The Psychometric Evaluation of a Summative Multimedia-Based Performance Assessment

Sebastiaan De Klerk[1,2(✉)], Bernard P. Veldkamp[2,3], and Theo Eggen[2,4]

[1] eX:plain, Amersfoort, The Netherlands
s.dklerk@explain.nl
[2] Research Center for Examinations and Certification, Enschede, The Netherlands
b.p.veldkamp@utwente.nl, theo.eggen@cito.nl
[3] University of Twente, Enschede, The Netherlands
[4] Cito, Arnhem, The Netherlands

Abstract. In this article, a case study on the design, development, and evaluation of a multimedia-based performance assessment (MBPA) for measuring confined space guards' skills is presented. A confined space guard (CSG) supervises operations that are carried out in a confined space (e.g. a tank or silo). Currently, individuals who want to become a certified CSG in The Netherlands have to participate in a one day training program and have to pass both a knowledge-based MC test and a practice-based performance-based assessment (PBA). Our goal is to measure the skills that are currently being assessed through the PBA, with the MBPA. We first discuss the design and development of the MBPA. Secondly, we present an empirical study which was used for assessing the quality of our measurement instrument. A representative sample of 55 CSG students, who had just completed the one day training program, has subsequently performed in the MC test, and then, depending on the condition they were assigned, the PBA or the MBPA. We report the psychometric properties of the MBPA. Furthermore, using correlations and regression analysis, we make an empirical comparison between students' scores on the PBA and the MBPA. The results show that students' scores on the PBA and the MBPA are significantly correlated and that students' MBPA score is a good predictor for their score on the PBA. In the discussion, we provide implications and directions for future research and practice into the field of MBPA.

Keywords: Performance-based assessment · Multimedia-based performance assessment · Psychometric evaluation · Design and development

1 Introduction

The growing capabilities and availability of technology enable a whole new generation of technology driven assessments, far more elaborated than computer-based transformations of formerly item-based paper-and-pencil tests [4, 10]. The new generation of technology-based assessments both expand and deepen the domain of assessment [9].

E. Ras and D. Joosten-ten Brinke (Eds.): CAA 2015, CCIS 571, pp. 1–11, 2015.
DOI: 10.1007/978-3-319-27704-2_1

Technology makes more flexible and context driven presentations of tasks and environments in CBA possible, which can lead to a broader and better understanding of what students have learned [4].

In this article, we discuss the design, development, and evaluation of a technology-based assessment that 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 being measured by a PBA. The purpose of the MBPA we discuss here is to measure the skills of confined space guards (CSG) after they have performed in vocational training. The CSG skills consist of 19 actions that a student has to take during the performance-based assessment (e.g., test the walkie-talkie, check the work permit, assess the wind direction, and register the number of people going in and out of the confined space).

Although PBA has been discussed and supported as a valuable tool for formative and diagnostic assessment of students [8, 11], the research is less supportive in cases where PBA was used as a summative assessment. This is foremost because PBAs are found to be prone to measurement error resulting from several sources; task, occasion and rater sampling variability [5, 6, 12]. Above that, task sampling and occasion sampling are confounded, which means that their combined effect strongly raises measurement error [13]. These findings indicate that students' scores resulting from performance in a PBA do not solely represent students' proficiency in a particular skill, but are influenced by the specific task they were assigned, the occasion of the assessment, and the raters judging their performance. Therefore, the purpose of the current study is to design, develop, and evaluate a multimedia-based equivalent of the PBA, for credentialing confined space guards in Dutch vocational training.

The first part of the paper focuses on design and development and in the second part of the paper an empirical study is presented that focuses on the psychometric functioning of the MBPA, and especially the empirical relationship between the MBPA and the PBA. We compared test scores resulting from the MBPA with students' test scores on the PBA, a paper-and-pencil (P&P) knowledge-based MC test and student ratings on questionnaires about computer experience and the usability of the MBPA. In the experiment, a random, yet representative, sample of students either first performs the PBA and then the MBPA, or vice versa. The central question of our study is: Is it possible to develop a multimedia-based performance assessment that produces valid and reliable estimates of the proficiency of confined space guards? We have the following hypotheses:

Hypothesis 1: The scores of students on the PBA will be positively correlated with the scores of students on the MBPA.
Hypothesis 2: The scores on the MBPA will not be correlated with students' background characteristics (i.e. age, education and ethnicity).
Hypothesis 3: The scores on the MBPA will not be correlated with students' answers on a computer experience questionnaire.
Hypothesis 4: The scores on the MBPA will be positively correlated with students' answers on a usability questionnaire.

