DESIGN CRITERIA OF VIDEO ELEMENTS AND THEIR EFFECTS ON SIMULATION BASED COMPETENCE MEASUREMENT

Janne Kleinhans, Thilo J. Ketschau

University of Goettingen (GERMANY)

Abstract

In recent years, vocational education and training (VET) has changed its perspective from transfer of knowledge towards competency development. This occurs, because the transfer of declarative knowledge is not sufficient to develop skills which are necessary to solve complex tasks in a working environment [1], [2]. Therefore, it is crucial for competence measurement to create authentic action situations to improve the validity of tests, which should result in a higher adaptability of test results on reality and everyday situations (cp. [3]). Video elements have the potential to increase the validity of tests for competency measurement significantly, and therefore ensure that test results reflect the test persons skills for actual problem solving ([4], [5]).

The quality of technical equipment used in institutions of VET is constantly improving. Computers and electronical test centres allow the application of video elements on a large scale for training and education. While there is a variety of research on the use of videos in teaching (cp. [6]), the determinants of design and use of video elements for assessments is hardly examined (cp. [7], [8]). This is the motivation of this contribution.

For a study on the vocational competencies of qualified medical employees, a computer-based competency test was designed and applied based on ILIAS, a digital learning management system. Nearly 1,000 test persons have been assessed. The purpose of this contribution is to empirically examine the effects of different aspects of video design on the behaviour of the participants, using the results of this competency test.

First, criteria for classification for the test items and videos are elaborated. Second, the effects of these criteria are tested statistically using correlation analyses. The analyses are based on the data of 21 videos and the related items. Statistical criteria include the parameters of the items and determinants of difficulty levels as well as motivational and emotional aspects. As a result, the contribution shows the effects of different aspects of video design – like the charisma of the actors or the presence of superiors – on the difficulty of the item and the motivation of the test persons. Furthermore, technical aspects – like the length of the videos or the format of the items – are considered.

The results are a contribution to the research on test design and the use of video elements in competence measurement. Practically, the findings may be used to improve the construction of video-based items. The results are not limited to the context of VET, they may be relevant for all areas of education.

Keywords: multimedia-based testing, video design, video-based testing, computer-based testing, test difficulty, test motivation, competence measurement, assessment.

1 INTRODUCTION

The importance of multimedia and especially video elements in vocational education and training (VET) has strongly increased in the last years. The most important reason is a general progress in VET from knowledge orientation towards a stronger competency perspective. In this change the focus is moved to the ability to solve situations in a real context, instead of offering declarative knowledge about such situations [2]. Therefore, competence definition exceeds factual knowledge and includes personal, activity and decision-making, subject and method, as well as social-communicative facets [1]. To measure these abilities, it is essential for competence diagnostics to create authentic action situations. If represented situations in testing environment become more accurate, a higher adaptability to everyday situations is possible (cp. [3]). At this point multimedia elements become an important tool to ensure that the created test situation addresses the test persons’ skills for problem solving. For example, audio files can simulate phone calls. They can be a very simple way to express emotional settings in conversations indirectly, which could not, or only very explicitly, be expressed in...
text form. For instance, in medical settings, this can be important to understand the situation of a patient adequately. Additionally, video elements offer further opportunities. They enable to transfer complex scenarios, including the behaviour of protagonists, directly to test persons. In medical settings for example, an emergency can be used to create stressful moments, which could hardly be transferred with the same urgency in text form. That way video elements can increase the validity of tests for competency measurement significantly [4], [5].

A second, more practical, reason for the growing importance of video elements in testing is the ongoing improvement of possibilities to use them on a large scale for training and education. Powerful computers and electronic test-centres become increasingly available in VET [4], [9], in research as well as in educational praxis.

Whereas the usage of videos in learning settings has a long tradition from at least 30 years, their use in assessment environments represents a relatively new approach. Therefore, current literature concentrates on aspects in combining video elements with lectures for learning purposes (cp. [6], [10]), whereas the influence of specific design elements in the videos are neglected (cp. [8]). For assessment purposes it is however important to understand which design elements of videos influence test persons' perception and the difficulty of an item.

