Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 26 Number 15, June 2021 ISSN 1531-7714

Eight Issues to Consider when Developing Animated Videos for the Assessment of Complex Constructs[1]

Anastasios Karakolidis, *Educational Research Centre*
Darina Scully, *Dublin City University*
Michael O'Leary, *Dublin City University*

As part of the growing interest in the measurement of complex constructs in recent years, a body of research examining the extent to which videos are a useful alternative to written text in tests and assessments has emerged. Early attempts to replace written text with videos featured actors, but lately, animated videos have become more popular. However, the few studies that have examined the use of videos (animated or acted) in assessment have focused purely on reporting the results of these endeavors, with little to no information provided about the process of transforming a test from text to video format. With this in mind, the aim of this paper is to outline the key issues that need to be considered when developing animated videos in an assessment context. Various decisions that need to be made are discussed and suggestions for overcoming challenges that may be encountered are offered. These considerations are intended to help anyone interested in the use of animated videos to enhance the validity of decisions made on the basis of assessments, including, but not limited to, educators, certification and licensure test developers, and those involved in personnel selection.

Background

The world is becoming increasingly interested in complex knowledge and skills that promote social transformation (United Nations Educational Scientific and Cultural Organization [UNESCO], 2014). There is now widespread acceptance that the knowledge and skills traditionally emphasized in education and valued in the workplace may no longer be optimal – or sufficient – to succeed in contemporary society (He et al., 2017). In recent years, increasing attention has been directed towards *21st century skills*, a combination of cognitive and

[1] This research was supported by a grant from Prometric to Centre for Assessment Research, Policy and Practice in Education (CARPE). The content of this paper has not been influenced in any way by Prometric, and is solely the responsibility of the authors. The authors gratefully acknowledge Prof. Steve Stemler (Wesleyan University) for granting them permission to adapt and use the Tacit Knowledge Inventory assessment, Ed Lozano (AnimatedScenarios.com) who developed the animated videos, and Pavlos Stampoulidis (Psycholate) who created the testing platform. The authors are also indebted to Vasiliki Pitsia for her assistance with various aspects of the project, and to colleagues at Dublin City University's Institute of Education for providing invaluable feedback on the pilot items.

non-cognitive *soft skills*, with the latter referring, in particular, to how people communicate and work in teams (Griffin et al., 2012; Riggio, 2014; Vandeweyer, 2016). Problem-solving, critical thinking, creativity, communication and collaboration are just some examples considered to be essential for success in the modern world (Binkley et al., 2012).

In the context of the growing interest in more complex knowledge and skills, the World Economic Forum (2015) highlighted the need for more effective measurement of such constructs. As Griffin and Care (2015) argued, traditional forms of assessment may not be suited to the measurement of many of these skills. In reality, the vast majority of tests in use today rely heavily on the use of text to present both stimuli and response options, and although written language is often a good fit for measuring traditional constructs such as knowledge of historical events, it may not a suitable medium for presenting the type of complex information that is needed to facilitate the measurement of more sophisticated, higher-order skills, such as problem-solving, communication or practical knowledge (Popp et al., 2015). Indeed, text-based test items designed to measure such skills often require the use of longer, complex pieces of text, and this linguistic complexity may introduce *construct-irrelevant variance*[2] for certain groups of test-takers, in that their performance can be negatively affected by factors, such as reading comprehension and proficiency in the language of the test, that are beyond the focus of the assessment (Abedi, 2010).

Examples of such tests are situational judgment tests (SJTs), where test-takers are provided with descriptions of challenging real-life situations and a number of possible ways to deal with the given problem (Motowidlo et al., 1990). These tests typically purport to assess skills such as leadership, interpersonal skills or emotional intelligence (Christian et al., 2010). However, the "noise" caused by heavy reading demands is likely to render them unsuitable for people who may have the relevant knowledge and skills, but lack sufficient language fluency and/or reading comprehension skills (Popp et al., 2015).

There is a growing body of research literature investigating the extent to which the use of videos can help alleviate this problem. Early attempts to replace written text with videos featured human actors (Chan & Schmitt, 1997; Kanning et al., 2006; Lievens & Sackett, 2006; Richman-Hirsch et al., 2000), but lately, animated videos have become popular (Bardach et al., 2020; Bruk-Lee et al., 2016; Dancy & Beichner, 2006; Karakolidis et al., 2021). Animated videos represent a distinct and relatively unexplored option that may have significant advantages over acted videos. In contrast to acted videos, animations can be changed and modified relatively easily making it possible to correct errors and/or to keep the instruments up-to-date over time. Moreover, animations can be adjusted, in terms of language, character features, clothing and location to make them more suitable for use across countries and cultures, something that is not as easily achieved when human actors are involved (Popp et al., 2015).