Hypothesis 5: The group of students who do not pass the PBA will score significantly lower on the MBPA than the group of students who pass the PBA.
Hypothesis 6: The group of students who first do the PBA will score significantly higher on the MBPA than the group of students who first do the MBPA.
Hypothesis 7: The group of students who first do the MBPA will score significantly higher on the PBA than the group of students who first do the PBA.

2 Design and Development

The start of building an MBPA is to determine the purpose of the assessment. As said, the MBPA was built to measure the skills of CSG's as defined by subject matter experts in the "final attainment objectives" so that it can be used as an assessment for certification of CSG's.

The design phase was started by determining the constructs and attributes that we wanted to measure and analyzing them for translation into the MBPA's tasks. This was done in collaboration with subject matter experts (SMEs) through multiple rounds of consultation. Of course, a lot about the tasks of CSG's was already known through the instruction material and final attainment objectives of the performance-based assessment. Furthermore, the first author took part in a one day course and performed the PBA to become a certified CSG. We used this material and knowledge to further work out the constructs and attributes for the MBPA.

Based on this knowledge, the tasks in the assessment could be designed and developed in collaboration with the SMEs. We first build what we have called an *assessment skeleton,* in which the general flow of the assessment was laid out, including the multimedia and the tasks. This was done on a relatively abstract level but it ensured that all constructs, final attainment objectives, and primary observables are incorporated in the tasks. In validity terms: the demands for content validity were met through the use of the assessment skeletons. Because the assessment skeleton is still a rather coarse-grained representation of the assessment it is not sufficient for actually building the assessment. Therefore, we further elaborated the assessment skeletons into *assessment templates*. In the assessment templates we showed – screen by screen – what was presented during the course of the assessment. The assessment templates enabled us to collect the 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. In addition, the templates served as a primary input for the designer to design the buttons needed in the assessment.

We hired a professional ICT system designer who was very experienced in designing intuitive, usable and efficient interfaces for interactive websites. Furthermore, the templates in combination with the buttons provided the necessary materials for the programmer to build the structure of the assessment on the online assessment platform. The next step was to test the assessment; first on its technical functioning and then on its psychometric functioning in an empirical study. The assessment is administered via the internet and through multiple test rounds we were able to solve the technical bugs, thereby ensuring that the assessment was technically functioning. We will now present our experiment.

3 Method

3.1 Participants

The participants in the pilot study were 55 confined space guard students (1 female, 54 male, mean age: 40.4 years ($\sigma = 11.5$), age range: 19–64 years). They were requested to do the MBPA after they had completed their training.

3.2 Materials

Multiple-Choice knowledge-Based test. Immediately after the training, the students did a knowledge-based P&P test, consisting of 21 MC questions with 3 alternatives.

Performance-Based Assessment. In the PBA, students perform a CSG's job tasks in a reconstructed, yet realistic situation. Figure 1 gives an impression of a PBA for measuring CSG skills.

The rater uses a rubric consisting of 19 criteria to evaluate the student's performance. All 19 criteria can be judged as *insufficient* or *sufficient* by the rater. From the 19 criteria (e.g. "tests the walkie-talkie"), 9 are considered to be a knock-out criterion (e.g. "recognizes and reacts to an emergency situation") which means that if a student's performance

Fig. 1. Performance-based assessment

on one of these criteria is insufficient he or she does not pass the PBA. Besides the 19 criteria rubric that focuses on 19 individual actions that a student can take during the PBA, a second rubric was used for assessing the communicative and behavioral skills of the student. For the second rubric, the rater scores students' communicative skills, proactivity, 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 strong"). Hence, students could get a minimum of 0 points and a maximum of 12 points on this rubric and a minimum of 0 points and a maximum of 19 points on the original rubric and both rubrics were filled out by the rater.

Multimedia-Based Performance Assessment. Clearly, another primary instrument in the study was the multimedia-based performance assessment. The scenario that students went through was the cleansing of a tank on a petrochemical plant by two workers, which was built in the online environment using multimedia. Students started in an office situation where the contractor handed the CSG and one of the workers the work permit. In this setting, students had to ask for explanation of the work permit by the contractor, check the work permit for blanks or errors, ask for a walkie-talkie and test the walkie-talkie. Then the setting changed to the confined space itself. In this setting, students were required to determine the right escape route in case of an emergency, students had to ensure that the environment was safe for the workers to work in and that there were no irregularities between the work permit and the actual situation at the confined space. In a next phase, students had to supervise two workers who were cleaning the interior of the confined space. Finally, students had to act upon a plant alarm. Students were required to watch the multimedia elements and to answer several types of questions (e.g. multiple choice, rank order, fill in the blank, etc.) during the administration of the MBPA. We also included so-called

Fig. 2. Multimedia-based performance assessment screenshot.

intervention tasks. The intervention tasks required students to intervene in two videos of workers performing cleansing tasks in a tank whenever their actions were incorrect. Students could intervene by clicking on a big and red "stop" button that was presented right beside the video screen. Students were told that they only had three possibilities to click on the stop button. That is, if they clicked the stop button when there were no faulty actions of the workers, then they had one less chance to press the button. The MBPA consisted of a total of 35 tasks. Figure 2 gives an impression of the MBPA.