The lack of research on the influence of design elements of videos in VET is the motivation of this contribution. The following research questions are formulated.

**RQ 1**: Which design elements influence participant’s motivation of video based items in a competence measurement environment?

**RQ 2**: Which design elements influence the difficulty of video based items in a competence measurement environment?

Since lots of current research focusses on the question of how video usage could promote participants' motivation in learning environments, it seems to be interesting, which design elements foster or prevent a high motivation. The second research question is formulated explicitly regarding the assessment context, in which the difficulty of an item gains high importance for the item and assessment design.

This study builds on an extensive research project for competence measurement with 997 collected data sets. The study was conducted in 2014 for medical assistants in vocational schools throughout Germany.

Remarks on status of literature and research outline follow in section 2. In section 3 the main study is presented. The contribution ends with a conclusion in section 4.

## 2 LITERATURE & OUTLINE

Most literature regarding the multimedia usage in VET focusses on learning processes. In general a positive influence on learning results and motivation of learners is pointed out [10], [11]. Videos can be more effective than texts in presenting specific situations through problem-based instructions and increase learners satisfaction, comprehension and retention [12]. Studies regarding the same effects in assessment context are rare. However, existing results confirm a higher efficiency in presenting complex situations [5], [13].

There is several research on the possibilities of multimedia enriched learning to increase motivation and attention for the learning process. Influencing factors can be found in emotional appealing design elements or decorative instructions [14]. Another factor is the interactivity of multimedia based situations [15], [16] and especially the fun to interact with the dynamic environment in more game-based settings [17]. However, one has to notice in this case that the primary motivating factor might not be the multimedia aspect itself but in fact the more game-based structure compared to traditional learning materials and lectures [18].

Concerning the difficulty of multimedia based learning processes, cognitive load theory [19] and the cognitive theory of multimedia learning [20] are widely used in literature to describe influencing factors from a cognitive point of view. However, they are still focused on learning processes and cognitive demands in learning processes [10], [21]. Furthermore, there still is a lack of research on elements in video design which may result in higher or lower cognitive load even for learning processes [22].
It may be argued that test motivation in assessment situations is neglectable for further research, as especially in exam situations there is normally a high extrinsic motivation. This must be rejected, since it has to be a pedagogical premise to motivate participants to gain best possible results. It can be assumed that a higher intrinsic motivation is useful to remain test persons concentrated. Especially in terms of competence measurement, a high degree of identification and involvement in the situation is also desirable to create an authentic environment to adapt the test results on action reality. But even if there is no positive effect on the validity of the simulated situation – which cannot be proved with current literature so far – results concerning a positive effect on motivation could expand the research for learning literature.

Possible influences from item design to item difficulty are however undoubtedly of great importance to understand requirements for item design in assessment context. For balancing the test difficulty for several items in assessment construction, it is essential to understand relevant design factors. Therefore, effects of video design on motivation and item difficulty are investigated.

3 MAIN STUDY

In this chapter the main study is described. Section 3.1 focusses on the data set. Section 3.2 depicts the indicators and criteria as dependent and independent variables for the statistical analysis following in section 3.3. The chapter ends with the discussion of the implications in 3.4.

3.1 Data set

The data set was collected through the large-scale competence measurement project CoSMed (Competence measurement based on Simulations and adaptive testing in Medical settings). The project focusses on medical and economic-administrative competences of medical assistants in vocational training [23]. The project focusses on medical and economic-administrative competences of medical assistants in vocational training [23]. The project focusses on medical and economic-administrative competences of medical assistants in vocational training [23]. The project focusses on medical and economic-administrative competences of medical assistants in vocational training [23]. Therefore, it requires the completion of an intermediate and final examination. ILIAS was used as test software (cp.[24]).

