-based assessment and performance-based assessment: Multimedia-based Performance Assessment (MBPA). The introduction of a new method of assessment in Dutch Vocational Education and Training (VET). *CADMO, 22*(1), 39-56.

Doust, J.A., Pietrzak, E., Sanders, S., & Glasziou, P.P. (2004). Identifying studies for systematic reviews of diagnostic test was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *Journal of Clinical Epidemiology, 58*(5), 444-449.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, *4*, 111–143.

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

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

Kerr, D., Chung, G. K. W. K., & Iseli, M. R. (2011). *The feasibility of using cluster analysis to examine log data from educational video games* (CRESST Report 790). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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).

Kim, J., & Chung, G.K.W.K. (2012). *Use of a survival analysis technique in understanding game performance in instructional games.* (CRESST Report 812). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kirriemuir, J., & McFarlane, A. (2004). *Literature review in games and learning.* Bristol: Futurelab.

Klinkenberg, S., Straatemeier, M., & Van der Maas, H.L.J. (2011). Computer adaptive practice on Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*, 1813-1824.

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.

Lamb, R.L., Annetta, L., Vallett, D.B., Sadler, T.D. (2014). Cognitive diagnostic like approaches using neural network analysis of serious educational videogames. *Computers & Education, 70*, 92-104.

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. (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).

Lucas, P. (2001). Bayesian networks in medicine: A model-based approach to medical decision making. *Proceedings of the EUNITE Workshop on Intelligent Systems in Patient Care*, Vienna, 73-97.

Margolis, M.J., & Clauser, B.E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D.M. Williamson, I.I. Bejar & R.J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123-168). Mahwah, 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., Steinberg, L.S., Almond, R.G., Breyer, F.J., & Johnson, L. (2001). *Making sense of data from complex assessment* (CRESST Report 538). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R.J., Almond, R.G., & Lukas, J. (2004). A brief introduction to evidence-centered design. *CSE Technical Report*. Los Angeles: The National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://www.cse.ucla.edu/products/reports/r632.pdf

Mislevy, R.J., Behrens, J.T., DiCerbo, K., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and data mining. *Journal of Educational Data Mining, 4*, 11-48.

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

Mitchell, A., & Savill-Smith, C. (2004). *The use of computer and video games for learning. A review of the literature*. London: Learning and Skills Development Agency.

Neapolitan, R.E. (2003). *Learning Bayesian networks*. New York, NY: Prentice-Hall.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

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

Poropudas, J., & Virtanen, K. (2007). Analyzing air combat simulation results with dynamic Bayesian networks. In *Proceedings of the 2007 Winter Simulation Conference*, pp. 1370–1377. Washington, DC: Institute of Electrical and Electronics Engineers, Inc.

Quellmalz, E.S., Timms, M.J., & Buckley, B. (2010). The promise of simulation-based science assessment: The Calipers Project. *International Journal of Learning Technology, 5*(3), 243-263.

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

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

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).

Rupp, A.A., DiCerbo, K.E., Levy, R., Benson, M., Sweet, S., Crawford, A., Caliço, T., Benson, M., Fay, D., Kunze, K.L., Mislevy, R.J., & Behrens, J. (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.

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.

Shute, V.J, Masduki, I., Donmez, O., Dennen, V.P., Kim, Y-J., Jeong, A.C, & Wang, C-Y (2010). Modeling, assessing, and supporting key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, and N.M. Seel (Eds.), *Computer-based Diagnostics and Systematic Analysis of Knowledge* (pp. 281-309). Boston, MA: Springer US.

Shute, V.J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias and 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.

Stevens, R., & Casillas, A. (2006). Artificial neural networks. In D.M. Williamson, R.J. Mislevy, & I.I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 259-312). Mahwah, NJ: Lawrence Erlbaum Associates.

Thomson Reuters (2013). EndNote. The most powerful tool for managing your research. Retrieved from http://endnote.com/

Van der Kleij, F.M., Timmers, C.F., & Eggen, T.J.H.M. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *CADMO, 19*, 21-39.

Vendlinski, T. & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *Journal of Technology, Learning, and Assessment, 1*(3). Retrieved from http://escholarship.bc.edu/jtla/vol1/3/

West, P., Wise Rutstein, D.., Mislevy, R. J., Liu, J., Choi, Y., Levy, R., Crawford, A., DiCerbo, K.E., Chappel, K., & Behrens, J. T. (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).

Chapter 4. A Framework for Designing and Developing Multimedia-based Performance Assessment in Vocational Education[5]

Abstract

The development of any assessment should be an iterative and careful process. Ideally, this process is guided by a well-defined framework (see for example Downing 2006; Mislevy et al. 1999; AERA et al. 2004), but such a framework is not always available when the instrument to be developed is new or innovative. Frameworks for the development of traditional computer-based tests have been published and experimented with since the late 1990s, by which time CBT had already existed for more than a decade. In an earlier empirical pilot study, we described a new type of assessment for Dutch vocational education, called multimedia-based performance assessment (MBPA). This CBT uses multiple media formats and interactive tasks to measure skills that are currently measured by performance-based assessment. In conducting that pilot study, deficits in the existing literature made it difficult to ground all developmental steps in sound scientific theory. To remedy those deficits, this article presents and validates a framework for the design and development of MBPA, combining a search of the relevant literature from several subfields of educational assessment and consultation with assessment experts. The first step in validating the prototype framework involved five semi-structured interviews with Dutch assessment and multimedia experts to produce a final version of the framework. Second, the pilot MBPA was revised in accordance with this finalized framework, resulting in an improved MBPA and demonstrating that the proposed framework is a useful and applicable tool for the design and development of MBPA in vocational education.

[5] This chapter is a minor revision of De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2015). *A Framework for Designing and Developing Multimedia-based Performance Assessment in Vocational Education.* Manuscript submitted for publication.

4.1 Introduction

Multimedia-based performance assessment (MBPA) is an innovative type of assessment that blends computer-based testing (CBT) and performance-based assessment (PBA) (self-revealing reference 2014). We have introduced the term MBPA for two reasons. First, it is enabled by technological and digital innovations employing multimedia. In contrast to more traditional forms of CBT, multimedia is used in MBPA to simulate a real-world environment in which tasks are administered as the student navigates that environment. Using multimedia (e.g., animation, video, virtual reality), students are "immersed" in a virtual environment in which they must complete tasks or achieve objectives. Second, MBPA is used to measure student skills that are currently measured by performance-based assessment, and so the term we coined was *multimedia-based performance assessment*. In this virtual environment, students may be provided with tools to help them complete their tasks, and because they have, (to varying degrees) more freedom to operate, there is usually more interaction between the student and the computer in this environment than in other types of CBT.

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

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

using multimedia and interactive tasks. In designing and developing this innovative MBPA, we identified an unmet need for a well-defined framework to guide this complex and multifaceted process. Increasing research and development on the measurement of vocational skills through CBT (Koenig, Lee, Iseli, and Wainess 2010) provides a strong empirical foundation for the present study.

There is also some theoretical basis for the framework, including Downing's twelve steps for effective test development (2006), the Standards for Educational and Psychological Testing (AERA et al. 2004) and the evidence-centered design (ECD) framework (Mislevy et al. 1999). The first two of these focus strongly on traditional testing formats and can be used as a structured, step-by-step approach to assessment development. Mislevy et al.'s framework is also applicable for simulation-based assessments but offers a more abstract approach to assessment development. In the present study, the proposed framework combines a step-by-step approach to designing and developing complex and innovative computer-based assessment—in an educational field (vocational education) that is traditionally less well researched than primary, secondary, and higher education—with attributes that are less common in those fields, such as the strong focus on performance-based assessment in acquiring and assessing vocational skills).

4.1.1 Assessment Design and Development

Sound and coherent assessment originates in a structured and well-defined approach to assessment development, which also ensures collection of sufficient evidence for the validation of future assessment scores (Downing 2006). Unstructured assessment development may result in a poorly functioning assessment that fails to support valid inferences about students' proficiency. Assessment design and development is a time-consuming, intensive, and laborious process of trial and error. Multiple specialists from different disciplines often work collectively to address different components of assessment development (Mislevy et al. 1999). Subject matter experts, assessment experts, and multimedia experts are all needed in the design and development of MBPA, and such a complex and highly interactive process is unfeasible without a unifying framework. Such a framework should also enable designers and developers to collect evidence for future validation of assessment scores.

In endeavoring to design and develop MBPA as a multimedia counterpart of PBA in vocational education and training, the complexity of the undertaking soon became clear. For example, it was found essential to involve multimedia experts at the earliest stage of design sketches because they know how to translate tasks into working ICT solutions. Without their early input, it is more likely that tasks will be designed that cannot be realized in a virtual environment. It is also important to collect assessment evidence, which is a complex and comprehensive process—especially where virtual environments are used, as the assessment results should reflect competency in a real-world setting. It was also found that the various processes—construct analysis, task design, building an evidence and measurement model in conjunction with actual development of multimedia and ICT assessment structure—had to be conducted iteratively and in parallel to ensure an assessment design that could be realized in an ICT environment.

It seems essential, then to develop a framework for assessment design and development, especially for any new or innovative type of assessment such as MBPA. The next section describes how the proposed framework was constructed.

4.2 Method

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

4.2.1 Step 1 - Literature Study

To begin construction of the prototype, sources that included Web of Science, Scopus, and Google Scholar were searched using relevant terms (e.g., "assessment design", "assessment development", "assessment guidelines",

"assessment framework", "test design", "test development", "test guidelines"). Following the literature review strategy of Petticrew and Roberts (2006), items were selected if (1) the main topic of the article or chapter related to assessment/test development and (2) the article or chapter provided a structured set of rules or guidelines (i.e., a framework) for assessment development. For example, the ECD framework (Mislevy et al. 1999) provided relevant and valuable input and their evidence model has therefore been adopted here.

4.2.2 Step 2 - Construction of the Prototype

The literature study was followed by consultation with three professional experts in the field of assessment, working respectively for the Dutch national institute for test development (Cito), the University of Twente in the Netherlands, and a private Dutch assessment development company. All three had more than ten years of experience in the design and development of assessments. During construction of the prototype, four rounds of expert consultation were organized to ensure that development remained on track and to avoid tunnel vision. Having first constructed the two general stages of the framework (as explained below), both stages were structured as sequential steps for design and development of MBPA, and those stages and steps were then connected to explicate their interrelationships. Finally, multiple sub-steps were added to the task design and development steps; these cannot stand on their own as separate steps because they are strongly connected to their parent step, but as they define quite specific processes during task design and development, it is reasonable to add them to the framework. For example, in accordance with the experts' view, the evidence model was placed appropriately within the framework, linking it to the relevant steps.

4.2.3 Step 3 - Validation of the Prototype

Participants. The prototype was validated by means of five semi-structured with experts in either assessment, multimedia design and development, or both, and the prototype was finalized version on this basis. To keep construction and validation of the prototype separate as processes, the participating experts were not involved in construction of the prototype. All had more than ten years of (leadership) experience in the design and development of innovative CBT. Three of the participants had a doctoral degree and two had a

master's degree in the areas of assessment or multimedia development. Their backgrounds were very diverse. One was primarily researching the incorporation of serious gaming elements in assessment for the purpose of personnel selection. Another was responsible for the implementation of technology-based assessment at Cito, while a third was involved in the development of multimedia for CBT. The two remaining experts were primarily involved in research on the innovative use of CBT in higher education.

Materials and procedure. The five identified experts were approached via e-mail and asked them whether they would be willing to read a manuscript describing the proposed prototype; all five replied positively. A semi-structured interview schedule was then constructed, based on a study in which an evaluation system for performance assessments had been validated on the basis of expert interviews (see Wools et al. 2011). Specifically, the interviews revolved around two concepts: the content and usability of the prototype. Content was characterized in terms of four categories: general quality, completeness, correctness, and coherence. Usability was characterized in terms of two categories: general usability and fitness for purpose. During the interviews, the experts were systematically questioned about all elements in the framework in respect of those concepts and categories. All interviews were conducted face-to-face at the experts' work location; the interviews were recorded, making it possible to carefully re-listen to and interpret the experts' statements without losing the context in which those statements were made.

A verbatim transcript was made of each interview. To keep experts' statements in their proper context, cues were written in the margin of the transcript to indicate any special circumstance that led to this statement (e.g., an example, anecdote, or personal experience). Text fragments that referred specifically to the content concept, the usability concept, or to one of the underlying categories were then filtered and selected from the full transcripts. This selection of text fragments was done on an individual and independent basis by each of the authors of this article. Subsequently, we collectively discussed what fragments were useful for revision of the prototype and which were not. A high degree of correspondence was found between the three authors in the fragments selected, and any fragments selected by all were automatically included. Statements that were selected by one or two of the three authors were collectively discussed for possible inclusion; overall, we were quickly able to reach agreement about these

statements. The rationale behind this strategy was that the experts' views as expressed in these text fragments would be both meaningful and useful for the further development of the prototype into a final framework. To follow up on our example, the evidence model of course formed part of the subject matter of the five interviews; the experts were questioned on the positioning of the evidence model within the framework and whether it was usable and correctly described.

4.2.4 Step 4 - Adjustment of the Prototype and the Final Framework

In the fourth step, statements made by the assessment experts were used to transform the prototype into a final framework. In total, 28 text fragments were extracted from the interviews that relate directly to the framework. As can be seen in the results section below, the role of the evidence model changed in the final framework as compared to the prototype, demonstrating how significantly the interview data impacted the finalization process.

4.2.5 Step 5 – Validation of the Final Framework

In the fifth and final step, the final version of the framework was used in a real situation. As discussed above, a pilot MBPA had already been developed for a Dutch vocational course for CSGs, and the degree of difficulty experienced in the design and development of that pilot MBPA was one reason for building this framework. In the last step of this research, the final framework was used to redevelop the CSG MBPA, which can be seen as an empirical step in validating our framework; if the end product (the new MBPA developed according to the such demonstration that a justifiable evidence model could be built for this MBPA would constitute a strong argument that the evidence model was correctly positioned within the framework. There follows a discussion of the results from each of the five steps.

4.3 Results

4.3.1. Step 1 - Literature Study

The literature search returned 14 articles or book chapters used in construction of the prototype, and each step in the prototype is grounded in a relevant area of the literature. The steps in both stages of the framework (design and development) were specifically linked to Downing's (2006) twelve steps for

effective test development, The Standards for Educational and Psychological Testing (AERA et al. 2004) (from here on referred to as the Standards), and the evidence-centered design (ECD) framework of Mislevy, Almond, and Steinberg (1999).