Questionnaire. After students had performed in the MBPA they were requested to fill out a questionnaire comprised of items (N = 15) addressing their background characteristics (e.g. "What is your highest level op completed education?"), computer use (e.g. "On a scale ranging from 1 (never) to 5 (every day) - How often do you play videogames on a computer?"), and MBPA interface (e.g. "On a scale ranging from 1 (strongly disagree) to 5 (strongly agree) - I was comfortable with the interface of the MBPA"). The questionnaire was based on a translated version of the System Usability Scale [1] and a questionnaire on the use of Internet and the computer at home, developed by Cito [3].

3.3 Procedure

Students participated in their training and completed the P&P test immediately afterwards. Then, depending on the condition they were randomly assigned, students either first performed the PBA and then the MBPA (N = 27) or reversely (N = 28). The students were not allowed to confer between both administrations, so that it was impossible that they exchanged 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. There was no time limit imposed on students; neither for the individual tasks nor for the whole assessment. After students finished the assessment they had to fill out the questionnaire that was upside down on their table.

4 Results

4.1 MBPA Performance

In this section, we will discuss the analysis of the sample data (N = 55). As mentioned above, the assessment is composed of 35 items. In total, students could get one point for each correct answer. The mean score on the test was 22.5 ($\sigma = 3.44$), 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, and 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 finish the assessment was 29 min ($\sigma = 8$). The minimum amount of time spent on the assessment was 19 min, the longest was 58 min. The high standard deviation and the wide bandwidth between minimum and maximum indicate that there is a lot of variance between students' time spent on the assessment.

The reliability of the MBPA is high (GLB = 0.94). We have looked at the best indicator of the reliability, the Greatest Lower Bound [16]. The GLB is the best indicator because the bias of the estimate is rather small [15], and compared to Cronbach's alpha, for example, the GLB is closer to the true reliability of the test [14]. The distribution of the test scores is not skewed (0.014), but many scores are distributed around the mean, thereby increasing kurtosis (0.488). Of course, the number of observations is limited, making it difficult to interpret these indices.

4.2 Hypotheses Testing

Our first hypothesis states that students' PBA score will be positively correlated with their MBPA score. Spearman's rho is used as a correlation index because the measures do not meet the assumptions of normality and linearity, while there is more of a monotonic relationship between the variables. For example, on the 19-point rubric, most students score 17 to 19 of the criteria as correct. The correlations are listed in Table 1.

Table 1. Correlations, means and standard deviations of measures (1000 sample bootstrapping performed)

Measure	1	2	3	4	5	6	7
1. MBPA							
2. PBA (19)	0.39†**						
3. PBA (12)	0.38†**	0.68***					
4. PBA (total)	0.43†**	0.84***	0.96***				
5. MC Test	0.30*	0.2	0.21	0.23			
6. MBPA (time)	0.01	−0.13	−0.2	−0.22	−0.05		
7. Q-Computer exp.	0.09	0.12	0.15	0.16	−0.01	0.1	
8. Q-MBPA usability	0.18†	0.15	0.09	0.16	−0.06	−0.18	0.42**

Note. *p < 0.05, **p < 0.01, ***p < 0.001, † (one-tailed)

The correlation between the MBPA and the rubrics used in the performance assessment is 0.38 ($p < 0.01$) and 0.39 ($p < 0.01$), respectively for the 19-point rubric and the 12-point rubric, which are both significant and thereby support our first hypothesis. 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 ($r_s = 0.43$ ($p < 0.01$)). Of course, there is also a strong significant correlation between both rubrics used in the assessment ($r_s = 0.68$, $p < 0.001$). Furthermore, we also performed a linear regression analysis to see to what extent the performance in the MBPA can predict performance in the PBA. To correct for the negative skew of the distribution of the 19-point rubric, we performed a log transformation [7]. For correct analyses, we did this for both the 12-point rubric, the 19-point rubric, and the total rubric score. The regression analysis for the 19-point rubric showed a significant effect ($F(1,53) = 4.365$, $p < 0.05$), which indicates that the MBPA score can account for 7.6 %