The video based test was just one element of the overall competence measurement test. In combination with a not multimedia-based adaptive test [24] it was used to measure professional competences. The data set was collected in the last year of the training in 37 vocational schools in five federal states. 1,155 apprentices were tested. After data cleaning 997 data sets remained for analysis. 79 different items were used [23]. The items were organized in five test sequences. As test sequences and other test parts were extensive, there was no possibility to test all sequences for every participant according to organizational constraints. Therefore, three different test outlines with three to four sequences were created. The test time for each outline was limited to 75 minutes. One sequence was fixed in each outline and was used as set of anchor items. The order of sequences was varied to minimize positional effects of items, so four different versions of each of the three outlines were presented. For each item at least 194 and on average 573 responses were collected.

24 different videos were used. The length of the videos was kept short to increase test efficiency. It varied between eleven and 74 seconds. Average length was 30 seconds. All videos were recorded by a professional director and cameraman. As location two real doctor's offices were used. To create authentic situations furthermore medical experienced stuff was casted for the role of doctors and medical assistants in the videos. All videos were shot from an outside neutral, narrative perspective. In the test the observer is presented a specific situation in a video watching different characters interacting with each other. One character is always a medical assistant, others could be a doctor or a patient. After the end of the video the instruction asks the participant to take the position of the presented medical assistant in order to answer the question.

Videos and items were integrated in the test on individual screen pages to increase attention for the video contents. However, it was possible to go back to a video in the test and repeat it as often as wanted. It must be noted that typically one video is followed by several items in the test to increase test efficiency. Therefore, two different kinds of items must be differentiated. There are items directly following and items not directly following a video. However, to generate a coherent multimedia appearance of the test, items not directly following a video were enriched with a graphic at the position of the video. Therefore, freezes from the video were taken. From a total of 24 items directly following a video in the assessment, three items had to be taken out of analysis due to bad scaling results. So,
the analysis builds up on 21 items, which are positioned directly after a video to minimize possible side effects. However, in a second step differences to the other items were explicitly analysed.

### 3.2 Methodology

Two indicators are used to identify the effects of the videos. The first indicator is the Skipping Rate (SR), which represents test persons’ motivation to answer or to skip an item. The SR was calculated as a quotient. The absolute frequency of missings by skipping is divided by the difference of the absolute number of test persons and the absolute frequency of system determined missings. System determined skipping becomes relevant because not all items were presented to the whole subject group (cp. 3.1), furthermore this includes missings caused by participants which did not finish the test. It must be noted that test results during data collection had no consequences for participants. Even if data collection was conducted in a realistic assessment environment, extrinsic motivation to complete the whole test for test persons was therefore lower than in an exam.

The second indicator is the Response Model Parameter Estimate (RMPE), indicating the item difficulty. The RMPE was calculated for every item of the assessment by statistical scaling of the assessment results using the Rasch model. In result, the RMPE is represented by a numerical figure which is not standardized, but describes the difficulty of an item in relation to the other items of the assessment for the sample [25].

In total, twelve criteria are used for analysis. The criteria have all been transformed in dichotomous or ordinal-scaled numeric figures. Each video and corresponding item were coded independently for each criterion by the two authors. Afterwards results were combined to one coding pattern. Few differences occurred, the interrater-reliability was at 0.92. Differences were extensively discussed and mean ratings established for differing aspects. In the following, the criteria are described.

As already mentioned, there are very few studies concerning the concrete design of videos or at least multimedia elements in assessment contexts. Existing studies focus on the possible combination with different question types [7], [8]. Therefore, the influence of different question types was examined. Item format stands for the technical type of item. A differentiation was made between multiple choice, fill-in-the-blank choice, image map and open-text format, assuming that there is a raising complexity of the item formats in this order. In a multiple-choice item format, pre-defined options must be selected directly. For fill-in-the-blank choice items there are typically more possible answers, however, they are pre-defined and can be selected from a dropdown menu. In image map questions, a pre-defined area of an image has to be selected. These areas are invisible at the beginning of a task. Once an area is selected, it is highlighted. Afterwards it can be deselected again, if a participant wants to change his choice. Therefore, there is still a support for test persons in identifying possible solutions. In free text-tasks there is no support. So, it can be assumed that the cognitive effort to handle the items is raising among the described formats [26].