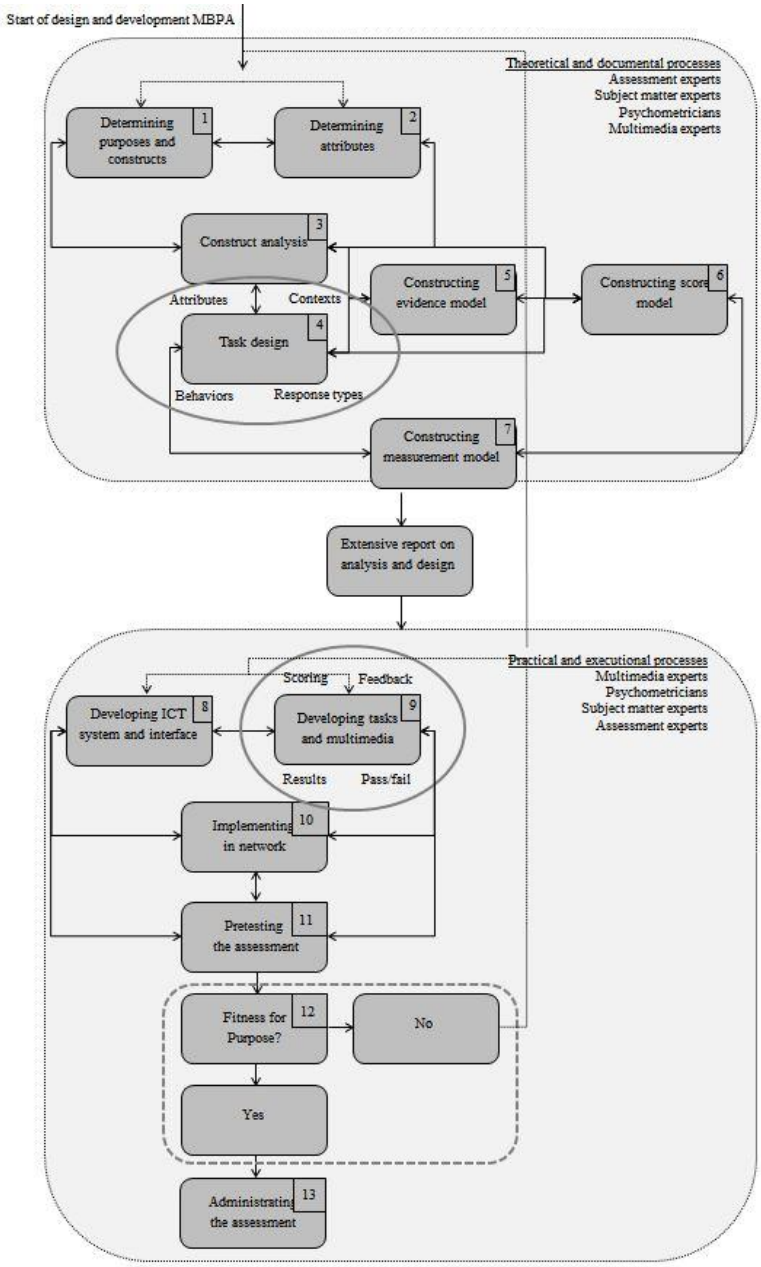
We refer in particular to these three frameworks for a number of reasons. First, Downing's *Handbook of Test Development* (2006) provides a step-by-step approach to assessment development, which is also an aim of the proposed framework. Secondly, while the Standards are the most influential guidelines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

4.3.2 Step 2 - Construction of the Prototype

Based on the literature search and input from the participating assessment experts, two processes were identified in building an MBPA: a design phase and a developmental phase. The framework therefore comprises two general stages: *analysis and design* and *development and administration*, both involving different processes. The analysis and design stage is guided mainly by assessment experts and subject matter experts and is for the most part executed mentally and on paper. The development and administration stage, on the other hand, is guided mainly by multimedia experts and practitioners and is for the most part executed practically in an ICT environment. For efficiency, we will refer in the following sections to the "assessment developer" as representing the whole team engaged in the design and development of the MBPA. Decisions made at the first stage influence the second stage, and conversely, the first stage is also influenced by the second stage, when possible hiatuses in the first stage may be detected. The full prototype framework is presented in Figure 4.1.

Figure 4.1

Prototype Framework



Analysis and design. Turning to a systematic discussion of all parts of the framework, the first stage involves seven steps, resulting in a detailed report for use at the development and administration stage. In this first stage, the general rationale is to design assessment tasks that are grounded in theory, are measurable, and elicit student behavior that reflects the construct (competencies, skills, knowledge, etc.) to be measured. These steps were identified on the basis of the literature informing each step, along with the expert inputs during the several rounds of consultation.

The first step, then, is *(1): determining the purpose(s) and construct(s) of the assessment*. Here, the assessment developer elaborates a comprehensive argument concerning the purpose of the assessment—what precise construct is to be assessed and why there is a need for that assessment. During this first step, an extensive overall plan should be made for systematic guidance of the developmental process (Step 1: Downing 2006). The Standards emphasize the interpretation of assessment scores that strongly relate to the purpose of the assessment, which may, for example, include certification of individuals (e.g., a yes/no decision), course placement, or curricular reform (RCEC 2014; Baker et al. 1993; Drasgow and Olson-Buchanan 1999; Schmeiser and Welch 2006). The assessment developer should state clearly the purpose of the assessment and what interpretations should follow from the scores produced (see Standard 1.1, 1.2, 3.2, and 14.1). Mislevy et al. (1999) refer to this step as one of the key ideas in educational measurement: "identifying the aspects of skill and knowledge about which inferences are desired".

The second step is *(2): determining the attribute(s) of the construct under measurement*. Some constructs comprise several attributes—in vocational education, it is not uncommon to refer to competencies (Baartman 2006), usually composed of knowledge attributes, skill attributes, and attitude attributes (Klieme et al. 2008). Sometimes, students must demonstrate that they have mastered one of the attributes; on other occasions, they may be required to demonstrate a combination of attributes in a single setting, usually in a performance-based assessment (Linn et al. 1991; Baartman 2006). For development of the assessment, then, it is very important to define which attributes of the construct are part of the assessment (and therefore operationalized) and which are not. For example, if the construct is writing, the attribute might be knowledge on writing or style, but it might equally be the student's writing skill or the use of style in a writing

assignment. Step 2 of Downing's (2006) twelve steps for effective test development stresses the importance of carefully delineated constructs. The assessment developer needs to consider which attribute in particular of the construct is to be measured, and the appropriateness of the assessment content for that particular attribute should be justified (see Standard 1.6). The second step of the framework can again be related to the key idea explicated by Mislevy et al. (1999): "identifying the aspects of skill and knowledge about which inferences are desired".

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

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

Logically, the assessment developer cannot include in the assessment anything outside of the domain. Within the domain, there is a universe of tasks that might be designed and incorporated into the assessment (Mislevy et al. 1999; Mislevy 2011). Through systematic analysis of actual job behaviors, the assessment developer is able to design tasks that will form part of the assess-

ment (Weekley et al. 2006). For example, by carefully observing the performance of qualified job incumbents, the assessment developer can isolate typical job behaviors that are the pillars of the vocation. This stage is generally characterized by a synthesis between subject matter experts (SMEs), and assessment experts (Downing 2006; Weekley et al. 2006).

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

Finally, using multiple perspectives (e.g., a competency-based profile, a qualification file, an analysis of job behavior, data from SMEs, and a cognitive analysis) the construct analysis delineated above informs a comprehensive argument explaining which factors of vocational behavior should be included in the tasks. In the task design step, then, the assessment developer should follow a strategy to select tasks that cover either the whole domain or the most important tasks within the domain. The latter strategy, which is fairly often used in vocational education, is also called the critical incidents technique (Flanagan 1954)—selecting the tasks that best predict future job behavior or are characterized as high risk, either for the student or for the organization, based on the construct analysis. This third step in the framework relates to another key idea of educational measurement as discussed by Mislevy et al. (1999): "identifying the relationships between targeted knowledge and behaviors in situations that call for their use".

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

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

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

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

The final part of Step 4 is the *response type* that characterizes the tasks. MBPA includes a whole range of new response types for logging actual student behavior in the tasks—for example, speed, clicking behavior, navigational behavior through the virtual environment, typing, eye-tracking, and accuracy. Of course, both innovative and traditional item types can be incorporated in the MBPA (for an overview of innovative scoring in CBT, see also Williamson et al. 2006; Mayrath et al. 2012; "self-revealing reference" 2012). Downing (2006) argues that the creation of effective assessment tasks with the right context and the appropriate cognitive level is one of the most difficult tasks in assessment development (see Step 4). Logically, the type of item and the response formats should be selected for the purposes of the assessment (see Step 1), the domain to be measured (see Steps 2 and 3), and the intended students (see also Standard 3.6). The fourth step in the framework also relates to another key idea of educational measurement as discussed by Mislevy et al. (1999): "identifying features of situations that can evoke behavior that provides evidence about the

targeted knowledge". We can also recognize some basic models from the ECD framework (Mislevy et al. 1999) in this step: the student model and the task model.

The fifth step is *(5): constructing the evidence model*. This step is schematically located between steps three and four, and relates to the exchange relationship between the former two steps. The evidence model implies that the assessment developer should construct and present a comprehensive and extensive argument that vindicates and explains why the constructed tasks (including attributes, context, student behavior and responses) should result in sound statements about students. In other words, there should be evidence that we can actually say something about students in real life (i.e., the criterion) based on their performance of the tasks in the assessment (i.e., the predictor) (see Standard 14.12). Often, the strength of the relationship can be determined after administration of the assessment has yielded results. However, it important to systematically analyze to what extent it seems plausible to expect valid results from performance of designed assessment tasks. For this reason, Downing (2006) remarked that systematic, thorough, and detailed documentation for validity arguments should be collected continuously (Steps 3 and 12). Mislevy et al. (1999) discern two models within the evidence model: the statistical model and the evidence rules. The evidence model in the proposed framework refers to and builds upon the evidence rules specified in the ECD framework, as the assessment developer should provide evidence of the relationship between student behavior in assessment tasks and the construct.

The sixth step is *(6): constructing the score model*. Student behavior in the assessment has to be scored in order to construct a measurement model that will lead us from collected observed variables to claims about the construct. All observed student behavior during administration that contributes to an overall score forms part of a score model. Scoring may be quantitative as well as qualitative, and scoring rubrics assist in attaching weights to the scores and combining them into an overall score or result (Shepherd and Mullane 2011). According to Downing (2006), perfectly accurate scoring results in valid meanings, as they are anticipated by the assessment developer. Furthermore, the assessment developer should specify the scoring criteria and procedures for scoring in sufficient detail and clarity to make scoring as accurate as possible (see Standard 3.22). In their ECD framework, Mislevy et al. (1999) classify scoring mainly

under the task model, but it also relates to their student model and evidence model because of the link between performance and evaluation.

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

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

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

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

mean to include a virtual interface that, for example, can incorporate movies, animations, and avatars.

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

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

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

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

The thirteenth and final step is *(13): administering the assessment*. The assessment can be administered when pretesting delivers the desired interpretation of

assessment scores. This does not mean that the assessment is complete and can be endlessly reused. The quality of the assessment and its fitness for purpose should be constantly monitored by educational and assessment experts (see also Standard 3.25).

4.3.3 Step 3 - Validation of the Prototype

We have validated the prototype discussed above on the basis of five semi-structured assessment expert interviews. We were able to filter and select 28 text fragments (i.e., statements) from the verbatim interview transcripts that specifically referred to the content concept, the usability concept, or one of the underlying categories. The distribution of the statements is shown in Table 4.1. In addition, we have added the questions that are answered by the statements in the first column of the table. These are not questions that were part of the interview, but they make the concepts and categories more explicit.

All the text fragments are listed in Table 4.2. Using these text fragments, it was possible to determine which steps or stages needed to be adjusted or refined for transformation of the prototype into the final framework. For efficiency, duplicate text fragments have been deleted; this only holds if the duplicate text fragments were used to refer to the same concept or category. For example, if two identical fragments from the same interviewee (or from two or more interviewees) referred to the completeness of the framework, then one was deleted. If one of the two statements referred to the completeness of the framework and the other to the usability of the framework, none was deleted.

Table 4.1
Classification of Number of Text Fragments in Concepts and Categories

Questions	Categories	N
To what extent is the quality of the framework sufficient according to experts?	Content: general quality	6
To what extent are steps and/or stages in the framework complete according to experts?	Content: completeness	8

Table 4.1 (continued)
Classification of Number of Text Fragments in Concepts and Categories

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

Table 4.2
Classification of Verbatim Text Fragments in Concepts and Categories

Questions	Text fragment
Content: general quality	*The framework needs to display a more dynamic process.*
	The framework needs to display a more fluid process rather than a sequential or linear process.
	The framework needs to display more of an iterative process.

Table 4.2 (continued)
Classification of Verbatim Text Fragments in Concepts and Categories

Questions	Text fragment
	The framework needs to incorporate more loops.
	The framework needs more balance between the first and second stage.
	The framework needs to be more of a cyclical concentrically designed process in which every step loops back to the previous steps and in which a prototype is continually updated throughout design and development.
Content: completeness	*The framework lacks a step that facilitates the process of suppressing bias that results from the interface of the MBPA.*
	The framework lacks a step that refers to a cognitive walkthrough of the MBPA.
	The framework lacks a step that refers to a paper-based walkthrough of the mockup of the MBPA.
	The framework lacks sufficient information about the feedback of performance in the MBPA to students.

Table 4.2 (continued)
Classification of Verbatim Text Fragments in Concepts and Categories

Questions	Text fragment
	The framework lacks information about the scoring of MBPA, e.g., is it fully automatic or blended.
	The design of tasks and feedback should be incorporated into one step of the framework.
	The second stage of the framework needs to be elaborated to maintain the balance between both stages.
	The framework lacks an exit step before the developmental phase starts.
Content: correctness	*The first stage of the framework needs to emphasize in what way MBPA design differs from traditional test design.*
	The first step of the framework should be 'determining purposes', rather than purposes and constructs.
	The framework should constantly update a prototype of the MBPA after every step.
	The eighth step of the framework, in the second stage, should be 'choosing an ICT interface', rather than developing one.

Table 4.2 (continued)
Classification of Verbatim Text Fragments in Concepts and Categories

Questions	Text fragment
	The first stage should be 'design' rather than 'analysis and design', and the second stage should be 'development' rather than 'development and administration'.
Content: coherence	*The first and second stage of the framework need to be more parallel processes in order to make the framework more coherent.*
	The final step of the framework, administration, is not part of design or development.
	ICT is much more important in the first stage of the framework in order to make the connection with the second stage.
	The framework needs to emphasize the relationship between design and development by incorporating more backward looping, also between stages.
Usability: general usability	*The framework needs to remain practical, in a sense that usability and the fulfillment of purposes is more important than a shiny layout.*
	To improve usability the framework needs a go/no go step before actual, costly development can start.

Table 4.2 (continued)
Classification of Verbatim Text Fragments in Concepts and Categories

Questions	Text fragment
	The framework is useful if ICT experts are part of the design stage as well as the development stage; otherwise the gap between design and development becomes too large.
Usability: fitness for purpose	*The framework is useful for experts leading a group of subject matter experts, however the framework needs to be simplified to make it useful for practitioners as well.*
	The framework needs to place more emphasis on the developmental phase as well to make the framework useful for multimedia experts

4.3.4 Step 4 - Adjustment of the Prototype and Final Framework

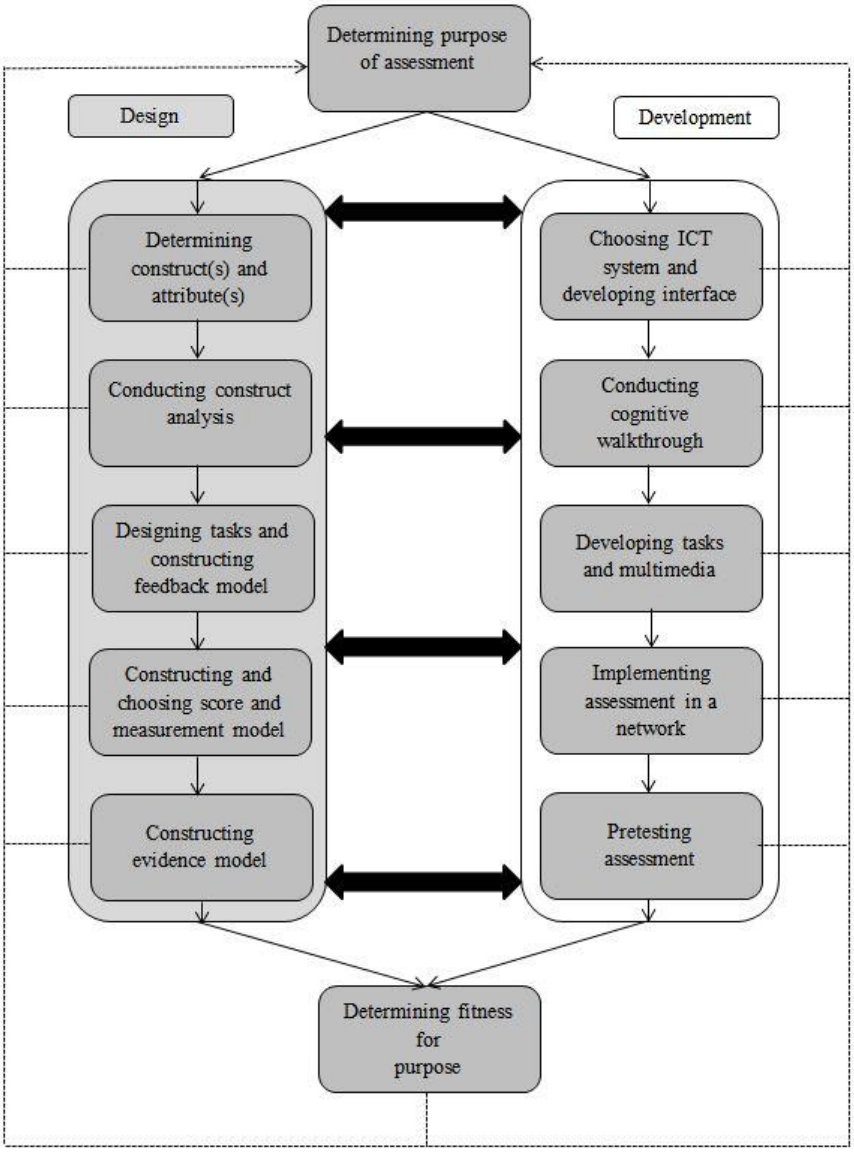
The text fragments refer either to general factors that are applicable to the complete prototype framework (e.g., "more iterativeness") or to specific factors that are applicable only to an element of the framework (e.g., "eliminate the final step"). In finalizing the prototype, we have addressed both types of text fragments, adjusting the general flow of the framework as well as specific elements within it. A schematic flow diagram of the amended prototype (the final version of the framework) is depicted in Figure 4.2.