of the variation in the PBA score. We performed the same analysis for the 12-point rubric, which was also significant ($F(1,46) = 5.544$, $p < 0.05$), with an explained variance of 10.1 %. Finally, we also performed a regression analysis for the total rubric score, which was also significant ($F(1,46) = 5.905$, $p < 0.05$), with an explained variance of 11.4 %. The total rubric score is the best predictor for performance in the MBPA. Unfortunately, the rater forgot to fill out the 12-point rubric on one assessment occasion, which explains the declined number of students in the second analysis. Furthermore, when we look at misclassifications at the 60 % cutoff percentage (as established by experts) for the MBPA, we see that 7 out of 8 students that failed their PBA also fail the MBPA. This indicates that the PBA score is a good predictor for the MBPA score.

We expected to observe no correlation between students' background characteristics and their score on the MBPA (H_2). The background characteristics are age, education, and ethnicity. Age was not correlated with assessment score ($r_s = 0.00$, $p > 0.05$). We calculated the biserial correlation coefficient for education. The biserial correlation coefficient is used when one variable is a continuous dichotomy [7]. First, we made two groups of students (low education vs. high education). The low education group consisted of students who have had education up to high school or lower vocational education ($N = 26$, $M_{MBPA} = 21.83$) and the high education group consisted of students who have had education from the middle level vocational education and upwards ($N = 27$, $M_{MBPA} = 23.08$). We calculated the point-biserial correlation (which is for true dichotomies [7]), and then transformed it into the biserial correlation. Although education and students' MBPA score were positively correlated, this effect was not significant ($r_b = 0.19$, $p > 0.05$). For ethnicity, we were especially interested in two groups: students with a Dutch ethnicity ($N = 40$, $M_{MBPA} = 22.8$) and students with another ethnicity ($N = 15$, $M_{MBPA} = 22.78$). Now, we calculated the point-biserial correlation between ethnicity (0 = Dutch, 1 = other) and the students' MBPA score. Again, we did not find a significant correlation ($r_{pb} = -0.01$, $p > 0.05$). These findings support our second hypothesis; there were no significant correlations between students' background variables and their score on the MBPA. Also, there was no significant correlation between the time spent on the MBPA and the score obtained ($r = 0.07$, $p > 0.05$).

We also found support for our third hypothesis, because there is no significant positive correlation between the students' MBPA score and their computer experience questionnaire ($r_s = 0.09$, $p > 0.05$). We could not find support for the fourth hypothesis, because there is no significant correlation between the MBPA score and usability questionnaire ($r_s = 0.14$, $p > 0.05$).

Our fifth hypothesis reflected our expectation that students who had failed their PBA would score significantly lower on the MBPA than students who had passed their PBA. Unfortunately, the group of students is 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 ($\sigma = 0.46$) and group of students who failed the PBA had a mean score of 20.1 ($\sigma = 1.1$). We used an independent samples t-test to check whether the groups differed significantly, which was the case ($t(53) = -2.563$, $p < 0.001$). We then performed a logistic regression analysis to check to what extent the MBPA score can predict whether a student will pass or fail in their PBA. The MBPA

score is treated as a continuous predictor in the logistic regression analysis and the dependent variable (success in PBA) is a dichotomous outcome variable (0 = failed, 1 = passed). The analysis demonstrated that the MBPA score is making a significant contribution to the prediction of students failing or passing their PBA ($\chi^2(1, 55) = 5.09$, $p < 0.05$). Furthermore, 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. The results of the logistic regression analysis are presented in Table 2.

Table 2. Logistic regression analysis of passing performance-based assessment

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

Hypothesis six and seven state that students' condition would positively influence the score on the second assessment they performed. However, students who first did the MBPA did not score higher on the PBA than students who started with the PBA [$F(1,53) = 0.96$, $p > 0.05$], and vice versa for the score on the MBPA [$F(1,53) = 0.05$, $p > 0.05$]. This indicates that there is no learning effect between both assessments.

5 Discussion and Conclusion

New forms of technology driven assessments are increasingly becoming part of the modern assessment culture. The aim of this study was to empirically investigate the design, development and evaluation of a multimedia-based performance assessment for credentialing confined space guards in Dutch vocational education. This study is one of the first endeavors in empirically determining the (psychometric) quality of an innovative computer-based assessment that aims to assess constructs normally associated with performance-based assessments.