Overall, the limited number of research studies that have used videos (acted or animated) have provided promising findings in relation to reducing construct irrelevant variance and/or improving test-takers' reactions to the test. However, research in the field, particularly with regard to animations, is still relatively rare, and some conflicting findings have emerged. Two of the most recent studies compared animated to conventional text-based SJTs in the context of pre-service teacher education. A study by Karakolidis et al. (2021) found that the use of animated videos slightly reduced the dependency of the test on construct-irrelevant factors (e.g., language and reading skills), and also had a positive impact on face validity and the test-taker experience. However, research published by Bardach et al. (2020) suggested somewhat different findings in that the use of animations, either instead of or in addition to written text, failed to reduce the adverse impact on minority groups' performance on the test, while mixed results were found in relation to takers' perceptions of the different versions of the test.

Of note in the context of this paper is that all of the published literature on the use of videos in assessment has been focused purely on the outcomes associated

[2] Construct-irrelevant variance can be defined as the measurement of phenomena that are not included in the definition of the construct of interest (Frey, 2018). It is considered to be one of the biggest threats to the validity (American Educational Research Association [AERA] et al., 2014).

with their use. What is lacking is specific information about the process of developing these instruments, particularly those involving animations, that would guide others interested in pursuing similar research.

This paper attempts to address this lacuna by documenting some of the key considerations that need to be borne in mind when developing animated SJTs. The discussion to follow is organized around eight issues: i. value added, ii. fidelity of representation, iii. facial expressions, voice, and movement, iv. cognitive load, v. situational contexts, vi. response options, vii. testing platform, and viii. financial cost. The paper draws on the relevant literature in the area combined with the authors' practical experience of developing an animated SJT of teachers' practical knowledge over the course of three years (see Karakolidis et al., 2021).

I. Consider the value added

Not all tests stand to benefit from the use of video technology. The first and one of the most critical decisions that has to be made is the extent to which the validity of the judgements made on the basis of the existing assessment is likely to be enhanced through the use of animations. It is critical to avoid technocentric thinking, i.e., trying to incorporate technology into assessment just because it is feasible or with the primary aim to make the assessment look more attractive. There should be a good reason and value to be added from the use of technological applications, in this case, animated videos.

To begin with, the test should go beyond the measurement of knowledge recall. The hypothesis that videos can improve assessment primarily applies to the measurement of more complex knowledge and skills that cannot be adequately captured by conventional text-based instruments. The incorporation of animated videos into straightforward, knowledge-based multiple-choice tests may not add any value to the quality of the assessment. Tests that require test-takers to process sophisticated information, on the other hand, are more likely to benefit from the incorporation of videos that can facilitate the communication of multi-faceted messages.

Assessment of complex skills is often undertaken through the use of long passages of text, providing test-takers with sophisticated information that needs to be fully comprehended before the questions are answered (Scully, 2017). However, such practices may introduce construct-irrelevant variance into the assessment. It follows that such tests can potentially be improved by the use of animated or acted videos.

SJTs are a good example of an assessment type that may benefit from the use of animations due to their heavy use of text. Moreover, as these assessments are typically designed to assess communication and interpersonal skills (Christian et al., 2010), human interactions are often a central feature of stimuli in SJTs. As such interactions are difficult to recreate authentically with text, it stands to reason that the benefits of using multimedia are potentially greater with respect to these types of assessment. Using videos to present these scenarios can enhance the fidelity of the stimulus (i.e., the way in which these scenarios are encountered in real life is more closely approximated). Indeed, many of the studies that have attempted to incorporate video technologies in assessment have used SJTs (e.g., Bardach et al., 2020; Bruk-Lee et al., 2016; Chan & Schmitt, 1997; Lievens & Sackett, 2006).

II. Consider the fidelity of representation

Animated characters can be presented in different formats that vary in their level of authenticity: (i) two dimensional (2D) animations, which have the lowest level of authenticity, (ii) three dimensional (3D) caricatured animations, which are more realistic than 2D animation but still not lifelike, and (iii) 3D realistic animations, which approach lifelike appearance (Popp et al., 2015).

Generally speaking, levels of fidelity can vary significantly among different types of tests that use multimedia (e.g., audio, acted videos, and animations). For example, adaptive simulations, which allow test-takers to interact with a virtual environment, have higher fidelity than static tests that present test-takers with a scenario and ask for a response to a series of selected-response items (Lievens & De Soete, 2012; O'Leary et al., 2018). Likewise, a test presenting candidates with a realistic virtual environment and humanlike characters is expected to have higher fidelity than a simulation that shows animated scenarios using unrealistic caricatured characters. However, higher fidelity, in terms of representation of the real world, does not necessarily lead to better assessments (Mislevy, 2011).