General adjustments. First, we have tried to make the framework more dynamic and fluid by transforming the sequential form of the framework into a more parallel form and by adding more backward loops between steps and stages. The steps in the left part of the framework relate to assessment design while the steps in the right part of the framework relate to development. Second, the backward loops between each step and the original purpose determination also exemplify the iterative nature of the design and developmental process. Now, progress is constantly monitored by relating each step to the original purpose that the assessment should fulfill. In that way, we have also placed more emphasis on ICT from the earliest moments of assessment development.

Third, we have removed the sub-steps of task design and development to make the framework more efficient and practical for its users. Fourth, we have removed the numbering of steps to emphasize the dynamic, fluid, and parallel nature of the framework. Fifth, we have slightly extended the development stage by adding a cognitive walkthrough step, which also relates to the interface of the assessment. In this way, we also provide more balance between both stages. Sixth, we have renamed the stages; the first stage is *design*, and the second stage is *development*. Finally, we have removed types of process and the constitution of the development team from the framework. Based on the interviews, we believe it is possible to develop MBPAs with relatively small teams, in which members take on different roles and work collectively through all steps of the framework.

Specific adjustments. We have also made some specific adjustments to the prototype framework. First, we have added steps to the final framework: a cognitive walkthrough and a separate step involving the construction of a feedback model. Second, we have also removed or changed steps. The eighth step of the second stage has been renamed *choosing ICT system and developing interface*, rather than *developing an ICT interface*. We have added the attributes determination to the first step of design—first, because it relates directly to the purpose of the assessment, and second, because it makes the framework more efficient and user-friendly. The score model and measurement model are also combined in one step in the final framework because the scores produced in an assessment strongly influence the choice of measurement model, and vice versa. Second, it also strengthens the efficiency and user-friendliness of the final framework. We have renamed *constructing a measurement model* to *choosing a measurement model*. The final stage, *administration*, has been removed, and we have removed the extensive reporting between first and second stages because the final version of the framework represents an integrated process involving both stages. By incorporating the adjustments discussed above in the prototype framework, we believe we have addressed most of the statements made by the experts during the interviews.

Figure 4.2
Definitive Framework



4.3.5 Step 5: Validation of the final framework

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

We started by defining the purpose of the CSG assessment, which was defined in the "final attainment objectives" for certification of CSGs (step 1). The purpose of the MBPA and the strategy for achieving that purpose was documented in a detailed project description. The project description also included a systematic developmental plan and choice of ICT system for the MBPA (step 2—development), which followed the framework's steps. A risk analysis was conducted to hypothesize about possible pitfalls in the project and how to avoid them, or how to handle them if they occurred. The project organization was described in order to assign project team members' roles and to ensure clear communication between members. By comparison with our pilot endeavor, the start of the project was much more structured, which in itself improved the chances of a successful outcome. For example, careful calculation of possible risks in the future steps of design and development can prevent mistakes and delays in building the assessment, which increases the possibility of an assessment that is fit for its original purpose.

The design phase commenced by determining the constructs and attributes to be measured and analyzing them for translation into the MBPA tasks (step 2—design). This was done in collaboration with subject matter experts (SMEs) through multiple rounds of consultation. Of course, a lot was already known about the CSG's tasks from the instructional material and the final attainment objectives of the performance-based assessment. Additionally, the first author took part in a one-day course and performed the PBA to become a certified CSG. This material and knowledge was used to further develop a structure of constructs and attributes for the MBPA (step 3—design).

Our framework ensured that attributes were mapped at the finest possible grain, enabling the design and development of assessment tasks that would

yield the most interesting and relevant information about students' skills (step 4—design). Of course, this was done in collaboration with the SMEs, first building what we called an *assessment skeleton,* in which the general flow of the assessment was laid out, including required multimedia and items or assignments. This was done on a relatively abstract level, but it ensured that all constructs, final attainment objectives, and primary observables were incorporated in the tasks. Because the assessment skeleton is still a relatively coarse-grained representation, it is not sufficient for actually building the assessment. For that reason, we further elaborated the assessment skeletons into *assessment templates*, showing (screen by screen) what was to be presented during the course of the assessment. These templates were based on a cognitive walkthrough of the assessment (step 3—development) and describe which buttons are present and which are not, which multimedia is presented, instructions for the student, what possible actions can be attempted and how these actions are scored (step 5—design). The assessment templates enabled collection of multimedia (video and photo) material in one day, at a reconstructed job site in the Netherlands that is used for practice and performance-based assessments (step 4—development). In addition, the templates served as a primary input for design of the buttons needed in the assessment. We hired a professional designer who was very experienced in designing intuitive, usable, and efficient interfaces for interactive websites. Furthermore, in combination with the buttons, the templates provided the necessary material for the programmer to build the structure of the assessment into our own assessment platform (step 5—development). The next step was to test the assessment—first for its technical functioning and then for its psychometric functioning, in a pilot study (step 6—development). The assessment was administered via the Internet, and multiple test rounds enabled any remaining errors to be resolved, so ensuring that the assessment was technically functional.

Finally, construction of the evidence model consisted of building an argument for the validity of the assessment (step 6—design). In this case, we already build an evidence model by using the framework itself, and professionals from several fields contributed to the process, with practical IT design and development of the assessment by an experienced web designer, multimedia expert, and programmer. The content was specified by subject matter experts and based on our previous experience of performance-based assessment. Another

important aspect of evaluating the assessment is the empirical analysis of its performance properties by use of a measurement/statistical model, ultimately determining whether the MBPA has really met its goal (step 7). This will be detailed in a future publication.

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

4.4 Discussion and Conclusion

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

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

The final framework has also been simplified by comparison with the prototype, making it easier to use and understand, not only for practitioners as well as researchers. We believe that the coming decades will be characterized by a growing emphasis on multimedia-based performance assessment and related types of assessment in vocational education, to which this framework can be hoped to contribute.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Stephens, D., Bull, J., & Wade, W. (1998). Computer-assisted assessment: Suggested guidelines for an institutional strategy. *Assessment & Evaluation in Higher Education,* 23(3), 283–294.

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

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

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

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

Chapter 5. The Design, Development, and Validation of a Multimedia-based Performance Assessment for Credentialing Confined Space Guards[6]

Abstract

In this chapter, we present a study on the design, development, and validation of a Multimedia-based Performance Assessment (MBPA) for measuring the skills of confined space guards in Dutch vocational education. An MBPA is a computer-based assessment that incorporates multimedia to simulate tasks. It is designed to measure performance-based skills. A confined space guard (CSG) supervises operations that are carried out in a confined space (e.g., a tank or silo). In the Netherlands, individuals who want to become certified CSGs have to participate in a one-day training program, and pass both a multiple-choice knowledge test and a Performance-based Assessment (PBA). In the first part of this chapter, we focus on the design and development of the MBPA, using a specific framework for design and development. In the second part of the chapter, we present a validation study. We use an extended argument-based approach to validate the MBPA (Wools, 2015). The extended argument-based approach to validation suggests using multiple sources of validity evidence to build a comprehensive validity case for the proposed interpretation of assessment scores, and to evaluate the strength of the validity case. We demonstrate that MBPA scores can be used for their intended purpose; students' performance in the MBPA can be used as the basis for making a CSG certification decision.

[6] This chapter is a minor revision of De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2015). *The design, development, and evaluation of a multimedia-based performance assessment for credentialing confined space guards.* Manuscript submitted for publication.

5.1 Introduction

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

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

A second group of highly innovative assessment techniques is found in so-called game-based assessment (GBA) (Mislevy et al., 2014). In GBA, test takers play a real video game, while all actions are logged in the background. This can

play a role in performance evaluation. The rationale is that game environments provide opportunities for students to demonstrate their skills in highly contextualized and interactive situations (Klopfer, Osterweil, & Salen, 2009). In this way, it is possible to measure new aspects of the students' skills and to measure other aspects better. An example of GBA is provided by Shute (2011). She uses the term *stealth assessment* for the measurement of student competencies in video games. As students are playing and progressing through the levels of a game, their actions, decisions, use of tools, navigation skills, and so on are being logged and then used to update beliefs about student competency in a particular skill, without students even noticing that they are in an assessment situation.

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

In this chapter, we discuss the design, development, and validation of a technology-based assessment that is positioned close to the right of the continuum of complexity, interactivity, and fidelity. The assessment we present incorporates images, animations, and videos for the purpose of creating complex and interactive tasks in a simulation of a real-world setting. We call this type of technology-based assessment Multimedia-based Performance Assessment (MBPA), because the tasks in the assessment are for a large part constructed of

multimedia and are used to measure student skills that were previously measured by a PBA. The purpose of the MBPA we discuss here is to measure the skills of CSGs after they have completed vocational training (De Klerk, Eggen, & Veldkamp, 2014). The skills that a student has to demonstrate during the assessment were determined by a commission of experts.

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

Although PBA has been discussed and supported as a valuable tool for formative and diagnostic assessment of students (Gulikers, Bastiaens, & Kirschner, 2004; Roelofs & Straetmans, 2006), the research is less supportive in cases where PBA is used as a summative assessment. This is foremost because PBAs are prone to measurement errors resulting from several sources, including task, occasion and rater sampling variability (Shavelson, Baxtor, & Gao, 1993; Cronbach, Linn, Brennan & Haertel, 1997; Dekker & Sanders, 2008). Shavelson, Ruiz-Primo, and Wiley (1999) provide an example in which task and occasion sampling are confounded. In such cases, their combined effect strongly increases the rate of measurement error. These findings indicate that students' scores in a PBA do not solely represent students' proficiency in a particular skill, but are influenced by the specific task that they are assigned, the occasion of the assessment, and the raters who judge their performance. In addition, this study found that it was difficult to define the exact source of measurement error because of the complex relationship between task and occasion sampling. In addition, PBAs are expensive and time-consuming when compared to CBA. Although much research has been carried out on (innovative) technology-based assessments in many educational fields, considerably less

research has been devoted to the use of technology-based assessment as an equivalent to PBA in a more practical educational field (i.e., vocational education and training). The purpose of the current study, therefore, is to design, develop, and evaluate an MBPA for credentialing CSGs in Dutch vocational training.

We aim to contribute to the theory and practice of innovative assessments by presenting, developing and validating a new type of assessment: the MBPA. We think that this computer-driven measurement instrument does not need to be a game or high-fidelity simulation, by definition. Instead, we argue that MBPA, a simulation type assessment, which is based only on audio/video material, animation, and interactive interface features, may suffice for measuring vocational constructs. Next, we discuss the design and development of an MBPA for CSGs in Dutch vocational training, before presenting the results of the empirical validation study in which we applied an extended version of the argument-based approach to validation (Wools, 2015).

5.2 Design and Development of a Multimedia-based Performance Assessment

The measurement instrument was designed and developed following a two-stage, twelve-step framework for designing and developing an MBPA (De Klerk, Veldkamp, & Eggen, *submitted for review*). The focus of the framework is on the integrative, iterative and adaptive character of the design and development of an MBPA (see Figure 4.2). In the framework, design and development are regarded as parallel operations, which are carried out simultaneously from the earliest moments of MBPA design and development. In addition, design and development are part of a continuous process of monitoring developmental progress in relation to the MBPA's desired final state. If needed, the design and development process can be adapted to retain the alignment between the current state of development and the original purpose of the assessment.

We now discuss the framework in greater detail, and explain how we used it to build our MBPA. The framework encompasses two parallel stages—design and development—that consist of five steps each. Including the two general steps of purpose determination and fitness for purpose determination, which go beyond the two stages, the framework comprises twelve steps in total. The 12-step framework revolves around the concepts of integrated, iterative and

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

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

Before we actually began designing and developing, we first laid out the purpose of the assessment. In other words, we defined the proposed interpretation of the assessment scores. The purpose of the MBPA is to measure CSG skills, as defined in the "final attainment objectives," so that it can be used as a tool for making a CSG certification decision. The desired interpretation can only be met if we can make a chain of inferences, from actual student performance to the certification decision. Because this chain of inferences is part of the *interpretive argument* in the extended argument-based approach to validation, we will come back to this during the discussion of the validation study (Wools, 2015).

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

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

rial and knowledge for further specification of the constructs and attributes of the MBPA.

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

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

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

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

The next step was to determine whether the MBPA is fit for purpose. To answer this question, we refer to the entire bottom part of the framework presented in Figure 4.2. In other words, to answer this question, we first have to do a pretest, in which a representative sample of students performs in the MBPA. We then have to decide how to score student responses in the MBPA, and we apply a measurement model to analyze student scores. Finally, evidence has to be collected that supports the supposed interpretation of the MBPA scores, as defined in the very first step of the framework. The validation study, which is presented below, can be regarded as the evidence model because it will determine whether or not the MBPA is fit for purpose.

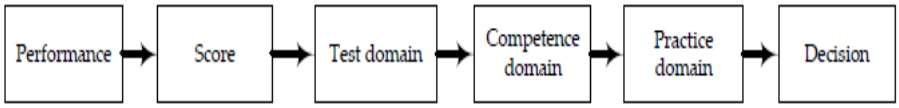
5.3 Validation of the Multimedia-based Performance Assessment

We applied an extended version of the argument-based approach to validate the interpretations assigned to the MBPA's scores (Wools, 2015). The original argument-based approach, developed by Kane (1992; 2004; 2006), presents a standardized framework for the validation process, consisting of a developmental stage, in which an interpretive argument is formed, and an appraisal stage, in which a validity argument is formed. The extended approach also includes an evaluation stage, in which the validation process itself is evaluated (Wools, 2015).

5.3.1 Interpretive Argument

The interpretive argument can be regarded as a chain of reasoning from student performance on the tasks in the MBPA to the decision whether or not to certify. Figure 5.1 shows the chain of reasoning in the interpretive argument of the MBPA. First, students' performance of the MBPA tasks is converted into a numerical score. Students get one point for each correct action or answer in the MBPA. The sum of the points is the total score. We consider the total score to be a representation of the test domain score, which consists of all possible tasks that could be administered in the MBPA. In the following step of reasoning, we extrapolate the test domain score as a representation of the skills or competences to be measured. In other words, we expect that a high MBPA score will correspond with a strong mastery of CSG skills and that a low MBPA score will correspond to weak mastery of CSG skills. Finally, the scores are extrapolated from the competence domain to the practice domain. In other words, can we regard students' MBPA scores not only as a representation of skill or competence within the assessment context, but also outside the assessment context, in their future professional life (Gulikers, 2006)? Only if we can make this extrapolation, can we take a meaningful decision about certifying CSG students, based on MBPA performance.

Figure 5.1
Chain of Reasoning in the Interpretive Argument (Adapted from Wools (2015))



5.3.2 Validity Argument

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

5.3.3 Analytical Validity Evidence

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

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

empirical evidence for the validity argument of the MBPA. Finally, we evaluate the validity argument.

5.3.4 Method

Participants. The participants in the empirical study were 55 CSG students (1 female and 54 male). Participants ranged in age from 19 to 64 years, with a mean age of 40.4 years (σ=11.5). They were asked to participate in the MBPA after they had completed training. Participation was voluntary and students did not receive a payment. All participants were recruited at one of the two locations where they had received their vocational training, the PBA and the pen-and- paper multiple-choice knowledge test. The 55 participants were a representative sample of the population for two reasons. First, all participants had just completed the course. Secondly, there is no reason to assume that the sample differs from the general population in age, ethnicity, or education. We can therefore assume that the sample of students included in this experiment is a representative sample.