Videos can be a tool to display situations neutrally [27]. Whereas in a text, the behaviour of persons must be described by using – necessarily – already categorizing attributes, test takers in videos can gather their own impression. In medical settings for example it can be important to detect a special emotional state of a patient for acting adequately in a situation. Therefore, the way, actors behave in a video, represents a significant design tool for video-based testing and has to be analysed. It is interesting to see, if participants decode such given emotional states in videos appropriately and if this information influences them in their response. Therefore, the criterion positive_perception_patient was created to gather if a patient leaves a friendly impression. Contrary the criterion negative_perception_patient refers to an unfriendly impression. Doctor_present indicates, whether the doctor and therefore the supervisor of the protagonist is present or absent in the video. Likeable_doctor indicates a friendly impression of the doctor, if he is present. Doctor_instruction examines, if the present doctor gives an instruction to the protagonist and, hence, the test-taker.

Participants were confronted with the situations from an outside perspective in the testing environment. By answering the tasks, they took the position of a medical assistant, shown in the video. That is why empathy_protagonist was constructed as criterion to analyse, if empathy shown by the protagonist influences the response to the items. Professionalism_protagonist means how professional or unprofessional the protagonist acts in the accomplishment of his or her tasks. Therefore, for both criteria a possible effect is investigated.

Medical assistants can be confronted with high emotional stress in their work. This aspect is adopted by two criteria. Emotional activation of compassion indicates whether the video scene is used to
trigger compassion in test persons or not. Creation of urgency considers, if the situation in the video provokes an urgent need to act for the protagonist, normally by presenting a medical emergency.

Furthermore, technical criteria are incorporated. Quantity prior videos is the number of videos which were presented in the test sequence prior to the referred video. At last, the length of the video in seconds (length_video) is used as criterion.

The number of cases for each criterion varies between the items, because not every video contains all the criteria, for example an urgency was only part of twelve videos. In the analysis of item_format one item was excluded, because the format of this item was used only once, offering no sufficient data base for this item type.

3.3 Results

Indicators and criteria were analysed statistically on correlations. Since both indicators are metric, Spearman coefficient was used for non-metric and Pearson coefficient for metric criteria [28]. The results are shown in table 1, with non-metric criteria in the upper and metric criteria in the lower part.

Table 1. Correlations of indicators and criteria.

<table>
<thead>
<tr>
<th>Non-parametric correlations (Spearman)</th>
<th>Skipping Rate (SR)</th>
<th>Response Model Parameter Estimate (RMPE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>item_format</td>
<td>coefficient of correlation 0.360</td>
<td>0.444*</td>
</tr>
<tr>
<td></td>
<td>significance 0.059</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>number of cases 20</td>
<td>20</td>
</tr>
<tr>
<td>doctor_present</td>
<td>coefficient of correlation 0.754**</td>
<td>0.433*</td>
</tr>
<tr>
<td></td>
<td>significance 0.000</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>number of cases 21</td>
<td>21</td>
</tr>
<tr>
<td>positive_perception_patient</td>
<td>coefficient of correlation 0.484*</td>
<td>0.490*</td>
</tr>
<tr>
<td></td>
<td>significance 0.047</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>number of cases 13</td>
<td>13</td>
</tr>
<tr>
<td>negative_perception_patient</td>
<td>coefficient of correlation 0.242</td>
<td>-0.267</td>
</tr>
<tr>
<td></td>
<td>significance 0.213</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>number of cases 13</td>
<td>13</td>
</tr>
<tr>
<td>likeable_doctor</td>
<td>coefficient of correlation 0.030</td>
<td>-0.120</td>
</tr>
<tr>
<td></td>
<td>significance 0.465</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>number of cases 11</td>
<td>11</td>
</tr>
<tr>
<td>doctor_instruction</td>
<td>coefficient of correlation -0.517*</td>
<td>-0.065</td>
</tr>
<tr>
<td></td>
<td>significance 0.050</td>
<td>0.425</td>
</tr>
<tr>
<td></td>
<td>number of cases 11</td>
<td>11</td>
</tr>
<tr>
<td>empathy_protagonist</td>
<td>coefficient of correlation 0.260</td>
<td>0.308</td>
</tr>
<tr>
<td></td>
<td>significance 0.128</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td>number of cases 21</td>
<td>21</td>
</tr>
<tr>
<td>professionalism_protagonist</td>
<td>coefficient of correlation 0.191</td>
<td>0.422*</td>
</tr>
<tr>
<td></td>
<td>significance 0.224</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>number of cases 18</td>
<td>18</td>
</tr>
<tr>
<td>compassion</td>
<td>coefficient of correlation 0.025</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td>significance 0.469</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>number of cases 12</td>
<td>12</td>
</tr>
<tr>
<td>urgency</td>
<td>coefficient of correlation -0.319</td>
<td>-0.205</td>
</tr>
<tr>
<td></td>
<td>significance 0.156</td>
<td>0.262</td>
</tr>
<tr>
<td></td>
<td>number of cases 12</td>
<td>12</td>
</tr>
</tbody>
</table>
There are several significant correlations between indicators and criteria. SR is correlated with a strong positive effect to doctor’s presence and with a negative effect, when the doctor gives an instruction. A medium positive effect is given to a positive perception of the patient.