There is a discourse in the research literature regarding the degree to which humanoid objects (e.g., robots) and animated characters should be realistic. It

has been hypothesized that humanlike objects, which are designed to be very realistic, may evoke a feeling of eeriness in some viewers. This hypothesis was first introduced by Japanese robotics professor Masahiro Mori in 1970 (Mori, 1970). He noticed that trying to make robots more humanlike would increase perceivers' affinity for them up to a point (when characters appear 80-85% humanlike). Once that point of similarity was exceeded, viewers experienced an eerie sensation. This phenomenon is called the *Uncanny Valley*. Although this theory originated in the field of robotics, it is applicable in the field of character-based animations as well. Indeed, animated characters that have been designed to be very realistic have also been shown to evoke a feeling of eeriness to some viewers (Dill et al., 2012; MacDorman et al., 2009). Some well-known animation productions, such as *The Polar Express*, have been criticized for having characters that are too realistic and make the audience feel uncomfortable (Misselhorn, 2009). MacDorman et al. (2009) argued that this might be the case because the more human a character looks, the easier it is to identify its imperfections. Indeed, there is research evidence suggesting that the key is not to try to imitate a real-human appearance but to design clearly non-human characters that have the ability to portray real-human emotions (Hawkes, 2012b, 2012a; Schneider et al., 2007).

Apart from human likeness, there are other factors that may influence how animated characters are perceived (MacDorman, 2006; Mathur & Reichling, 2016). Hanson (2005, 2006) argued that very abstract and cosmetically atypical robots and animated characters can be uncanny, regardless of their degree of human likeness. He demonstrated that well-designed characters with large expressive features, clear skin, well-groomed hair and other characteristics that are often considered to be aesthetically pleasing could eliminate the phenomenon of the Uncanny Valley. Hanson (2005, 2006) admitted, though, that the design of very realistic characters is more challenging because they trigger higher expectations from the perceiver's point of view.

Avoiding the Uncanny Valley may be quite important because negative perceptions of any aspect of a testing experience may ultimately impact on test-takers' overall attitudes towards both the assessment process and the body organizing the assessment (Popp et al., 2015). Nevertheless, the most important question remains whether the quality of a measure or test-takers' performance can be affected by the nature of the animations. Indeed, it has been evidenced that the choice of the multimedia used in a test can impact not only test-takers' attitudes towards the assessment, but also their responses to and engagement with the assessment process (Bruk-Lee et al., 2016; MacDorman et al., 2010). This probably renders the use of unappealing characters in testing problematic, especially in areas where engagement and empathy are essential, such as in interpersonally-oriented assessments.

Figure 1 presents some examples of the main animated characters used by Karakolidis et al. (2021). The 2D caricatured animated characters used were expected to provide satisfactory authenticity without leading to negative reactions. The ultimate goal should always be to offer test-takers a pleasant virtual experience which does not evoke any eerie sensations that could distract them from focusing on the assessment.

III. Consider animated characters' facial expressions, voice, and movement

Animated characters' facial expressions are of great importance for conveying non-verbal messages. The importance of facial expressions for making characters aesthetically more pleasant and less eerie was examined by Tinwell et al. (2010, 2011). These authors concluded that characters who lacked facial expressions in the middle (i.e., cheeks) and upper part of the face (i.e., eyes and forehead), which are the areas primarily involved in the transmission of non-verbal signals, were perceived as being less relatable. According to the authors, the lack of facial movement was an obstacle for viewers when interpreting animated characters' emotions. Consequently, they perceived these characters as eerie and strange. Exaggeration of the mouth expressions, though, was also found to increase the uncanniness.

Despite the importance of facial expressions, it should be acknowledged that animated characters might not always be able to express complex emotions in a non-verbal way; this is something that depends highly on the animation technology used. For instance, 2D animated characters are usually less versatile than 3D ones. However, through the use of other means, such as thought bubbles, characters' emotions and thoughts can be conveyed in non-verbal way (see for example Figure 2).