Materials. *Knowledge Test.* Immediately after training, the students performed a knowledge-based pen-and-paper test, consisting of 21 multiple-choice questions, each offering 3 alternatives. According to the assessment regulations, as determined by the assessment commission, the students had 30 minutes to complete the test and had to answer 14 or more questions correctly to pass the test. The test is composed of randomly selected items from an item bank, and reliability cannot therefore be calculated.

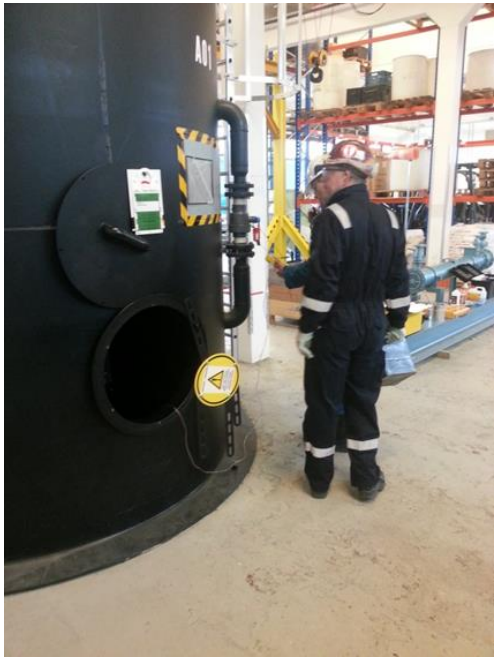
Performance-based Assessment (PBA). In the PBA, students performed CSG tasks in a reconstructed, yet realistic situation. Before the PBA started, students were randomly assigned to one of four scenarios that would be played during the PBA. A scenario always started with receiving the work permit from 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 then had to follow the correct emergency procedures. The assessment ended when the student and the worker were both at the emergency gathering point and had notified the operator that they were safe. Figures 5.2 and 5.3 give an idea of what the PBA looks like.

The rater used a rubric consisting of 19 criteria to evaluate the student ($\alpha = .35$). All 19 criteria were marked as *insufficient* or *sufficient* by the rater. From the 19 criteria (e.g., "tests the walkie-talkie"), 9 are considered to be knock-out criteria (e.g., "recognizes and reacts to an alarm"). If a student's performance in any one of these criteria is insufficient, he or she will fail the PBA. Because the rubric consists of rather narrowly-defined actions, we decided to develop an additional rubric of 12 criteria ($\alpha = .8$) that focused on the upper-level (behavioral) constructs of the CSG vocation. These constructs comprise communication, proactivity, environmental awareness, and procedural efficiency. Raters were asked to assess students on a scale ranging from 0 (e.g., "Student does not demonstrate any communication skills") to 3 (e.g., "Student communicates very well"). Hence, students could get between 0 and 12 points for the new criteria, and between 0 and 19 points for the original criteria. Both rubrics were marked by the rater. We also calculated the combined score of both rubrics ($\alpha = .73$). Translated versions of these rubrics can be found in Appendix 5A.

Figure 5.2 and 5.3

The Confined Space Guard is Checking the Environment of the Confined Space for Potentially Dangerous Features (5.2) And the Confined Space Guard (with White Helmet) Discusses Communication with Worker (5.3). The Rater Observes from a Distance.



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

Students were required to watch multimedia elements and to answer questions in the MBPA. For example, students were presented with a digital work permit that they could inspect using the zoom and navigation buttons in the MBPA. They then had to answer the question accompanying the work permit (e.g., "Is the first column of the work permit completed correctly?"). In this case, both the work permit and the question were presented simultaneously. Where students had to watch a film fragment, the film fragment was presented first, followed by the question. In this way, students were confronted alternately with multimedia and different types of questions. The MBPA consists of 35 questions: 18 yes/no questions, 5 multiple-choice questions (with 4 options), 4 fill-in-the-blank questions, 1 multiple-response question, 1 rank order question, and 6 so-called intervention questions. The intervention questions required students to watch two videos of workers cleaning a tank, and to intervene whenever their actions were incorrect. Students could intervene by clicking on a big red stop button that was located beside the video screen. Students were told that they could only click the stop button three times. In other words, if they clicked the stop button when there were no faulty actions, they had one less chance to press the button when it was required. Figures 5.4 and 5.5 illustrate the MBPA.

Figure 5.4 and 5.5
MBPA Screenshots

Questionnaire. After students had performed in the MBPA, they were asked to complete a questionnaire comprised of 15 items (N=15) addressing the following: 1) their background characteristics (e.g., "What is the highest level of education you have completed?"); 2) computer experience (e.g., "On a scale ranging from 1 (never) to 5 (every day) - How often do you play videogames on a computer?"); and 3) MBPA usability (e.g., "On a scale ranging from 1 (strongly disagree) to 5 (strongly agree) - I was comfortable with the interface of the MBPA"). Reliability for computer experience and usability were $\alpha = .64$ 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 the Internet and computers at home, developed by Cito (Cito, 2014). As a result, students' computer use and the usability of the MBPA could be classified as subscales of the questionnaire. A translated version of the questionnaire can be found in Appendix 5B.

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

5.3.5 Empirical Validity Evidence

The first element of empirical validity evidence was constructed around the internal structure of the MBPA. The internal structure of the assessment was assessed through a psychometric analysis (based on classical test theory (CTT)) of the answers that the students gave. As mentioned earlier, the assess-

ment was composed of 35 items. In total, students could get one point for each correct answer. The mean score on the test was 22.5 (σ=3.44), with a 95% confidence interval [21.6, 23.6], which indicates that the test was quite difficult for the students. The maximum score (obtained by two students) was 30, whereas the minimum score was 14 (N=1). The standard deviation is rather low, which means that most students achieved a score around the mean. The average time that students needed to complete the assessment was 29 minutes (σ=8). The minimum amount of time spent on the assessment was 19 minutes, and the longest was 58 minutes. The high standard deviation and the wide bandwidth between minimum and maximum indicate that there is a lot of variance between the time students spent on the assessment. Table 5.1 provides mean, standard deviation, and confidence intervals for the MBPA.

Table 5.2 provides other characteristics of the MBPA. The variation of the scores is relatively high (11.9). The distribution of the scores is not skewed (.014), but the kurtosis is high (.488). This indicates that a relatively large portion of the variance is caused by extreme values on both sides of the distribution, with most students' scores being clustered around the mean. In addition, the standard error of measurement (SEM), which was calculated by multiplying the standard deviation of the assessment scores by the square root of 1, the reliability, is .83. The SEM was defined as the standard deviation of the mean of a hypothetical normal distribution of many administrations of the same test. In other words, it represents the possible distance between the observed score and the true score in a CTT context. For the MBPA, the SEM was relatively low, which means that the students' true score was relatively close to their observed score. To be precise, 95% of the scores on the hypothetical distribution for a student with a mean score of 22.5 fall between 20.8 and 24.2. The reliability of the MBPA is high—with a Greatest Lower Bound (GLB) of .94. Of course, high reliability is in accordance with a low SEM. We looked at the best indicator of reliability, the GLB ((see Verhelst (2000) and Sijtsma (2009)).

To establish further support for the internal structure of the MBPA, we also looked at CTT indices for the individual tasks and assignments in the MBPA. Table 5.3 displays the CTT indices for the 35 items in the test. The second and third columns indicate what part of the content was assessed with the particular item, and which item format was used. The p-value of the test is the proportion of students that answered the item correctly. That is, a high p-

value is associated with a relatively easy question, whereas a low p-value points to a difficult question for this group of students.

Table 5.1
Mean, Standard Deviation, and 95% Confidence Interval for Measures (1000 Sample Bootstrapping Performed)

Measure	Mean	σ	Rel.	95% CI	
				Lower	Upper
MBPA (35p)	22.54	3.45	.94	21.59	23.57
PBA (19p)	17.35	1.52	.35	16.87	17.72
PBA (12p)	9.11	2.45	.80	8.39	9.87
PBA (total)	26.25	3.78	.73	25.15	27.25
MC Test (21p)	17.89	1.85	-	17.35	18.39
MBPA Time (minutes)	29.2	8.33	-	26.91	31.85
Q-Computer exp.	2.88	0.91	.64	2.6	3.12
Q-MBPA usability	2.93	0.76	.75	2.72	3.14

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

Table 5.2
MBPA Test Characteristics (1000 Sample Bootstrapping Performed)

M	GLB	S^2	SEM	Skewness	Kurtosis
22.54 (σ = 3.45)	.94	11.1	0.83	.014	.488

Note. M = the mean score on the test; GLB = Greatest Lower Bound; S^2 = the variance of the scores; SEM = the standard error of measurement

The mean p-value of the test was .62, which indicates that the assessment was rather difficult for this group of students. The r_{it}, the correlation between the item scores and the total test score, gives an indication of the discriminative power of an item. The higher the value, the better the item can discriminate between good and poor performers. A low r_{it} means that students that score

high on the overall test score low on the item, whereas students that score low on the overall test score high on the item. Conversely, a high r_{it} means that good performers do well on the item and poor performers do worse on the item. The mean r_{it} of the MBPA is .22. There is reason to believe that the mean r_{it} could be improved, as the quality of the individual items in the MBPA fluctuate (see Table 5.3). To summarize, we have provided evidence that the MBPA has a strong internal structure. Although there is room for improvement, all indices fall within acceptable levels.

Table 5.3
CTT Indices of 35 Items in the Multimedia-based Performance Assessment

Item	Content	Type	P-value	r_{it}
1	Explain WP	MC-4	.36	.38
2	Ask for addition WP	Yes/No	.76	.26
3	Ask for addition WP	Yes/No	.82	.40
4	Ask for addition WP	Yes/No	.44	.29
5	Ask for addition WP	Yes/No	.89	.31
6	Ask for addition WP	Yes/No	.72	.37
7	Ask for addition WP	Yes/No	.13	.01
8	Ask for addition WP	Yes/No	.89	.24
9	Check WP	Yes/No	.36	.33
10	Check WP	Yes/No	.86	.05
11	Check WP	Yes/No	.87	-.03
12	Check WP	Yes/No	.44	.37
13	Explain WT	MC-4	.87	.22
14	Channel WT	MC-4	1	-
15	Test WT	Fill in	.95	.23
16	Battery check WT	Fill in	.15	.01
17	Escape plan	MC-4	.49	.53
18	Work preparation	Yes/No	.96	-.07
19	Work preparation	Yes/No	.95	-.16
20	Work preparation	Yes/No	.18	.23
21	Work preparation	Yes/No	.86	.05
22	Work preparation	Yes/No	.60	.12
23	Work preparation	Yes/No	.33	.46
24	Work preparation	Yes/No	.89	.07
25	Environment check	Multiple sel.	.34	.37
26	Report to operator	MC-4	.96	.05
27	Agree communication	Fill in	.53	.13
28	Error intervention	Intervention	.18	.26
29	Error intervention	Intervention	.63	.40
30	Error intervention	Intervention	.31	.24
31	Error intervention	Intervention	.27	.56
32	Error intervention	Intervention	.29	.35

Table 5.3 (continued)
CTT Indices of 35 Items in the Multimedia-based Performance Assessment

Item	Content	Type	P-value	r_{it}
33	Error intervention	Intervention	.87	.06
34	React to emergency	Rank order	.81	.16
35	Report to operator	Fill in	.89	.16
Mean			.62	.22

Note. P-value = the proportion of the students who have answered the item correctly; r_{it} = the correlation between the item score and total score.

The next element of empirical validity evidence is used to support the external structure of the MBPA, in particular the convergent validity (based on the PBA scores) and the discriminant validity (based on the questionnaire and the multiple-choice knowledge test). In other words, the MBPA scores are expected to correlate with the PBA scores, but not with the questionnaire and the multiple-choice test. The correlations (Spearman's rho) between the measures of the experiment are presented in Table 5.4. Spearman's rho is used because there is a monotonic relationship between the variables and because the measures do not meet the assumptions of normality and linearity. For example, on the 19-point rubric, most students score 17 to 19 of the criteria as correct. It is therefore better to look at the rank order of the scores on the different measures than at the linear correlation. As can be seen, the correlations between the MBPA and the rubrics used in the performance assessment are .37 (p<.01) and .38 (p<.01), respectively, for the 19-point rubric and the 12-point rubric, which are both significant. We have also combined students' scores on both rubrics to get a *total rubric score*. The correlation between the total rubric score and the MBPA score is strongly significant (r_s=.43 (p<.01)). We also applied a correction for attenuation, and found that the correlation then improves considerably (respectively to .92, .59, and .70). This indicates that the correlation is strongly diluted by measurement error. Measurement error may result from one or both parameters (i.e., the MBPA score or one of the PBA scores).

Of course, there is also a strong significant correlation between both rubrics used in the assessment (r_s=.68, p<.001). We also performed a linear regression analysis to see the extent to which performance in the MBPA could predict performance in the PBA. Because of the negative skew of the distribution of the rubrics, especially the 19-point rubric, we first subtracted each score

from the highest score obtained, plus one, and then performed a log transformation (see Field, 2009).

Table 5.4
Correlations, Means, and Standard Deviations of Measures (1000 sample bootstrapping performed)

Measure	mean	σ	1	2	3	4	5	6	7
1. MBPA	22.5	3.5							
2. PBA (19)	17.4	1.5	.39**						
3. PBA (12)	9.1	2.5	.38**	.68***					
4. PBA (total)	26.3	3.8	.43**	.84***	.96***				
5. MC Test	17.9	1.9	.30*	.2	.21	.23			
5. MBPA (time)	29.2	8.3	.01	-.13	-.2	-.22	-.05		
6. Q-Computer exp.	2.9	.9	.09	.12	.15	.16	-.01	.1	
7. Q-MBPA usability	2.9	0.8	.18	.15	.09	.16	-.06	-.18	.42**

Note. *p<.05; **p<.01; ***p<.001

We did this for the 12-point rubric, the 19-point rubric, and the total rubric score to get a reliable comparison. The regression analysis for the 19-point rubric showed a significant effect ($F(1,53)=4.365$, $p<.05$), which indicates that the MBPA score can account for 7.6% of the variation in the PBA score. We performed the same analysis for the 12-point rubric, which was also significant

($F(1,46)=5.544$, $p<.05$), with an explained variance of 10.1%. Finally, we performed a regression analysis for the total rubric score, which was also significant ($F(1,46)=5.905$, $p<.05$), with an explained variance of 11.4%. The total rubric score was the best predictor for performance in the MBPA. Unfortunately, the rater forgot to complete the 12-point rubric on one assessment, which explains the lower number of students in the second analysis.

To provide further evidence, if MBPA performance is related to PBA performance, then we would expect students who had failed their PBA to score significantly lower on the MBPA than students who had passed their PBA. Unfortunately, the group of students was rather small (N=8), which makes it quite difficult to interpret the results and draw definitive conclusions. The group of students who passed the PBA had a mean score of 23.2 (σ=.46) and the group of students who failed had a mean score of 20.1 (σ=1.1). We used an independent samples t-test to check whether the groups differed significantly, which was the case ($t(53)=-2.563$, $p<.001$). We then performed a logistic regression analysis to check the extent to which the MBPA score could predict whether a student will pass or fail their PBA. The MBPA score is treated as a continuous predictor in the logistic regression analysis. The dependent variable (success in PBA) is a dichotomous outcome variable (0=failed, 1=passed). The results of the analysis can be found in Table 5.5. The analysis demonstrated that the MBPA score made a significant contribution to predicting whether students failed or passed the PBA ($\chi^2(1, 55) = 5.09$, $p<.05$). The odds ratio (e^{β}) for the BPA score is 1.39 with a 95% confidence interval [1.04, 1.86]. This suggests that a one unit increase in the MBPA score increases the probability of being successful in the PBA (i.e., passing the PBA), with 1.39. To summarize, the overall correlations and regression analysis provide evidence for the convergent validity of the MBPA.

The absence of a correlation between students' MBPA scores and their background characteristics, the questionnaire ratings (computer experience and usability of the MBPA) and multiple-choice test results should provide evidence for discriminant validity. The background characteristics are age, education, and ethnicity. Age was not correlated with assessment scores (r_s=.00, $p>.05$). We calculated the biserial correlation coefficient for education. The biserial correlation coefficient is used when one variable is a continuous dichotomy (Field, 2009).