The reliability of the MBPA is good; the GLB [16] is the best estimate of the reliability and gives the greatest lower bound of the reliability. That means that the reliability of the test is at least as high as the GLB indicates. In our case, the GLB is 0.94.

Students' scores on the PBA (rubrics independently 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 students' performance on the PBA do not show much variance in sum score. We had foreseen this problem already for the 19-point rubric and therefore developed the 12-point rubric; to induce more variation in students' PBA scores. Indeed, it does produce slightly more variance in students' scores, yet it might be too less to really make a difference. It is statistically difficult to establish strong relationships between two variables when one of the variables almost has no variance.

Thereby, we might have found the reason why there isn't a stronger relationship between PBA and MBPA. As discussed in the introduction, performance-based assessments generally suffer from measurement error. This might also be the case for the PBA in our study. In a future study, generalizability theory could be used to determine the psychometric quality of the PBA. Future research in this area should also try to find criteria that are out of the assessment domain. An external criterion could for example be students' future job appraisals, made by their managers. Also, a future study on the subject could include a strong analysis on the quality of the PBA, for example through generalizability theory [2].

Of course, there are some limitations to our study. First, the sample size is rather small. It was difficult to get a substantial number of students to participate in the study, because many assessment locations do not have internet or computers and the locations itself are spread all over The Netherlands. Also, the assessment itself takes place, on average, 15 times per year per location. Sometimes, a group consists of less than five students, which indicates that it can be quite difficult to get a sufficient number of students to participate. On the other hand, because there are not many students per year, we can say that we have included a substantial amount in our study. Furthermore, if we look at background, the sample does not systematically differ from the population.

As already mentioned, another limitation is the quality of the performance-based assessment. Although the PBA is professionally organized, only one rater is being used, who is also playing a part in the assessment (the operator). The 19-point rubric, used for rating a students' performance, shows little to no variance at all, which makes it difficult to draw firm conclusions regarding the MBPA – PBA comparison. Furthermore, another limitation is that this is a first version of the MBPA. If we look at the test and item characteristics presented, then there is room enough for qualitative improvement.

To conclude, with this study we make strong theoretical and practical contributions to advanced technology-based assessment. To our knowledge, we are the first to make an empirical comparison between a computer-based assessment and a practical or manual performance-based assessment in vocational training.

References

1. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Int. J. Hum.-Comput. Interact. **24**, 574–594 (2008)
2. Brennan, R.L.: Generalizability Theory. Springer, New York (2001)
3. Cito: The use of internet and the computer at home questionnaire. Dutch version (2014). http://toetswijzer.kennisnet.nl/html/internetvaardigheid/vragenlijst.pdf
4. Clarke-Midura, J., Dede, C.: Assessment, technology, and change. J. Res. Technol. Educ. **42**(3), 309–328 (2010)
5. Cronbach, L.J., Linn, R.L., Brennan, R.L., Haertel, E.H.: Generalizability analysis for performance assessments of student achievement or school effectiveness. Educ. Psychol. Meas. **57**(3), 373–399 (1997)
6. Dekker, J., Sanders, P.F.: Kwaliteit van beoordeling in de praktijk [Quality of rating during work placement]. Ede: Kenniscentrum handel (2008)
7. Field, A.: Discovering Statistics Using SPSS, 3rd edn. SAGE Publications Inc, Thousand Oaks (2009)

8. Gulikers, J.T.M., Bastiaens, T.J., Kirschner, P.A.: A five-dimensional framework for authentic assessment. Educ. Technol. Res. Dev. **52**(3), 67–86 (2004)
9. Levy, R.: Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. Educ. Assess. **18**(3), 182–207 (2013)
10. Quellmalz, E.S., Pellegrino, J.W.: Technology and testing. Science **323**, 75–79 (2009)
11. Roelofs, E.C., Straetmans, G.J.J.M. (eds.) Assessment in actie [Assessment in action]. Cito, Arnhem (2006)
12. Shavelson, R.J., Baxter, G.P., Gao, X.: Sampling variability of performance assessments. J. Educ. Meas. **30**(3), 215–232 (1993)
13. Shavelson, R.J., Ruiz-Primo, M.A., Wiley, E.: Note on sources of sample variability in science performance assessments. J. Educ. Meas. **36**(1), 56–69 (1999)
14. Sijtsma, K.: On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika **74**(1), 107–120 (2009)
15. Ten Berge, J.M.F., Sočan, G.: The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika **69**, 613–625 (2004)
16. Verhelst, N.D.: Estimating the reliability of a test from a single test administration. Measurement and Research Department Reports 98-2. National Institute for Educational Measurement, Arnhem (2000)