For RMPE there are more effects. RMPE is correlated with a medium positive effect to item formats, doctor’s presence, positive perception of the patient, professionalism of the protagonist and emotional activation of compassion. Before detailed implications are made in 3.4, the results are analysed further comparing items directly following a video and the other items.

Therefore, the dichotomous variable *directly_following_video* was created, conducting if an item directly follows a video. It was correlated to SR to gain evidence, whether test persons respond to items significantly more often, when items directly follow a video. Furthermore to get evidence whether the complexity of the item format is influenced by a direct position behind a video, the statistical correlation to RMPE was checked for all items of the assessment. Table 2 shows the results.

There are two possible explications: The videos might not influence the following items at all. In this case it is not plausible, why there are effects of several criteria on SR and RMPE as mentioned in table 1. Therefore, it seems more likely that videos influence the whole following sequence of items. This effect was also intended in test construction by enriching items, not directly following a video, with graphics.

Participants were asked at the end of the competence measurement, if they liked the video test. A likert-scale with five steps was used. As the question was at the end of the test, many test persons failed to answer this question in time and motivation was lower than in the content related tasks before. However, 533 answers were collected. Participants showed high satisfaction with a mean value of 4.4 of the multimedia based test. As there were other, not multimedia enriched, test parts (for example the adaptive test) test persons were asked, how they evaluated electronic tests in general. The average value was 3.2. Furthermore, questions were posed, if videos were helpful to get into the situations. This was clearly affirmed with a value of 4.2. Only a value of 1.8 was measured according to the question, if videos were redundant and therefore partially obsolete. These results underline the high positive influence on participants’ satisfaction by using multimedia elements.

### 3.4 Implications

Regarding test persons’ motivation to answer or to skip an item, calculated as indicator SR, there is a positive correlation with the doctor’s presence. In contrast, the doctor giving an instruction has the effect of a decreasing SR. One possible interpretation could be that presence of the superior leads to lower received self-responsibility. On the other hand, concrete instructions lead to the opposite effect.
It becomes clear that the integration of superiors in videos has great influence on the behaviour of test participations. This is an indication for the assumption that the video usage is effective in creating authentic situations, as participants seemed to react on the presented principal.

More difficult to interpret is the correlation of SR and a positive perception of the patient, since it seems more plausible that sympathy stimulates the motivation to deal with a task. This aspect has to be researched further. One explanation could be that a positive perception of the patient is perceived less challenging and therefore less motivating in the test situation. This would, however, indicate in contrast to the paragraph before that people are still highly aware of the test as an artificial situation.

Regarding the difficulty of the items, RMPE is positively correlated to complex item formats. As mentioned, there is no difference if an item is directly following to a video or not. This indicates that the question type influences item difficulty independent from video design. Further research should be conducted to investigate, if videos influence the difficulty of some item formats in a special way. In this study no corresponding evidence was found.