Table 5.5
Logistic Regression Analysis of Passing Performance-based Assessment

Predictor	β (SE)	Wald's χ^2 (df=1)	p	e^{β}	e^{β} (95% CI) L	U
Constant	-5.4	3.05	0.08	0.00		
MBPA Score	0.33	5.09	0.02	1.39	1.04	1.86

Note. The dependent variable in this analysis is Performance-based Assessment success coded so that 0 = failed and 1 = passed

First, we divided the students into two groups (low education vs. high education). The low education group consisted of students who had continued education up to high school or lower vocational education (N=26, M_{MBPA}=21.83), whereas the high education group consisted of students who have had continued education from middle level vocational education and upwards (N=27, M_{MBPA}=23.08). We calculated the point-biserial correlation (which is for true dichotomies (Field, 2009)), and then converted it into the biserial correlation. Although education and student MBPA score were positively correlated, this effect was not significant (r_b=.19, p>.05). For ethnicity, we were especially interested in two groups: students with Dutch ethnicity (N=40, M_{MBPA}=22.8), and students with another ethnicity (N=15, M_{MBPA}=22.78). We calculated the point-biserial correlation between ethnicity (0=Dutch, 1=other) and the students' MBPA scores. Again, we did not find a significant correlation (r_{pb}=-.01, p>.05). Overall, student's background characteristics were not related to their MBPA performance, which supports the discriminant validity of the MBPA.

We found further support for discriminant validity, because there is no significant correlation between student MBPA scores and their computer experience (r_s=.09, p>.05). Additionally, the MBPA score and student rating on the usability of the MBPA are not correlated (r_s=.14, p>.05). It is interesting to note that there is a significant correlation between students' computer experience and their rating of the usability of the MBPA (r_s=.42, p<.01), but that there is no significant correlation between the time spent on the MBPA and the score obtained (r=.07, p>.05).

However, there is a significant correlation between the multiple-choice knowledge-based test and the MBPA (r_s=.3, p<.05), which may indicate that, at least to some extent, the multiple-choice test and the MBPA do measure the same construct(s). Interestingly, there is no significant correlation between the PBA scores and the multiple-choice test scores (r_s=.09, p>.05).

Finally, we determined the number of misclassifications at six different levels of MBPA cutoff scores (50%, 55%, 60%, 65%, 70% and 75%). No misclassifications would mean that all students (N=8) that failed their PBA would also fail the MBPA, and that all that passed the PBA would also pass the MBPA (N=47). The results are presented in Table 5.6. Although the lowest cutoff percentage (50%) results in the least number of misclassifications, which can be explained by the fact that we have a small group of students who failed their PBA, it is most interesting to note the difference in fail-fail classifications between the cutoff points at 55% and 60%. At the 55% cutoff point, only two students who failed their PBA would also fail the MBPA, whereas this number increased to 7 at the 60% cutoff score. Therefore, a cutoff score at approximately 60% would be most defensible empirically. In addition, we looked at the number of misclassifications at different levels of cutoff scores using Cronbach's alpha in TiaPlus (Cito, 2006). The analysis indicates that the least misclassifications take place when the cutoff score is placed at 50% (see also Table 5.6). In TiaPlus, the GLB reliability coefficient is the point of departure to estimate the misclassifications at the different cutoff levels.

Table 5.6
Number of Misclassifications MBPA–PBA at Different Cutoff Score Levels

Cutoff percentage		Performance-based Assessment		N misclassifications
		Fail	Pass	
MBPA-50%	Fail	2	2	8
	Pass	6	45	
MBPA-55%	Fail	2	3	9
	Pass	6	44	

Table 5.6 (continued)
Number of Misclassifications MBPA–PBA at Different Cutoff Score Levels

Cutoff percentage		Performance-based Assessment		N misclassifications
		Fail	Pass	
MBPA-60%	Fail	7	12	13
	Pass	1	35	
MBPA-65%	Fail	7	30	31
	Pass	1	17	
MBPA-70%	Fail	7	35	36
	Pass	1	12	
MBPA-75%	Fail	8	39	39
	Pass	0	8	

5.3.6 Validity Evaluation

In the previous paragraph, we presented validity evidence. In this paragraph, the validity evidence is used and evaluated. The argument-based approach is applied to prove the proposed interpretation of the MBPA. Wools (2015) distinguishes three criteria to evaluate the validity and the process of validation for an assessment. The first criterion evaluates the interpretive argument, the second criterion evaluates the different elements of validity evidence, and the third criterion evaluates the validity argument as a whole.

With regards to the first criterion, we can say that there are a substantial number of inferences. Following the chain of inferences, we have to go from a student's performance and accompanying raw scores to meaningful statements regarding performance in a practice domain, and then to a final certification decision. This is an indication of the complexity of the inferences that we wish to make with the MBPA. Nevertheless, these inferences are required to ensure

that the MBPA can be used for its intended purpose. The question concerns whether the interpretive argument addresses the correct inferences and assumptions (Wools, 2015). We specified the interpretive argument in sufficient detail so that the chance on possible voids or inconsistencies in our inferred reasoning is kept at a minimum. According to the extended argument-based approach to validation, each inference in the chain (the arrows in Figure 5.1) should have at least a warrant, a supporting warrant (or backing), and rejected rebuttals. A rebuttal indicates a circumstance in which the warrant or backing would not hold (Wools (2010)). We can demonstrate this by looking at each inference in the chain individually. The first inference is from performance to score, or the scoring inference. The same performance always leads to the same score (warrant), but this will only hold if the MBPA is correctly programmed (rebuttal). Furthermore, there has to be an objective scoring system (backing), which needs to be used objectively (rebuttal). In our case, the MBPA has a standardized and objective scoring structure. Scoring has already been addressed in the *assessment skeleton*, which was made in collaboration with SMEs.

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

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

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

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

The second criterion for validity evaluation relates to the validity evidence. Is the presented validity evidence plausible and representative for the assumptions that we wish to make with the MBPA scores? In other words, are the inferences justified by our validation study (Wools, 2015)? Each element of validity evidence should relate to and substantiate one or more inferences in the chain of reasoning. Wools indicates that an evaluation status should then be assigned to the inference as a whole. The status is justified when warrants and backings on the validity elements are accepted, and possible rebuttals are rejected. With the evidence presented above, we argue that the validity elements give enough support for all the inferences in the interpretive argument.

Finally, the third criterion focuses on the outcome of the validation process or the validity argument as a whole. The question to be answered is: Is the validity argument as a whole plausible (Wools, 2015)? The validity argument can only be plausible when both the first and second criteria are met. As with the second criterion, the third criterion is somewhat subjective but boils down to taking all elements of validity evidence into account and then deciding whether the argument is strong enough to substantiate the validity of the assessment scores and final interpretations. In our case, we can say that all criteria have been met. The validity evidence provided in this chapter is plausible be-

cause every inference in the chain of reasoning, from performance to decision, can be substantiated by evidence.

5.4 Discussion and Conclusion

The aim of this study was to investigate the design, development and validation of an MBPA for credentialing CSGs in Dutch vocational education. The first part of the chapter focused on design and development. In particular, a 12-step framework was used. This was specifically built for the design and development of the MBPA. The second part of the chapter focused on the validation of the developed MBPA, using an extended version of the argument-based approach to validation, as presented by Wools (2015).

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

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

The first analytical element of validity evidence referred to the content validity of the instrument. We demonstrated that the structured design and development process, through the use of the framework, ensured that the content of the MBPA was a full and representative sample of the content domain. We have also shown how different steps in the framework secured the correct complexity of tasks and assignments in the MBPA, which was the second ana-

lytical validity element. More specifically, the process from construct analysis to task design in assessment templates, through a cognitive walkthrough of the assessment, offered the chance to design tasks at the correct level of (cognitive) complexity. These elements of validity evidence followed naturally from using a comprehensive framework for the structured design and development of the MBPA. The other three elements are based on the empirical data produced by students performing in the MBPA.

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

A second empirical element of validity evidence was based on the external structure of the assessment. The convergent and discriminant validities of the MBPA were used as indicators of the strength of the external structure. Students' MBPA scores were correlated with their PBA scores, as evidence for the convergent validity of the instrument. Students' scores on the PBA (by independent rubric and total rubric score) moderately correlated with their scores on the MBPA. The fact that the correlation is not stronger may be because of several reasons.

First, the rubrics used for rating student performance on the PBA do not show much variance in sum score. We had foreseen this problem for the 19-point rubric and therefore developed the 12-point rubric to induce more variation in student PBA scores. And, indeed, it does produce slightly more variance in students' scores, but not enough to make a real difference. It is statistically difficult to establish strong relationships between two variables when one of the

variables has almost no variance. Secondly, the correction for attenuation on the correlations indicated that there is a stronger relationship between the measures, but that it is diluted by measurement error. Finally, although the PBA is the best external criterion, it may not be the ultimate criterion, because PBA might have (psychometric) limitations, and we cannot be certain about the quality (i.e., validity and reliability) of the scores. Future research in this area should also try to find criteria that are outside of the assessment domain. One external criterion could, for example, be the students' future job appraisals, made by their managers. A future study on the subject could include a strong analysis of the quality of the PBA, for example, through generalizability theory (Brennan, 2001).

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

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

The third empirical and fifth overall element of validity evidence related to the number of misclassifications at different cutoff score levels. Of course, the number of misclassifications is related to the reliability of the MBPA, which is high. We used the GLB index (Verhelst, 2000; Sijtsma, 2009), which provides the best estimate of reliability by giving the greatest lower bound of the reliabil-

ity. That means that the reliability of the test is at least as high as the GLB indicates. In this case, the GLB is .94.

There are some general limitations to our study. First, the sample size is small. It was difficult to get a substantial number of students to participate in the study, because many assessment locations do not have internet connections or computers and the locations themselves are spread all over the Netherlands. The assessment itself takes place, on average, 15 times per year, per location. Sometimes, a group can consist of less than five students, which suggests how difficult it can be to get a sufficient number of students to participate. On the other hand, because there are not many students per year and we have collected data for seven months, we can say that we have included a substantial amount of data in our study. Furthermore, if we look at background, the sample does not systematically differ from the general population. Another limitation is the quality of the PBA. Although the PBA is professionally organized, only one rater is used. The rater also plays the part of the operator in the assessment. The 19-point rubric, used for rating a students' performance, shows little to no variance, which makes it difficult to draw firm conclusions regarding a comparison between the MBPA and the PBA.

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

References

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

Brennan, R.L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Cito (2006). *TiaPlus. Test and Item Analysis*. Arnhem: Cito.

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

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

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373–399.

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

De Klerk, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2014). A blending of computer-based assessment and Performance-based Assessment: Multimedia-based Performance Assessment (MBPA). The introduction of a new method of assessment in Dutch vocational education and training (VET). *CADMO, 22*(1), 39–56.

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

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M (2014). *A framework for designing and developing Multimedia-based Performance Assessment.* 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.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: SAGE Publications Inc.

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

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

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*, 135–170.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport: American Council on Education and Praeger Publishers.

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

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).

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).

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

Parshall, C.G., Spray, J.A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

Roelofs, E.C., & Straetmans, G.J.J.M. (Eds.) (2006). *Assessment in actie* [Assessment in action]. Arnhem: Cito.

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).

Quellmalz, E.S., & Pellegrino, J.W. (2009). Technology and testing. *Science, 323*, 75–79.

Quellmalz, E.S., Timms, M.J., & Buckley, B. (2010). The promise of simulation-based science assessment: The Calipers Project. *International Journal of Learning Technology, 5*(3), 243–263.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved [March 20, 2012] from http://www.jtla.org.

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

Shavelson, R.J., Ruiz-Primo, M.A., & Wiley, E. (1999). Note on sources of sample variability in science performance assessments. *Journal of Educational Measurement, 36*(1), 56–69.

Shute, V.J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias and J.D. Fletcher (Eds.), *Computer Games and Instruction* (pp. 503–523). Charlotte, NC: Information Age Publishing.

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

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

Wools, S. (2015). *All about validity. An evaluation system for the quality of educational assessment* (Unpublished doctoral dissertation). RCEC, Enschede.

Appendix 5A

19-point rubric

No.	Criterion	Insufficient	Sufficient
1	Carries out last minute risk analysis correctly.		
2	Wears the prescribed personal protective equipment (PPE).		
3	Asks the operator for explanations about the work permit.		
4	Asks the operator about the functioning of walkie-talkies.		
5	Discusses communication with workers in the confined space.		
6	Discovers and reports deviations in the work permit.		
7	Carries out actions in accordance with the work permit.		
8	Determines wind direction and the correct escape route.		
9	Explores the confined space environment before operations.		
10	Asks about the dangers of the last substance stored in the confined space.		
11	Tests the walkie-talkies at the confined space.		
12	Registers the workers entering and leaving		

	the confined space.		
13	Recognizes and corrects incorrect PPE of workers.		
14	Recognizes and reacts to an alarm.		
15	Alerts workers in the confined space during an alarm.		
16	Verifies the number of workers leaving the confined space during an alarm.		
17	Reports the alarm to the operator via the walkie-talkie.		
18	Leaves the confined safe and tidy.		
19	Attaches a "do not enter" sign to the confined space.		

12-point rubric

Construct/Points	0	1	2	3	Rater judgment
Communication	Student shows (almost) no communication skills	Student shows poor communication skills	Student shows sufficient communication skills	Students shows good communication skills	

Proactive attitude	Student has (almost) no proactive attitude	Student has poor proactive attitude	Student has sufficient proactive attitude	Student has good proactive attitude	
Environment awareness	Student (almost) doesn't observe environment	Student observes environment poorly	Student observes environment adequately	Student observes environment well	
Procedural efficiency	Student (almost) doesn't follow procedures	Student follows procedures poorly	Student follows procedures adequately	Student follows procedures well	
Total number of points:					

Appendix 5B

Questionnaire
General questions (6 questions)

1. *Sex.*

   O M
   O F

2. *Place of residence.*

   _____

3. *Age*

   ____

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

   O None
   O Elementary school
   O Lower level high school
   O Lower vocational/technical education
   O Middle level high school
   O Middle level vocational/technical education
   O Higher level high school
   O Higher level vocational/technical education
   O University
   O Postgraduate education

5. *What is your ethnic origin?*

   O Dutch
   O Turkish

O Moroccan
O Surinam
O Antillean
O Other. Please specify:_____

Computer use (3 questions)

6. *Do you have a computer?*

O No
O Yes, but no internet
O Yes, with internet

7. *Do you use a computer frequently?*

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

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

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

Multimedia-based Performance Assessment (7 statements)

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

O Strongly disagree
O Disagree
O Neutral

O Agree
O Strongly agree

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

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

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

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

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

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

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

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

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

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

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

O Strongly disagree
O Disagree
O Neutral
O Agree
O Strongly agree

Chapter 6. A Methodology for Applying Students' Interactive Task Performance Scores from a Multimedia-based Performance Assessment in a Bayesian Network[7]

Abstract

Computer-based simulations are increasingly being used in educational assessment. In most cases, the simulation-based assessment (SBA) is used for formative assessment (Mislevy et al., 2014), but as research on the topic continues to grow, possibilities for summative assessment are also emerging (De Klerk, Veldkamp, & Eggen, 2015). The current study contributes to research on the latter category of assessment. In this chapter, 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 chapter: the *evidence identification challenge* (i.e., scoring interactive task performance), and the *evidence accumulation challenge* (i.e., accumulating scores in a psychometric model). The methodology is illustrated by analysis of the interactive task performance data of 57 students who completed the MBPA.

[7] This chapter is a minor revision of De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2015). *A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian network.* Manuscript submitted for publication.

6.1 Introduction

The use of interactive computer-based simulations in educational assessment is growing together with research on the topic (see for example, Quellmalz & Pellegrino, 2009; Clarke-Midura & Dede, 2010; Shute, 2011; Mislevy et al., 2014; De Klerk, Veldkamp, & Eggen, 2015). 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 and GBAs (Prensky, 2001; Squire, 2003; Mislevy et al., 2014). 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 et al., 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.

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 chapter, 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. The MBPA is 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 chapter 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.