Figure 1. Sample animated characters

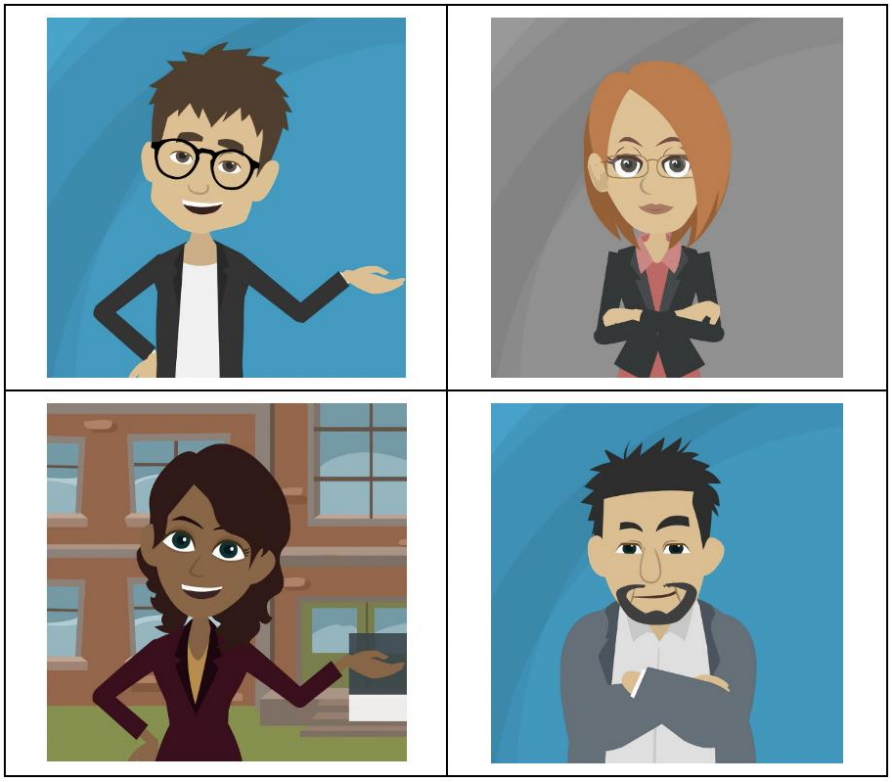
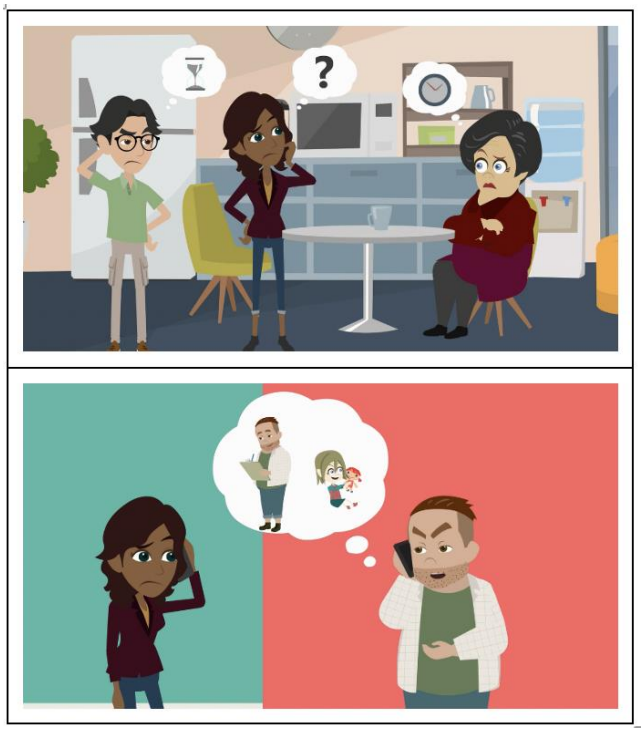


Figure 2. Examples of the use of thought bubbles in the animations



Tinwell et al. (2010) also investigated how various aspects of characters' speech impacted on viewers' perceptions. The findings indicated that monotony and slowness of speech were factors that increased the uncanniness of animated characters. Generally, it could be concluded that both voice and facial expressions should correspond to the degree of human-likeness of the virtual characters to achieve the most favorable outcomes (Mitchell et al., 2011; Tinwell et al., 2010). For instance, it would not be advisable to give a human voice and expression to an animated character that looks more like a robot than a human. It should also be noted that it is not always necessary to give the characters their own voices. The characters used in Karakolidis et al.'s (2021) study, for example, did not speak for themselves, rather, there was a voiceover describing the situation in addition to what they were thinking and feeling. This was reinforced through their facial expressions and the images in the thought bubbles, as described above.

Regarding body movement, White et al. (2007), who designed 3D animated characters with different levels of realistic motion, concluded that smooth and controlled movements were always preferred to more abrupt ones. As with speech quality outlined above, animated characters' appearances should match their behavior so that the Uncanny Valley is not exaggerated. Creating human-like characters who are not able to act like humans can exacerbate the unfamiliarity of the animated characters (Tinwell, 2014).

IV. Consider the cognitive load

Animations, and video-based tests in general, offer higher levels of fidelity than written text. However, this does not necessarily mean that they are always the optimal way of presenting complex information. As Wouters et al. (2008) argued, the fact that animations can present aspects of a situation simultaneously may not always render them better than static representations (i.e., text and images), whereby learners are able to digest information at their own pace. In animated videos, multiple sources of information can interact and simultaneously convey sophisticated messages. This can create substantial extraneous cognitive load, which can place excessive demands on perceivers' working memories, affecting their ability to comprehend the material – c*ognitive load theory* (Sweller et al., 2011).