RMPE is positively correlated to a positive perception of the patient and especially compassion regarding the patient. The same occurs, when an authority – the doctor in this case – appears in the video. It can be assumed that all of these effects increase pressure to react adequately for the test persons and therefore the difficulty. Moreover, it is plausible that the feeling of compassion causes an emotional involvement or emotional stress as an interference of cognitive performance. Participants were confronted with the situations from an outside perspective in the testing environment. By answering the tasks, they took the position of a medical assistant, shown in the video. If the depicted competence of the medical assistant in the video is higher, then RMPE is raised, too. This promotes the view that there is transmitted some kind of identification process between the test taker and the presented protagonists in the videos.

It must be noted that a correlation was not found for a negative perception of the patient and the positive or negative perception of the doctor or the protagonist. So, it cannot be stated in general that the perception of characters in videos as likeable or unlikeable is a determinant of item difficulty. The presence or absence of an authority seems to have more influence than his behaviour. One explanation could be that videos and sequences were too short to develop different attitudes between different authorities in the videos. Equally interesting is that pressure on the test persons, caused by an urgency in the video situation, has no effect. Further research is needed to differentiate this result. It could be argued on the one hand that medical assistants in general are quite experienced in handling stress situations and therefore reacted very professionally in the test situation. On the other hand, this could be another cue that participants still were aware that the created situation was just simulated and the authenticity of the emergency situations was too low.

Besides these effects, there are also cases which are interesting because they show no significant correlations. The length of the video and the quantity of prior videos display no effects, though it may be argued that long or many videos tend to demotivate the test persons, for example by boring them. This could be an indication that the target to keep the videos short for maximum test efficiency payed off. However, it must be stated that the variance in the related variables was low, so a generalizability of these assumptions would not be legitimate at this point.

4 CONCLUSION

This research paper investigated the influence of video design elements on difficulty and motivation. Therefore, results from a competence measurement project in medical settings with 997 data sets was analysed. Many contributions state that videos elements can increase participants’ motivation and satisfaction in learning situations. This result can generally be confirmed. It was found that enriching tests with multimedia elements has a general positive influence on user satisfaction even in assessment settings. However, the effect was not limited to items, which directly refer to a video. This could lead to the conclusion that the positive motivation effects of videos can be used for whole sequences, even if items do not directly follow a video.

There is few research concerning the question, which specific elements have effects on participants’ motivation. Results of the current investigation demonstrate that the integration of superiors – in this case the doctor in relation to the medial assistant – increases motivation, if instructions are given. Without instructions, however, less motivation was identified. One possible explication is a lower perceived self-responsibility if superiors are present.
Regarding the difficulty of video items several elements were found, which determine item difficulty. A connection of these elements is likely to the perceived pressure to react adequately in the given situation. A positive recognition of a patient and especially compassion regarding the patient are examples. The same occurs, if a superior appears in the video. However, further research is needed to validate these assumptions.

Neither on difficulty nor on participants’ motivation consequences could be found from the length of the video or number of preceded videos.

In general, the results indicate that there were significant effects of video design on test motivation and item difficulty. The results of this study promote the view that videos enable some kind of identification process between test persons and the presented main actor in the videos, in this case the medical assistant, whose perspective later was taken to answer the items. Also, the above-mentioned effects on difficulty and motivation underline the potential of videos in creating authentic situations. Very few of these effects would be plausible, if the test situation was just observed from a neutral external point of view. Furthermore, these results – and especially the impact of perceived compassion for patients – underline the assumption that videos can transfer emotions to test persons in assessment settings.

There are some limitations of the study. Considering the design of this study and the relatively small sample of eleven to 21 items (depending on the criterion) it should be emphasized that all results cannot be generalized without further appropriate research. Such studies should be conducted as experiment, allowing the controlled modification of video criteria. Another limitation of the current data set is that most important information from the video were explicated additionally in text form in the items. Therefore, effects from memorization or a high and low information density in the videos could not be observed.

REFERENCES