6.2 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 (Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991; Shavelson, Baxter, & Gao, 1993; Brennan,

2000). Furthermore, PBA performance is usually scored on a rubric by a human rater. Both rubrics (Shepherd & Mullane, 2011) and human raters have been found to negatively influence the reliability of the assessment (Dekker & Sanders, 2010). 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, Chung, & Delacruz, 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 et al., 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 tasks is high and dependencies across observations are often caused by the constantly changing state of the game (Rupp, Gushta, Mislevy, & Williamson, 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) (Mislevy, Almond, Yan, & Steinberg, 2000; Levy & Mislevy, 2004; West et al., 2010; Levy, 2014). 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, 2001).

Shute, Ventura, Bauer, and Zapata-Rivera (2011) 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 chapter, 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.

6.3 Method

6.3.1 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 (σ=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. Sec-

ondly, the sample in the experiment reflects the general student population with respect to age, ethnicity and education.

6.3.2 Materials

Multimedia-based Performance Assessment (MBPA). The MBPA is a flexible and interactive assessment that is accessible through the internet. In the MBPA, the authentic work setting and real-life equipment are simulated by images and video fragments. The MBPA has been designed according to the design framework presented in Chapter 4. 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 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 tasks 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. A screenshot of the MBPA is shown in Figure 6.1.

Figure 6.1

Interface of the MBPA. The Toolbox with a Few Tools is Shown at the Right Side of the Screen. The Stop button, at the Top Right is Shown for the Intervention Questions. Below is a Bar in which Instructions for the Student are Shown.

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 video-

games 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 5B.

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. Figures 5.2 and 5.3 give an idea of what the PBA looks like.

The rater used a rubric made up of four criteria (α = .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 translated version of the rubric can be found in Appendix 5A.

6.3.3 Procedure

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.

6.4. Results

6.4.1 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 dur-

ing 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 x 2 ratings x 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 6.1. 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 (Fleiss, 1981; Cicchetti & Sparrow, 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.

Table 6.1
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/TOT	Interrater Reliability	Interrater Agreement	α
Office	ESS	.31	.34	.73
	DIF	.38	.42	.79
Outside	ESS	.34	.37	.75
	DIF	.20	.22	.60
Alarm	ESS	.35	.39	.76
	DIF	.36	.41	.77

In Their average ratings for both essence and difficulty for all the actions in the MBPA are shown in Table 6.2. 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.

Table 6.2
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

Table 6.2 (continued)
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}
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

These average ratings were then used to calculate total scores for each of the actions in the MBPA. In Table 6.2, 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 6.2, 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 6.3, 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).

Table 6.3

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	.26	.24	.68
Outside	.29	.31	.71
Alarm	.26	.26	.67

Overall, the reliability and agreement indices indicate poor to fair agreement among the experts (Fleiss, 1981; Cicchetti & Sparrow, 1981). In Table 6.4, the mean ratings are presented. As can be expected, the rank order of the actions in Table 6.4 resembles the rank orderings in Table 6.2. However, the orderings are not identical. The reason is that the probability ratings in Table 6.4 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.

Table 6.4

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)

Table 6.4 (continued)
Experts' Average Probability Ratings that a Minimally Competent Student Would Successfully Complete the Action (from 1-highly improbable to 5-highly probable)

Setting	Action	$M^{PROB}(\sigma)$
Alarm	D3	2.17 (0.75)
	E3	3.00 (1.27)
	F3	3.50 (0.55)
	Avg. alarm	3.06 (0.70)

6.4.2 Application of a Bayesian Network (BN) on Student Scores—Evidence Accumulation Challenge

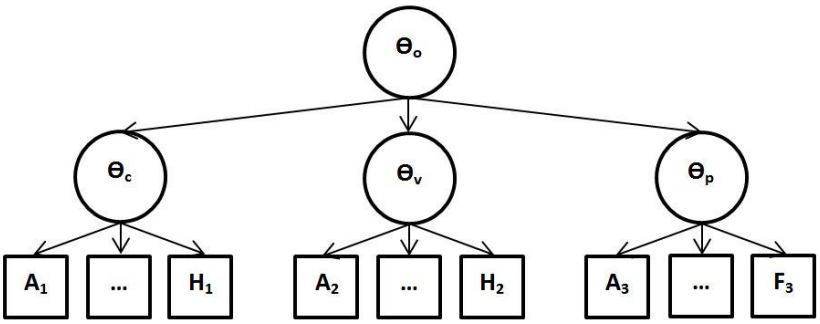
The ultimate goal of the MBPA is to make an informed and valid decision, based on student performance in the MBPA, about a student's overall KSAs to work as a minimally competent CSG. We therefore need to synthesize and aggregate student scores (i.e., the correct actions they took during the MBPA) to 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 (Rupp, Nugent, & Nelson, 2012; Mislevy et al., 2014). 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)"

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 Figure 6.2.

Figure 6.2

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.



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 Figure 6.2, 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 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 x least essential and most difficult x most essential). The ratings of the experts on difficulty and essence can be seen in Table 6.2. 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 6.5 for Model 1 (most difficult x least essential), and in Table 6.6 for Model 2 (most difficult x most essential).

Table 6.5
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

Table 6.5 (continued)

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

Setting	Action	C^{SCORE1}	*z-score*	Percentile
Alarm	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 6.6

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

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 6.5, we calculated the CPTs, which can be seen in Table 6.7.

Table 6.7
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

The difference between both models is demonstrated in Table 6.7. The experts' ratings indicated that action A1 (i.e., collecting the work permit by clicking on the work permit in the image) is highly essential, yet very easy, and the experts have rated action G1 (i.e., setting the walkie-talkie on the right channel after actions E1 and F1) as considerably less essential but a bit more difficult than action A1. The difference between both models can also be seen in Table 6.7. In Model 1, correctly performing action A1 in the MBPA slightly increases the probability that a student is in the sufficient category for θ_c,

whereas incorrect performance slightly decreases the probability that the student is in the sufficient category for θ_c. That is, action A1 is not a very informative OV regarding student proficiency in information communication (SMV). Action G1 is much more informative; correctly performing this action strongly increases the probability of a student being in the sufficient category, whereas incorrect performance strongly decreases this probability. Because the most essential actions have a stronger influence in Model 2, correctly performing action A1 increases the probability that a student is in the sufficient category for θ_c more than in Model 1. Yet, correctly performing less essential actions, like G1, have less influence on a student's 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 Figures 6.3 and 6.4) 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 Figure 6.3, for example, the line between ThetaC and A1 is thicker than in Figure 6.4. This shows that correct performance of action A1 in Model 1 influences the joint probability distribution of ThetaC 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 Figure 6.4 that the OV question marks (see Fig. 6.3) 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.

Figure 6.3

Influence Diagram of the Model 1 Bayesian Network for the Confined Space Guard MBPA

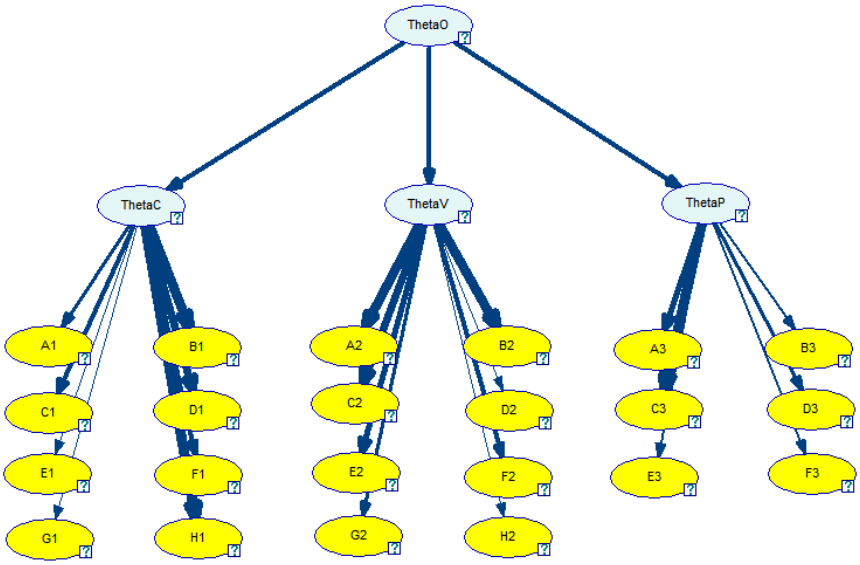
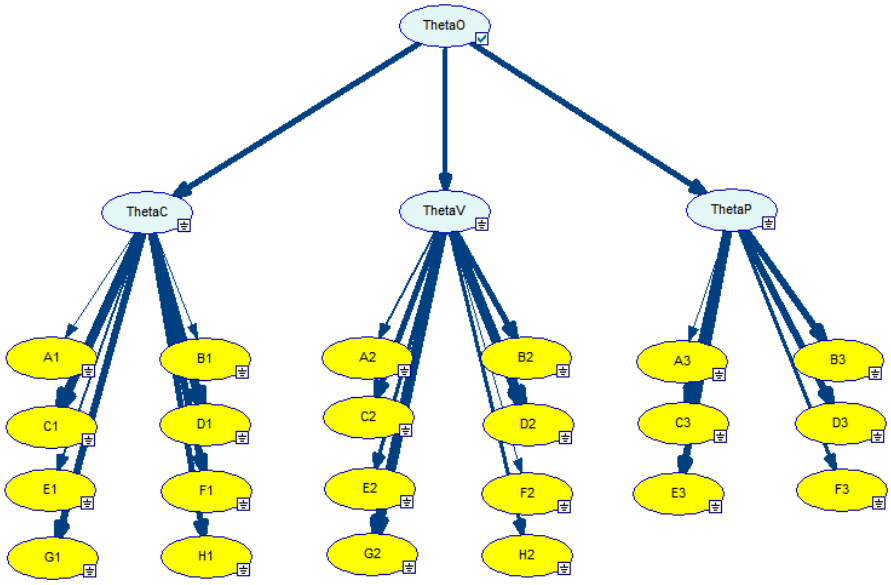


Figure 6.4

Influence Diagram of the Model 2 Bayesian Network for the Confined Space Guard MBPA



The clustering algorithm is the most basic and most widely-used algorithm for BNs (Almond, Mislevy, Steinberg, Yan, & Williamson, 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 6.4) 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 .66 (according to the experts' input). Therefore, the cutoff point is at .66. If the value of θ_c is .66 or higher, then we consider the student to be sufficient in that SMV ($\theta_v \geq .58$, $\theta_p \geq .62$). 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 6.8). 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 Figures 6.4 and 6.5 between the lower-level SMVs and the upper-level SMV are therefore equally thick.

Table 6.8
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

As the whole network has now been defined, we can enter the 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 6.9. 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 6.9. 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 least essential actions, but less well on the most essential actions. There is a strong correlation between both models, except for the CSG vigilance SMV (θ_c: $r(57) = .62$, $p < .01$, θ_v: $r(57) = .16$, $p > .05$, θ_p: $r(57) = .89$, $p < .01$, θ_o: $r(57) = .535$, $p < .01$). Notice that the marginal probabilities for θ_v in Table 6.9 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 6.9 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.

Table 6.9

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
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	9
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
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

Table 6.9 (continued)

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
M1 $\mu(\sigma)$	0.66 (0.46)	0.31 (0.42)	0.51 (0.48)	0.47 (0.30)	S 11.09	(3.20)
M2 $\mu(\sigma)$	0.45 (0.43)	0.28 (0.39)	0.46 (0.43)	0.36 (0.30)	P 9.33	(2.61)

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 6.9. 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 ($r(57) = .68$, $p = .00$ and $r(57) = .63$, $p = .00$, 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 ($r(57) = -.01$, $p > .05$) was significantly lower than the correlation between θ_o for Model 1 and the PBA rubric score ($r(57) = .225$, $p < .05$). 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.

6.4.3 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 .01 level for Model 1 ($r = .384$, $n = 57$), and also significant for Model 2 ($r = .265$, $n = 57$, $p < .05$). 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 num-

ber 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 ($r = .279$, 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 ($r = -.189$ and $r = -.175$, 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 chapter. 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 tasks 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.

6.5 Discussion and Conclusion

In this chapter, 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 chapter, 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.

A specific question for future research 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 are still rare (especially in a vocational setting) (Sinharay, 2006).

To conclude, this chapter 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 chapter 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.

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. 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 T.J.H.M. Eggen & B.P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 137–150). Enschede: RCEC.

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 Bayesian network example. *Computers & Education, 85*, 23–34.

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

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., & 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. (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).

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. *Proceedings of the EUNITE Workshop on Intelligent Systems in Patient Care*, Vienna, 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., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

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

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*, McGraw-Hill, New York, NY, USA.

Quellmalz, E., & Pellegrino, J. (2009). Technology and testing. *Science, 323*, 75–79.

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).

Rupp, A.A., DiCerbo, K.E., Levy, R., Benson, M., Sweet, S., Crawford, A., Caliço, T., Benson, M., Fay, D., Kunze, K.L., Mislevy, R.J., & Behrens, J. (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.

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., 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.

Shute, V.J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias and 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.

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., Crawford, A., DiCerbo, K.E., Chappel, K., & Behrens, J. T. (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).

Chapter 7. Epilogue

This thesis has made contributions to the theory and practice of Multimedia-based Performance Assessment (MBPA) in Dutch vocational education. In doing so, we hope to help future researchers and practitioners study and use MBPA. Ultimately, students should benefit from the research discussed here because improved and expanded assessment possibilities should lead to more reliable and fairer judgment of students' knowledge, skills, and abilities (KSAs). This chapter discusses the general question and the research questions that were presented in the General Introduction, but in a broader context. In addition, the one remaining question is answered. Finally, the strengths, limitations, and practical implications of the thesis are discussed, as well as directions for future research.

7.1 Research Questions

The general research question in this thesis was:

Can we develop and use a Multimedia-based Performance Assessment for which students' performance in the MBPA provides valid inferences regarding their knowledge, skills, and abilities?

This question could only be answered by addressing the research questions in Chapters 2 to 6. We now discuss the research questions, and then answer the general research question.

Research question 1: Why should we use MBPA?

This question was addressed through a literature study of Performance-based Assessment (PBA) and innovative CBT, and a pilot study on MBPA. A good reason for using MBPAs in vocational education is to diminish the measurement error found in PBA. Measurement error in PBA results from a complex of variables: the raters, the tasks, and the occasion. Because MBPAs are standardized, enable full objective scoring, and offer the possibility to present dynamic and multiple tasks, they can remedy the measurement error in an assessment program using PBA.

Research question 2: Are there psychometric models to analyze the performance data of MBPA?

The second research question has been answered through a systematic review of the literature on MBPA. Research on using innovative simulations for assessment purposes has grown considerably during the last few years. We presented many examples in Chapter 3. One specific psychometric model has excellent qualities for analyzing the student performance data that results from MBPA performance. We tested this model, which shows that there are indeed modern psychometrical methods for assessing student KSAs in MBPA.

Research question 3: How do we build an MBPA?

We built a framework for designing and developing the MBPA to answer the third research question. The framework is a particularly helpful tool for practitioners when building an MBPA. Using a framework for structured design and development not only makes this an efficient and effective process, it also helps to build a validity argument from the earliest moments of the developmental process.

Research question 4: What is the relationship between scores in an MBPA and scores in a PBA?

Where an MBPA is used to assess student KSAs that are normally assessed in a PBA, there should be a relationship between scores in both types of assessment. However, the relationship may be impaired because it is affected by the measurement error of the PBA. We have demonstrated that there is indeed a moderate correlation between student scores for a group of students who performed in both the MBPA and the PBA. The empirical results provide strong arguments for answering the general research question.

Research question 5: How can we score complex and interactive behavior in an experimental MBPA, and then apply those scores in a psychometric model?

The fifth research question actually went beyond the general research question. Although we were satisfied that the design and development of the MBPA answered the fourth research question, we wanted to innovate further by building a more experimental MBPA in which the student had (some) open space in the virtual environment in which to act. We presented a methodology that can be used to score the complex and interactive behavior that students show in such an assessment. In addition, we have shown how these scores can be used in a psychometric model.

Combining the answers to the research questions of this thesis, we can verify that we were able to develop and use an MBPA for making valid inferences regarding student KSAs. MBPA offers the possibility to measure skills which could not previously be measured via a TBA. In this way, the MBPA proves to be a valuable tool for assessment in vocational education.