For this reason, van Merriënboer and Sweller (2005) suggested that, when animations are developed, designers should gradually present information from simple to complex in order to help perceivers to fully comprehend the messages that they receive visually. Furthermore, as Mayer and Pilegar (2014) argued, in animated representations, designers should familiarize perceivers with new concepts before they are exposed to interactions involving these unfamiliar concepts. Finally, a meta-analytical study conducted by Ginns (2005) indicated that, when images and animations are used as learning materials, explanations of the illustrations, when necessary, should be provided in audio rather than in written format.

As has been mentioned, little has been written about the actual development process of animated videos for assessment purposes. However, there are some useful resources that offer a detailed framework of guidelines for designing multimedia and, mostly, videos for instructional purposes. Koumi (2006) suggested that video developers should, first of all, avoid using too much text in the videos. Most of the time, text simply duplicates a message that multimedia tries to convey. As a result, viewers end up processing the animated scenes and the text within them synchronously, losing part of the message. By keeping the use of written text in the animations to a minimum, concerns about test-takers missing critical information communicated via the animations can be minimized. This was the main factor underlying Karakolidis et al.'s (2021) decision not to use subtitles. Although subtitles are usually thought of as being helpful for non-native speakers, in the context of animated assessments, they may create excessive cognitive load and thus distract test-takers from the messages conveyed. The elimination of text is also consistent with the overarching goal of reducing construct-irrelevant variance related to language and reading proficiency.

Koumi (2006) also noted that, in a multimedia environment communicating a lot of information in multiple ways, there is a risk that users may be distracted by peripheral information and miss message being conveyed by the animations. Therefore, efforts should be made, directly or indirectly, to indicate to the user where they should focus. For instance, when new characters were introduced in Karakolidis et al.'s scenarios or when test-takers needed to focus on certain characters' reactions, various effects, such as zooming and highlighting shapes, were employed in an attempt to

capture test-takers' attention and ensure they would not miss any critical information.

Finally, users should always be given enough time to perceive a scene before moving to the next one; especially when complicated information is communicated. Therefore, pauses between the different scenes of a scenario, or even between different messages conveyed in the same scene, can be introduced. Such pauses allow test-takers to reflect on and better comprehend the content of the scenarios.

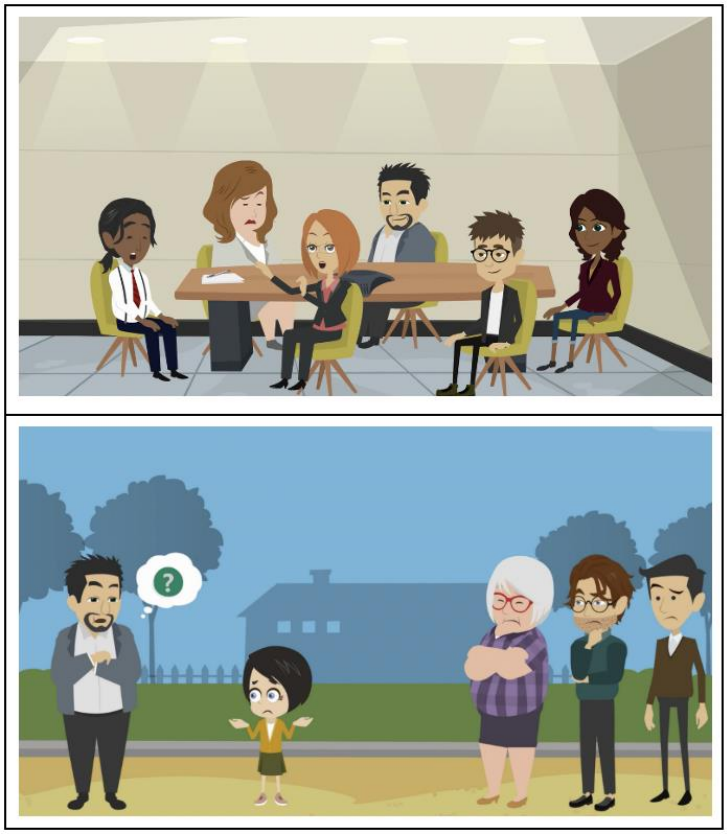
While videos should be paced to ensure that the content is engaging and understandable for examinees, they should not be overly long. In reality, there are no definitive rules about how long each animated clip should be as this will depend on such variables as the age and educational level of examinees, the complexity of the information communicated and the purpose of the assessment. However, excessively long clips should be avoided as this can overload working memory. In Karakolidis et al.'s study, for example, none of the animated videos were more than two minutes long, and participants were able to pause, resume and replay each video as many times as they wished.

V. Consider contextual issues

A particular challenge in the development of animated assessments lies in the fact that some peripheral aspects of the item stimuli suddenly become salient when animations are introduced. Animations necessitate the visualization of aspects that, in text-based assessments, are not necessary to describe. For instance, written scenarios do not always provide any information about the physical appearance of characters and their environments (i.e., skin color, facial characteristics, body, clothes, setting).

Great care needs to be taken with respect to ensuring balanced gender representation, inclusion of characters from a range of ethnic and cultural backgrounds and those with disabilities, and avoiding the reinforcement of stereotypes. A number of stills illustrating gender balance, age diversity and ethnic diversity in animations are presented in Figure 3.

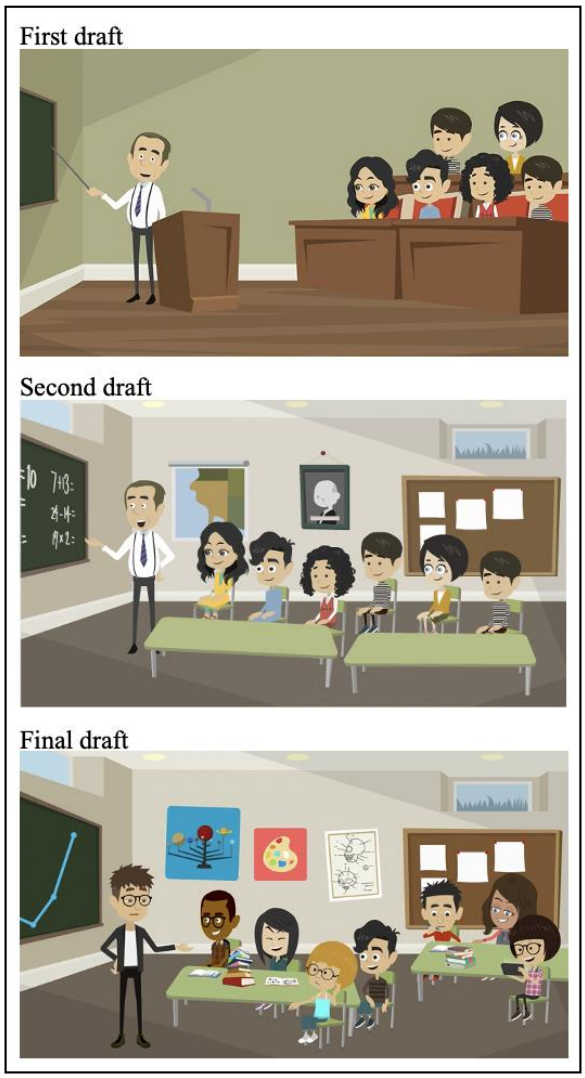
Figure 3. Examples of gender balance and age and ethnic diversity



In text-based tests, adjectives such as disappointed, frustrated, worried and destructive can be used to describe human emotions and behaviors. However, animating such reactions can be very challenging, in terms of determining the intensity of each given reaction. The animation of such reactions or feelings runs the risk of designing characters who either overreact or whose feelings and reactions are not pronounced enough. Achieving an optimal, or at least, acceptable, outcome with respect to this issue may ultimately involve a certain amount of trial-and-error. However, test developers might consider providing additional directions to animators from the offset, especially in situations where the intensity of the emotion displayed may have an impact on the test-takers' response.

The involvement of subject-matter-experts as advisors also becomes particularly important in the development of animated assessments, to ensure that other aspects of the characters' appearance and behavior and details of the background are appropriately depicted, taking into account nuances of the context with which the animators may not be familiar. Figure 4, for example, shows the evolution of an animated scenario set in a primary school classroom (Karakolidis et al., 2021), following input from subject-matter-experts (in this instance, primary school teachers and teacher educators). The final scenario best represents a contemporary classroom – pleasant and colorful, students engaged in collaborative work and using technology, and the teacher playing a facilitative rather than purely instructive role.

Figure 4. The first, second, and final draft of an animated scenario



Finally, it is also important to bear in mind that various scenarios, images, emotions and expressions may be perceived differently across different cultures. In the context of Karakolidis et al.'s (2021) animated SJT, which was administered in Ireland and Greece, differences between the two countries were not expected to be large. For assessments that are intended to be administered across cultures that are clearly very different, however, more attention and work may be needed. Keller et al. (2017) provide some valuable insights into this topic.

Although it can be challenging to control for all these factors, it could be argued that the animation of written text may ultimately result in enhanced standardization of the assessment, as all test-takers are being presented with exactly the same stimuli. In other words, by presenting a given scenario through animation, assumptions and thus, arbitrary interpretations on behalf of test-takers about the characters and settings involved in the scenario are likely to be avoided.

VI. Consider if response options should be animated

Another important consideration in the process of developing an animated assessment is whether the assessment should be partially (item stimulus only) or fully (both stimulus and response options) animated. A review of the relevant literature reveals that the majority of the studies in which a video-based version of a text-based SJT was developed focused on the scenarios, with the response options retaining their original text-based format. However, in most cases, no rationale was provided for such a decision. One exception was a study conducted by Kanning et al. (2006), which compared acted-video SJTs with and without recorded response options. The findings indicated that the use of acted videos in the response options, on top of the video-based scenarios, did not have a statistically significant impact on the face validity of the test. However, the potential impact on other aspects, such as performance or construct-irrelevant variance, was not explored.