7.2 When to Use MBPA?

In the six preceding chapters of this thesis, we discussed a four year PhD research study on MBPA in vocational education. We believe that many questions have been answered, at least to some extent, by the preceding chapters; nevertheless, one question remains. The first chapters focused on the growing influence of technology in testing, the measurement concerns related to PBA, a rationale *why* we should use MBPA and *who* should use it, and a discussion of *what* has already been done. In Chapter 4, a framework is presented that explains *how* to build MBPA. Chapters 5 and 6 present empirical research of an application *where* we have used MBPA. The only question that remains is: *when* should we use MBPA? That is, when is using MBPA as a method of assessment in an assessment program more efficient and/or more effective than using other forms of assessment? More specifically, when do the benefits of MBPA outweigh the costs of MBPA?

The question of when to use MBPA is related to several factors. In testing practice, besides ensuring quality of measurement, it is also important to look at the costs of developing an MBPA, which can be substantial. It also depends on the profession or vocation that is being measured. For example, very expensive simulations are built for airline pilots, simply because everyone agrees that the benefits outweigh the costs. A pilot's errors can result in the loss of human life and human life is in general the most valuable asset in our society. Such expen-

sive simulations are not built for forklift truck drivers, although to do this would be possible and might result in a more valid and reliable assessment of the skills of forklift truck drivers. If a forklift truck driver makes an error, he or she may destroy a pallet of goods (which may be of substantial monetary value), or cause a personal accident (which may even be fatal). Nevertheless, in contrast to an airplane crash, the consequences of an accident with a forklift truck are considered to be limited. In other words, the *value of the damage* that someone can do by making mistakes is one factor that influences the decision whether or not to build an MBPA. In general, if the risk of high value damage is substantial, then we are prepared to invest in more expensive ways of assessment. On the other hand, it is also questionable whether MBPA is really expensive. Of course, initial developmental costs are high compared to developing a multiple-choice test or a scenario for a PBA, but performing real-life scenarios is also expensive. However, once the multimedia structure has been built, it is less expensive to make alternative scenarios in the MBPA. At least, that is what we experienced during our research. In addition, although some costs for maintenance and improvement still exist after development, the MBPA itself remains usable for a long period of time.

Another factor, related to the one above, is the opportunity to measure skills that are difficult to measure in a paper-based test or PBA. For example, *dangerous situations* are often very difficult to measure. MBPA offers a solution to this problem because it is possible to provide an immersive virtual environment that realistically represents a real-world environment. A good example is the *Navy Damage Control* simulation presented in a research report by Iseli, Koenig, Lee, & Wainess (2010) (see also Chapter 3). Of course, there are questions concerning the validity of the assessment when the situation is actually not dangerous at all. A test taker's reaction to a genuinely dangerous situation might be different from the reaction in the MBPA. However, with the research presented in this thesis, we have shown that students' reactions in a virtual environment are related to their reactions in a real environment. In addition, it is not always possible to use a form of assessment other than an MBPA (e.g., a plane crash).

On the other hand, some skills cannot be measured using MBPA. This is often the case when the limitations of (accessible) technology are reached. For example, in the MBPAs presented in this thesis, we simulate a conversation by using text input, whereas it would be better to use a student's own voice. Alt-

hough there is progress in this field (McNamara, Graesser, McCarthy, & Cai, 2014), it is still not possible to fully and realistically simulate a live conversation between the student and a virtual character inside the simulation. Thus, communicative skills are difficult to measure. Another aspect of performance that is still difficult to measure is a student's work attitude or motor skills. Again, in this field there is also technological progress, especially in the medical field (Vankipuram, Kahol, McLaren, & Panchanathan, 2010). However, although technology for the assessment of these types of skills is becoming available, it will take quite some time for it to be available on a broader scale. Finally, students sometimes have to make a so-called masterwork, which is required to reach the rank of master in a guild or other craft organization as a proof of competence. This masterwork can also be part of a final assignment for creative vocational training. Currently, we see no possibilities to integrate this type of assignment into an MBPA. The true value of such masterworks can only be defined by how they are evaluated by connoisseurs and society at large.

In general, and taking in consideration the paragraphs above, MBPA should be used when it possible to expand or strengthen the domain of assessment, or both. With regards to expanding the domain of assessment, we mean that MBPA can reveal certain aspects of student performance that cannot possibly be revealed through paper-based assessments or PBA. With regards to strengthening the domain of assessment, we mean that particular student skills can be assessed in a more valid and reliable way than was possible with traditional assessment methods.

Another way to make a decision about when to use MBPA is to look at assessment taxonomy, for example Bloom's taxonomy. We used Bloom's revised taxonomy (Krathwohl, 2002) to analyze which categories of the taxonomy it is best to use in MBPA. Table 7.1 shows a simplified version of Bloom's revised taxonomy.

Table 7.1
Bloom's Revised Taxonomy

	Cognitive process dimension					
Knowledge dimension	Re-member	Under-stand	Apply	Analyze	Evaluate	Create
Factual						
Conceptual						
Procedural						
Metacogni-tive						

Any item, task or assignment in an assessment can be indexed in one of the taxonomy categories. It is a crossed design in which the knowledge dimension presents a noun and the cognitive dimension presents a verb. For example, in the item "In which year did Napoleon become emperor?" the noun is a fact (the year) and the verb is to remember, which would put this item in the upper left category of the taxonomy. On the other hand, if the assignment is "Create a strategy to build your own enterprise", then the noun is metacognitive (a strategy) and the verb is to create, which would put the assignment at the lower right category. Both the question and the assignment would not be very suitable for MBPA. Although it might be possible to use them in an MBPA, it would not yield the added value that MBPA can provide. A multiple-choice item would suffice for the question and is recommended because of its efficiency. It would best be to use an essay assignment.

So the question remains as to when to use an MBPA. In other words, when can MBPA add something to the measurement perspective? MBPA is most useful for the higher knowledge dimensions (procedural and metacognitive), because the lower knowledge dimensions can generally be very well tested

using multiple-choice tests or essay questions. Within these two dimensions, we would say that the middle two cognitive dimensions (apply and analyze) are then best suited for MBPA. Again, the lower two dimensions can best be tested using a traditional knowledge test, whereas the upper two dimensions are rather advanced cognitive processes that, at least for a large part, cannot be tested through an MBPA. We have therefore colored the categories of the taxonomy gray where we think MBPA is most suitable.

What you can do very well with MBPA, in our opinion, is to have students execute certain procedures, integrate and check procedures, use strategies, make observations, and analyze strategies. This makes sense as these are also the constructs that are usually tested through PBA in vocational education. In fact, you could say that MBPAs are primarily concerned with applying and analyzing certain types of behavior, appraising the outcomes of behavior, and making decisions based on analysis and appraisal. To conclude this section, the answer to the question when to use MBPA depends on a number of factors. It should be used to expand and/or strengthen the domain of assessment, but only where used to measure constructs for which MBPA really has added value.

7.3 Strengths, Practical Implications, and Limitations of the Research Presented in this Thesis

Research on innovative computer-based testing applications is growing, and has shown that the technological developments enable test designers to build interactive virtual assessments. The studies reported in this thesis provide theoretical as well as practical contributions for researchers, practitioners and policy makers.

The first contribution is that all literature on MBPA has been brought together in a single chapter of the thesis. Up to a few years ago, studies in this area of research were still scarce, and research was struggling to keep up with the pace of technological developments. Now, research on the topic is growing fast. For researchers working in this field, this thesis provides a reference work and a point of departure for new research. Providing a solid framework for designing and developing MBPA is a second contribution of the thesis. The framework is especially helpful for practitioners who are currently working (or will be in the future) on building an MBPA or related type of assessment. A third contribution is that the empirical research presented in this thesis is one

of the first attempts in vocational education to assess KSAs that had previously been measured in a PBA. In this way, we hope that future researchers can build on the work presented in the preceding chapters. The fourth, and perhaps most essential, contribution that this thesis provides is its demonstration that an innovative computer-based assessment can indeed be used to measure skills that were in the realm of PBA for a long period.

Of course, there are also some limitations to the research presented here. A primary limitation remains that performance in a virtual environment is not the same as performance in a real-world environment. For example, when we want to observe students' attitudes or ways of communication, we cannot resort to MBPA. Our research also shows that specific constructs within a PBA are still difficult to measure in an MBPA. Nevertheless, we have been able to show that there is a relationship between performance in a virtual and in a physical environment. This relates to the *criterion validity* of the MBPA, more specifically its *predictive validity*. In fact, one might argue that predictive validity is the most important form of validity, because predicting whether or not a student is capable of performing a vocation is the essence of credentialing/certification assessment in vocational education.

Another limitation is that the use of computers and especially the more complex features in MBPA can cause *construct irrelevant variance* (CIV) when students are not sufficiently trained in operating the assessment. In other words, it might be that a students' score is not a clear representation of the construct we want to measure, but instead measures their skills in being able to handle different features in the MBPA (e.g., navigating, zooming, etc.). Although we used practice assignments and questionnaires on computer use and MBPA usability, CIV is difficult to detect. Overall, we have shown that usability and computer experience do not influence scores in MBPA, although this might be true for individual cases. Some students, especially those who do not work with computers on a regular basis, had difficulties with the functionalities of the MBPA. However, explaining how the assessment worked in more detail, improved their ability to handle it. We think this objection to the validity of the assessment could be neutralized with students who have already engaged with computer-based simulations during their education, for example through the use of trial scenarios.

7.4 Future Research

We expect MBPA to evolve into an interactive and immersive virtual environment in which a student can freely wander around and fulfil tasks and assignments, while his or her actions are scored *on the go*. A challenge will be to use this new type of MBPA for summative rather than formative purposes. Future research should therefore focus on investigating the extent to which these assessments can be used in a summative setting. Of course, the research presented in this thesis already contributes to this question.

Another interesting stream of research will be on the psychometric analysis of the data that students produce while performing in an MBPA. Performance in MBPAs, like other SBAs, can produce huge amounts of process data. We have already shown, in Chapter 6, that process data may also provide interesting information regarding student KSAs. Future research could apply (educational) data mining techniques to find meaningful relationships in the process data, which in turn could be used for the overall evaluation of student performance. As an example, future research could try to apply response time models on MBPA performance data to make better estimates of student proficiencies.

7.5 Conclusion

The research presented in this thesis is one of the first endeavors to measure practical constructs in a computer-based environment, especially in vocational education. This thesis has revolved around multiple aspects of MBPA: literature on the topic, design and development, and empirical research. The results from our research address a timely topic, from a both a practical and a theoretical point of view. Of course, in a field that evolves so quickly, the results may be outdated within a decade. We can never be sure about the progression of technology or psychometrics in time, but it is definitely not unthinkable that the use of more static multimedia, photographs or video material, for example, will be a thing of the past within a few years. When we look at the commercial film industry, we also see an increasing use of high fidelity reconstructions of reality.

Based on scientific inquiry, the use of technology in assessment grows layer by layer. Sometimes a layer is thick and sometimes it is thin. New layers can only be built and placed on top because of the layers that have already been constructed. Our research contributes to a new layer of technology in assess-

ment. This research will help researchers and practitioners to make that layer stronger, and future layers more tangible.

References

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

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212−218.

McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: University Press.

Vankipuram, M., Kahol, K., McLaren, A., & Panchanathan, S. (2010). A virtual reality simulator for orthopedic basic skills: A design and validation study. *Journal of biomedical informatics, 43*(5), 661−668.

Summary

In this thesis, a new method of assessment, which we have called Multimedia-based Performance Assessment (MBPA) is presented and tested in the context of Dutch vocational education. An MBPA is an assessment that incorporates multiple types of multimedia and is used to assess the skills that are usually measured through Performance-based Assessment (PBA). The goal of this research project was to investigate whether or not MBPA could be a more efficient and effective way of assessing students' skills than a traditional PBA. In this thesis, we present both qualitative and quantitative evidence to substantiate this claim.

In **Chapter 1**, a general introduction, unfamiliar readers are introduced to several fields. First, a short background on vocational education and training (VET) in the Netherlands is given. One of the most common ways to assess students' skills in VET is to use a PBA. In a PBA, a student is confronted with tasks and assignments that resemble activities that take place in the real work environment. PBAs can therefore be administered during students' work placements, but they can also be presented in simulated work environments. The chapter also includes an introduction to the history of computer-based testing (CBT) and the most recent innovations in CBT, which can be called simulation-based assessment (SBA). SBA is an umbrella term for all innovative types of CBTs, and MBPA is one of these. The MBPA distinguishes itself from other SBAs for two reasons. First, it is strongly focused on assessing the skills that are currently being measured by PBAs. Secondly, whereas the most innovative SBAs already present game-like environments, the MBPA tries to simulate an immersive environment by using multimedia rather than interactive game environments. This is explained in detail in the first chapter, which ends with an outline of the thesis. One could say that the first chapter answers the question: *"Where does MBPA originate?"*

PBA is characterized by being prone to several sources of measurement error, which are presented and discussed in detail in **Chapter 2**. A rationale for using MBPA is also presented in this chapter. Several arguments are given that position MBPA as a realistic alternative to PBA. Although we do not necessarily present MBPA as an alternative to PBA, we state that it has some attractive characteristics that can at least diminish the influence of measurement error

typical of PBA in an assessment program. In addition, a pilot example of an MBPA is presented in this chapter and future research is explained. The second chapter of the thesis answers the question: "*Why should we use MBPA?"*

Some work on SBA has already been done and **Chapter 3** therefore presents a systematic review of the literature on SBA. Solid methodology is used for the systematic review, which lists all scientific articles in which an SBA is presented and the data was psychometrically analyzed. The results of the papers are presented and indexed according to evidence-centered design categorization. The task model variables that were measured with the SBA are specified. This includes details of which task model variables were used, and how the evidence model was constructed, including the measurement model. In addition, the Bayesian network (BN), the most frequently-used psychometric model, according to the review, is explained in depth and an example is presented in the second part of this chapter. The third chapter answers the questions: *"How do we psychometrically analyze SBA and MBPA performance data and what has already been done in this field?"*

In **Chapter 4**, a framework for designing and developing an MBPA is presented. This framework was built on a careful analysis of literature and several rounds of expert consultation. After a prototype framework had been built, multiple interviews with new experts were organized to validate whether they would agree with the prototype framework or could add suggestions for improvement. The five interviews yielded enough information to further update the prototype to a final version, which is presented in the chapter. This chapter answers the question: *"How do we build MBPAs?"*

This framework was then used to build a fully functioning MBPA for a profession in Dutch vocational education. This work, including empirical research on its functioning, is presented in **Chapter 5**. This chapter begins with an in-depth discussion of exactly how we built the MBPA. The MBPA is used to assess the skills of students who want to become *confined space guards*. A confined space guard (CSG) supervises operations that are performed in confined spaces, which are spaces with limited entry, for example a silo or a tank. After the design and development discussion, we then present an empirical study in which a group of students perform in both the MBPA and the PBA. Several instruments are used to collect data and test our hypotheses. The results show that there is a significant relationship between students' scores on the MBPA

and the PBA. Furthermore, MBPA test and item characteristics are reported in this chapter. This chapter answers the questions: *"What is the (psychometric) functioning of an MBPA and how do students' MBPA scores relate to their PBA scores?"*

Because the MBPA presented in Chapter 5 is still rather linear in nature, we decided to develop another, more experimental and interactive MBPA for assessing CSG skills in **Chapter 6**. In this chapter, we looked specifically at how to psychometrically score interactive student behavior in an MBPA. This chapter answers the question: *"How can we score complex, interactive student behavior in an MBPA?"*

Finally, in **Chapter 7**, the epilogue, we make remarks about the research carried out for this thesis. We try to predict future trends in MBPA and in SBA in general. And we try to answer one last question: *"When is it efficient and effective to use MBPA?"*

Samenvatting

In dit proefschrift wordt een nieuwe examenmethode, die wij het multimediale praktijkgerichte computerexamen (MPCE) noemen, in het Nederlandse beroepsonderwijs gepresenteerd en onderzocht. Een MPCE is een examen waarmee aan de hand van verschillende vormen van multimedia, praktijkgerichte studentvaardigheden worden getoetst. Het doel van dit onderzoeksproject was om te onderzoeken of MPCE's een efficiëntere en effectievere examenmethode kunnen zijn dan traditionele praktijkexamens. In de hoofdstukken van dit proefschrift presenteren we zowel kwalitatieve als kwantitatieve argumenten die aantonen in hoeverre we het doel behaald hebben.