Animation of response options as well as the stimuli may in fact create a number of additional complexities. For example, test-takers' responses may be affected by the way the response options are animated. Consider the last option in Figure 5. Respondents' inclination to agree with this practice may be influenced by how upset the child appears when this option is animated. If the animation presents a teacher who was angry and a student who was very upset, almost crying, respondents may be less inclined to select this option, not because of the strategy in question, but because of the depiction of the characters' reaction to this strategy. Indeed, a test-taker might agree that sending a student to the principal would be a good practice, given the situation, but might nonetheless avoid selecting it, because they believe that they would have implemented it in a different way than how it was presented in the animation.

Other issues linked to the animation of the response options are more practical in nature. The animation of the responses, on top of the scenarios, is expected to significantly increase the time required to complete the assessment. This might create an unnecessary burden on test-takers, which could cause ethical concerns (Lingler et al., 2014) and also lead to fatigue and thus, less accurate responses. Last but not least, the animation of the response options can considerably increase the cost of a project.

VII. Consider the testing platform

After finalizing the animated SJT, the next step is to find a platform that can best facilitate the administration of the assessment. The following are identified as key requirements for creating an assessment environment that is engaging and takes full advantage of the animation technology:

- The platform should support high quality image and sound.
- Test-takers should be able to give their responses one by one after watching the animated scenario.
- Test-takers should have the option to watch the video again at any point while dealing with the given scenario.
- The videos and the response options should be easily accessible by the test-takers with minimum scrolling.

Off-the-shelf versions of commercial platforms may not meet all these requirements. As such, it should be borne in mind that the design of a video-based SJT may also necessitate the design of a tailored platform that can support it. Figure 6 provides some screenshots

Figure 5. A sample text-based scenario along with its response options

Mr. Smith is teaching the 6th class this year. For the most part, the students are interested in the topics and listen to what he has to say. Some of the students in the class understand new topics easily, while other students have difficulties understanding basic concepts and ask questions that show they don't understand the content. William, one of the brighter students, is obviously bored with the pace of the class, so he has begun to laugh and make fun of students who ask questions.

Given the situation, rate the extent to which you agree or disagree with each of the following statements.

1	2	3	4	5
Strongly Disagree	**Disagree**	**Neither Agree Nor Disagree**	**Agree**	**Strongly Agree**

Mr. Smith should…

1. tell William, in front of the class, that any further disruption will be punished.
2. ignore William's inappropriate behaviour.
3. speak with William's other teachers to see if he is above average in other subjects, and if this has led to more disruptive behaviour.
4. go over his class rules with the class, emphasizing the importance of respect.
5. talk to William in private and tell him that he recognizes how smart he is and will find assignments that will really challenge him.
6. speak to William privately about his rudeness in class.
7. send William to the principal.

Note. The original SJT items were developed by Stemler et al. (2006) and were adapted for the purposes of Karakolidis et al.'s (2021) study.

of the platform used for the purposes of Karakolidis et al.'s (2021) research.

VIII. Consider the financial costs

Creating animated videos is a particularly challenging process that requires a lot of time, effort and financial resources. This is probably the main reason why, initially, test developers who were concerned about the limitations of text-based tests decided to use acted videos and not animations. More recently, however, technologically advanced software that includes a range of ready-to-use characters, movements, facial expressions, environments, and objects has made the animation process much easier and much more affordable. Although the development of simple acted videos, which do not include many characters and scenes, can be generally quicker and less expensive than producing an equivalent animation, as the scenario becomes more complex, relying on multiple locations, characters, and events, it may take significantly longer to be captured when compared to an animated video. Therefore, shooting acted videos can ultimately cost more than creating equivalent animations (Hawkes, 2013).

The costs involved in developing even relatively simple 2D animated videos can be high if the work is done by professionals. For example, the price offers for animating 15 SJT scenarios using 2D technology for Karakolidis et al.'s research ranged from €7,000 to

Figure 6. Examples of the platform environment

€45,000 in 2017; each animated scenario was one-minute long, on average. This large range of prices was due to the fact that some companies offered to design new characters for the scenarios, whereas others intended to draw on characters that had been developed for other projects. Companies working with software that includes a range of ready-to-use characters and environments submitted much more affordable offers. Such software makes it possible for non-experts to create their own animations at very reasonable costs. However, it is also likely that more sophisticated work will need to be done by professional animators. Additionally, the most expensive offers came from companies that were involved with the movie industry, while more reasonable offers came from companies that had undertaken similar projects of animating scripts in the past. Generally speaking, the development of 3D animations costs significantly more than 2D animations. Last but not least, it should be highlighted that, the expertise and financial resources required for the development of a testing platform that maximizes the potential of animated videos should not be underestimated.