In **Hoofdstuk 1**, een algemene introductie, worden verschillende termen en onderzoeksvelden geïntroduceerd voor lezers die onbekend zijn met het onderwerp. Eerst wordt een korte achtergrond gegeven over het Nederlandse (middelbaar) beroepsonderwijs en het praktijkexamen. Het praktijkexamen is een van de meeste gebruikte examenmethodes in de examenmix van veel kwalificaties in het beroepsonderwijs. In een praktijkexamen worden studenten geconfronteerd met taken en opdrachten die overeenkomen met activiteiten zoals deze in de beroepspraktijk plaatsvinden. Daarom kunnen praktijkexamens ook vaak tijdens de beroepspraktijkvorming (de stage) afgenomen worden. Een andere methode om praktijkexamens af te nemen is in een gesimuleerde beroepsomgeving, bijvoorbeeld op het regionaal opleidingscentrum (ROC). Daarnaast wordt er in dit hoofdstuk een ander, voor dit proefschrift belangrijk, onderwerp geïntroduceerd: computer-based testing (CBT). De historie van CBT wordt geschetst en de chronologie wordt gevolgd tot de meest innovatieve vorm van CBT: het gebruik van computersimulaties voor examinering. Simulation-based assessment (SBA) is feitelijk gezien een verzamelterm voor allerlei vormen van innovatieve computertoetsing en het MPCE is een van deze vormen. Het MPCE onderscheidt zich van andere vormen van SBA op twee manieren. Eén, MPCE's richten zich op het meten van praktijkgerichte vaardigheden. Twee, MPCE's zijn opgebouwd uit verschillende vormen van multimedia maar bieden nog geen volledige game omgeving waarin een student vrij kan rondlopen met een virtueel karakter. Deze laatste vorm wordt vaak *serious gaming* genoemd. Het hoofdstuk eindigt met de verdere indeling van het proef-

schrift. Men kan zeggen dat dit hoofdstuk de volgende vraag beantwoordt: *"Waar ligt de oorsprong van het MPCE?"*.

Praktijkexamens zijn bevattelijk voor verschillende bronnen van meetfout, die allemaal worden gepresenteerd en behandeld in **Hoofdstuk 2**. Verder zal in dit hoofdstuk een uitgebreide argumentatie worden opgevoerd voor het gebruik van het MPCE in het beroepsonderwijs. Verschillende argumenten worden gegeven waarmee wordt uitgelegd dat het MPCE een realistisch alternatief is voor het veelvoorkomende praktijkexamen. Hoewel we het MPCE niet per se als een vervanging van het praktijkexamen presenteren, laten we zien dat het verschillende interessante eigenschappen kent waarmee het ten minste de invloed van de door praktijkexamens geproduceerde meetfouten kan verminderen. Ook zal een eerste pilot voorbeeld van een MPCE getoond worden en wordt uitgelegd hoe toekomstig onderzoek kan bijdragen aan het beantwoorden van verschillende onderzoeksvragen. Men kan zeggen dat dit hoofdstuk de volgende vraag beantwoordt: *"Waarom zouden we gebruik maken van het MPCE?"*.

Er is natuurlijk al onderzoek gedaan naar SBA en daarom wordt in **Hoofdstuk 3** een systematische review van de literatuur gepresenteerd. We hebben een solide methodologie gebruikt om alle wetenschappelijke literatuur naar dit onderwerp in kaart te brengen. Uitgangspunt was dat de schrijvers van de artikelen tenminste ook de data die de SBA produceert (psychometrisch) hebben geanalyseerd. De resultaten van alle artikelen die we gevonden hebben, hebben we geïndexeerd naar de categorieën van het *evidence-centered design* framework. Dat betekent dat per artikel gekeken is naar de studentmodelvariabelen (wat wordt er gemeten?), de taakmodelvariabelen (hoe wordt er gemeten?) en het bewijsmodel (welk statistisch meetmodel wordt er gebruikt?). De resultaten van deze systematische review lieten duidelijk een trend zien voor het gebruik van het Bayesiaanse netwerk voor het analyseren van data uit SBA. Daarom geven we in het tweede gedeelte van dit hoofdstuk een introductie over Bayesiaanse netwerken en presenteren we een beknopt voorbeeld van een Bayesiaans netwerk voor het MPCE die verder wordt behandeld in Hoofdstuk 5. Men kan zeggen dat dit hoofdstuk de volgende vragen beantwoordt: *"Hoe analyseren we (op een psychometrische manier) de data die wordt geproduceerd door SBA en dus MBPA en wat is er allemaal al aan onderzoek gedaan op dit gebied?*".

In **Hoofdstuk 4**, wordt een raamwerk voor het ontwerpen en ontwikkelen van het MPCE gepresenteerd. Het raamwerk is gebouwd door middel van een

grondige analyse van de literatuur en verschillende rondes van expert consultatie. Dit resulteerde ten eerste in een prototype van het raamwerk. Vervolgens hebben we via 5 interviews met nieuwe experts gecontroleerd in hoeverre zij overeenstemmen met het prototype. Deze strategie kan gezien worden als een validering van het prototype. De 5 interviews leverden nieuwe informatie op waarmee het prototype op verschillende manieren is aangepast en verbeterd, waarmee een definitief ontwerp- en ontwikkelingsraamwerk ontstond, die uiteraard wordt gepresenteerd in het hoofdstuk. Men kan zeggen dat dit hoofdstuk de volgende vraag beantwoordt: *"Hoe bouwen we een MPCE?"*.

Het raamwerk dat gepresenteerd wordt in Hoofdstuk 4 is vervolgens gebruikt om een volledig functionerende MPCE te ontwikkelen voor een kwalificatie in het Nederlandse beroepsonderwijs. Dit werk, tezamen met een empirische studie naar het (psychometrisch) functioneren van het MPCE, wordt gepresenteerd in **Hoofdstuk 5**. Dit hoofdstuk begint met een uitgebreide discussie over de ontwerp- en ontwikkelingsfase van het MPCE. Het MPCE wordt gebruikt om de praktijkgerichte vaardigheden te meten die behoren tot de kwalificatie tot buitenwacht uit de opleidingengids van de Stichting Samenwerken voor Veiligheid. Een buitenwacht houdt toezicht op werkzaamheden die plaatsvinden in besloten ruimtes. Een besloten ruimte is een ruimte met beperkte toegang en weinig ventilatie. Hierbij kun je bijvoorbeeld denken aan een tank of een silo op een petrochemisch bedrijfsterrein. Nadat we uitvoerig hebben besproken hoe het MPCE is ontwikkeld, presenteren we een empirische studie waarin een aselecte groep van kandidaten naast hun praktijkexamen deelneemt aan het MPCE. Verschillende meetinstrumenten worden gebruikt om data te verzamelen en de hypotheses te testen. De resultaten tonen een significante relatie tussen de scores op het MPCE en de scores op het praktijkexamen. Daarnaast presenteren we de toets- en itemkarakteristieken van het MPCE. Men kan zeggen dat dit hoofdstuk de volgende vragen beantwoordt: *" Hoe functioneert het MPCE (psychometrisch) en wat is de relatie tussen scores op het MPCE en scores op het praktijkexamen?*".

Omdat het MPCE gepresenteerd in Hoofdstuk 5 nog relatief lineair is, hebben we besloten om nog een tweede MPCE te ontwikkelen waarmee de praktijkgerichte vaardigheden van buitenwachten kunnen worden gemeten. Het doel was om een experimentele en interactieve versie te ontwikkelen, waarin kandidaten een vrijere ruimte krijgen om te handelen. Het resultaat wordt ge-

presenteerd in **Hoofdstuk 6**. In dit hoofdstuk wordt specifiek gekeken op welke manier interactieve handelingen in het MPCE (psychometrisch) gescoord kunnen worden. De resultaten laten zien dat aan de hand van expertbeoordelingen, handelingen in een complexe en interactieve MPCE gescoord kunnen worden en vervolgens toegepast kunnen worden in een Bayesiaans netwerk. Men kan zeggen dat dit hoofdstuk de volgende vraag beantwoordt: *"Hoe kunnen we complex, interactief gedrag van kandidaten in een MPCE scoren en toepassen in een psychometrisch model?"*.

Tot slot, in **Hoofdstuk 7**, de epiloog, worden de resultaten van het onderzoek in dit proefschrift geduid en in het juiste perspectief geplaatst. Ook worden toekomstige trends in onderzoek naar het MPCE en SBA in het algemeen voorspeld. Daarnaast beantwoorden we nog een laatste vraag: *"Wanneer is het efficiënt en effectief om het MPCE te gebruiken in het beroepsonderwijs?"*.

Dankwoord

Ik wil mijn beide promotors, Theo Eggen en Bernard Veldkamp, bedanken voor hun waardevolle inzichten, commentaren en de nodige gezelligheid. Gedurende de afgelopen vier jaar heb ik met beiden elke twee weken overleg gehad en ik kan mij niet anders herinneren dan dat dit mij hielp om verder te komen met mijn onderzoek. Een promotieonderzoek kenmerkt zich volgens mij door het bewandelen van vele paden, waarvan sommigen je het bos uit leiden naar succes en anderen je alleen maar dieper het bos insturen. Theo en Bernard hebben er altijd voor gezorgd dat ik weer op het juiste pad terecht kwam.

Daarnaast wil ik eX:plain bedanken voor het initiëren, faciliteren en financieren van mijn promotieonderzoek en wil ik mijn collega's bij eX:plain bedanken voor hun ondersteuning tijdens mijn promotieonderzoek. Mijn leidinggevenden tijdens dit project, Hester Brenninkmeijer en Frank Hubert, hebben alles in het werk gesteld om mij dit proefschrift te laten voltooien. Een speciaal dankbericht gaat uit naar Peter van Dijk, die mij door vele discussies en overleggen, soms over het promotieonderzoek soms over van alles en nog wat, altijd het vertrouwen heeft gegeven dat ik dit project kon voltooien.

Ik wil ook graag alle collega's van het RCEC en Cito bedanken voor de inspirerende onderzoekomgeving die jullie hebben geboden. Elke woensdag was ik op Cito en ik ben ervan overtuigd dat dat heeft bijgedragen aan de uiteindelijke vorm en inhoud van mijn proefschrift.

Tijdens de dataverzamelingsfase van mijn onderzoek heb ik gebruik gemaakt van de diensten van VCA Infra en Examenbank, SAIO, Ascend, TCC, Falck, PBNA, de SSVV en de waarborgcommissie buitenwacht. Ik wil al deze organisaties bedanken voor de hulp bij de praktische uitvoering van mijn onderzoek.

Tot slot wil ik graag mijn familie en vrienden (waaronder mijn twee paranimfen) bedanken voor hun ondersteuning en voor de afleiding die ik soms nodig had. Ook zij hebben er altijd vertrouwen in gehad dat ik deze prestatie kon behalen. Een speciaal dankbericht gaat uit naar mijn vriendin, Iris, die mij ook altijd heeft helpen doorzetten om het proefschrift te voltooien.

Curriculum Vitae

Sebastiaan de Klerk was born on July 17, 1986 in Haarlem, the Netherlands. He completed his high school (VWO) at College Hageveld in Heemstede. After finishing his VWO, he studied psychology at the University of Amsterdam. During his bachelor he started to focus on work and organizational psychology, which he continued doing in the research master. He completed his research master in December 2010. Besides majoring in work and organizational psychology, he also followed a minor in social psychology. His internship in the research master was concerned with doing research on the influence of high-power group collaboration on group members' individual negotiation style and their well-being.

In December 2011, he started with a four year PhD project at eX:plain and the Research Center for Examinations and Certification, under supervision of prof. dr. ir. Theo J.H.M. Eggen and prof. dr. ir. Bernard P. Veldkamp. The studies reported in the chapters of this thesis are the result of this PhD project. The PhD project focused on the efficiency and effectiveness of the use of multimedia-based performance assessment as a new way of performance assessment in Dutch vocational education and training. Furthermore, Sebastiaan has worked as a project advisor and test developer at eX:plain for the last four years.

Research Valorisation

1. Publications in Scientific Journals

1.1 Published

De Klerk, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2014). A blending of computer-based assessment and performance-based assessment: Multimedia-based Performance Assessment (MBPA). The introduction of a new method of assessment in Dutch Vocational Education and Training (VET). *CADMO, 22*(1), 39-56.

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

1.2 Submitted

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.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (submitted). *The design, development, and evaluation of a summative multimedia-based performance assessment for credentialing confined space guards*. Manuscript submitted for publication.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (submitted). *A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian network*. Manuscript submitted for publication.

2. Professional Publications

De Klerk, S., Van Dijk, P., & Van den Berg, L. (2015). Voordelen en uitdagingen voor toetsing in computersimulaties: Van ontwikkeling tot analyse. *Examens, 12*(2), 11-17.

3. Book Chapters

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

4. Conference Contributions, Presentations, and Lectures

De Klerk, S. (2012, March). *The introduction of multimedia-based performance assessment in Dutch vocational education.* Presentation at the Foundation Cooperation for Safety quality assurance committee, Breda, the Netherlands.

De Klerk, S. (2012, July). *The introduction of multimedia-based performance assessment in Dutch vocational education.* Presentation at eX:plain working visit at Employment Technologies Corporation, Winter Park, FL.

De Klerk, S., Boonman, K., & Wagemakers, W. (2012, October). *The Foundation Cooperation for Safety opts for computer-based assessment.* Interview given for *VCA Actueel*, Breda, the Netherlands.

De Klerk, S., & Boonman, K. (2012, October). *Innovative test items in computer-based testing.* Presentation at conference of the Safety, Health and Environment Checklist Contractors, Rotterdam, the Netherlands.

De Klerk, S., & Van den Berg, L. (2013, April). *The future of performance-based assessment.* Presentation at the eX:plain performance-based assessment rater's day, Amersfoort, the Netherlands.

De Klerk, S., & Kloppenburg, M. (2013, May). *Two day introductory course on assessment and psychometrics*. Course given at eX:plain, Amersfoort, the Netherlands.

De Klerk, S. (2013, May). *The use of multimedia in assessment*. Presentation at the Dutch council for trainers and educators, Utrecht, the Netherlands.

De Klerk, S., Veldkamp, B.P., & Van Dijk, P. (2013, May and June). *Qualitative assessment during work placement: Pitfalls and opportunities*. Presentation at the conference of the Dutch Vocational Education Service Institution, Utrecht, the Netherlands.

De Klerk, S. Veldkamp, B.P., & Eggen, T.J.H.M. (2013, June). *A framework for designing and developing multimedia-based performance assessment.* Poster presented at the conference of the Dutch Interuniversity Graduate School of Psychometrics and Sociometrics, Groningen, the Netherlands.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2013, October). *A framework for designing and developing multimedia-based performance assessment*. Paper presented at the conference of the International Association for Educational Assessment, Tel Aviv, Israel.

De Klerk, S. (2013, November). *Multimedia-based performance assessment*. Experimental workshop given at the conference of the Dutch Association for Assessment, Nunspeet, the Netherlands.

De Klerk, S., & Van Dijk, P. (2014, June). *Challenges and solutions for performance-based assessment: The use of a multimedia-based performance assessment for credentialing confined space guards*. Presentation at the conference of OnTrac, Ede, the Netherlands.

De Klerk, S., & Veldkamp, B.P. (2014, October). *Assessment: Cycle of testing and innovations in testing* (Lecture, Educational assessment, Pre-master Educational Science and Technology), University of Twente, Enschede, the Netherlands.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2014, November). *Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example*. Paper presented at the conference of the European Association for Educational Assessment, Tallinn, Estonia.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2014, November). *Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example*. Paper presented at the conference of the Research Center for Examinations and Certification, Enschede, the Netherlands.

De Klerk, S. (2014, November). *A summative multimedia-based performance assessment for credentialing confined space guards*. Presentation at a working visit of the Dutch Association for Assessment, Rotterdam, the Netherlands.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2015, June). *The psychometric evaluation of a summative multimedia-based performance assessment for credentialing confined space guards*. Paper presented at the Computer Assisted Assessment conference. Zeist, the Netherlands.

De Klerk, S., Veldkamp, B.P., & Eggen, T.J.H.M. (2015, November). *The design, development, and validation of a multimedia-based performance assessment for credentialing confined space guards*. Paper presented at the conference of the European Association for Educational Assessment, Glasgow, Scotland.