Conclusion

The use of animated videos in assessment has the potential to enhance the validity of the inferences drawn from the scores of certain groups, while at the same time significantly improving the experience of test-takers.[3] These are important justifications for why many organizations now consider animations when seeking to improve existing assessments or to develop innovative new ones. However, developing an animated assessment is a particularly painstaking and expensive process and the issue of value added must be at the forefront of considerations when deciding if an investment in the technology is worthwhile. In addition, there are numerous other important decisions that need to be made throughout the process once the decision is made to proceed with development.

Despite the growing interest in animated videos for assessment, there is a knowledge gap in relation to some of the practicalities involved in developing them. In this paper we have set out to provide some guidance for others based on what we have learned over the course of a three-year project to animate a situational judgement test of teachers' practical knowledge. Specifically, we have highlighted a number of considerations related to the development of animations themselves, the platform used to administer them in a test, and the financial costs involved. Our paper was written to be useful to assessment practitioners while at the same time to add something worthwhile to the literature. In time, our hope is that others will build on this work by adding their insights gained from their own unique experiences of different assessment projects across education and the workplace

References

Abedi, J. (2010). Linguistic factors in the assessment of English language learners. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE Handbook of Measurement* (pp. 129–150). SAGE Publications Ltd. https://doi.org/10.4135/9781446268230.n8

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Bardach, L., Rushby, J. V., Kim, L. E., & Klassen, R. M. (2020). Using video- and text-based situational judgement tests for teacher selection: A quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. *European Journal of Work and Organizational Psychology*, *30*(2), 251–264. https://doi.org/10.1080/1359432X.2020.1736619

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 17–66). Springer.

Bruk-Lee, V., Lanz, J., Drew, E. N., Coughlin, C., Levine, P., Tuzinski, K., & Wrenn, K. (2016). Examining applicant reactions to different media types in character-based simulations for employee selection. *International Journal of Selection and Assessment*, *24*(1), 77–91. https://doi.org/10.1111/ijsa.12132

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, *82*(1), 143–159. https://doi.org/10.1037/0021-9010.82.1.143

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational Judgement Tests: Constructs assessed and a meta-analysis of their criterion related reliabilities. *Personnel Psychology*, *63*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Dancy, M. H., & Beichner, R. (2006). Impact of animation on assessment of conceptual understanding in physics. *Physical Review Special Topics - Physics Education Research*, *2*(1), 1–7. https://doi.org/10.1103/PhysRevSTPER.2.010104

Dill, V., Flach, L. M., Hocevar, R., Lykawka, C., Musse, S. R., & Pinho, M. S. (2012). Evaluation of the Uncanny Valley in CG characters. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents* (pp. 511–513). Springer. https://doi.org/10.1007/978-3-642-33197-8_62

Frey, B. B. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE

[3] It must be acknowledged that animations may not be suitable for all test takers, especially those with visual impairments. In this case, other accommodations and modifications to SJT assessments may be more valuable (e.g., screen readers and large print).

Publications, Inc. https://doi.org/10.4135/9781506326139.n143

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, *15*(4), 313–331. https://doi.org/10.1016/j.learninstruc.2005.07.001

Griffin, P., & Care, E. (2015). The ATC21S method. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach* (pp. 3–33). Springer.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and school. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 1–15). Springer Berlin Heidelberg.

Hanson, D. (2005). Expanding the aesthetic possibilities for humanoid robots. *IEEE-RAS International Conference on Humanoid Robots*.

Hanson, D. (2006). Exploring the aesthetic range for humanoid robots. *The ICCS/CogSci-2006 Symposium: Toward Social Mechanisms of Android Science*.

Hawkes, B. (2012a). Multimedia situational judgment tests: Are animation and live action really equivalent? *The Annual Meeting of the Society for Industrial and Organizational Psychology*.

Hawkes, B. (2012b). Test-takers' empathy for animated humans in SJTs. *The Annual Meeting of the Society for Industrial and Organizational Psychology*.

Hawkes, B. (2013). Simulation technologies. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for Personnel Selection* (pp. 61–82). Springer.

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem solving measures in the Programme for International Student Assessment (PISA). In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative Assessment of Collaboration* (pp. 95–111). Springer. https://doi.org/10.1007/978-3-319-33261-1

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of Situational Judgment Items. *European Journal of Psychological Assessment*, *22*(3), 168–176. https://doi.org/10.1027/1015-5759.22.3.168

Karakolidis, A., O'Leary, M., & Scully, D. (2021). Animated videos in assessment: Comparing validity evidence from and test-takers' reactions to an animated and a text-based situational judgment test. *International Journal of Testing*. https://doi.org/10.1080/15305058.2021.1916505

Keller, L., Keller, R., & Nering, M. (2017). *Cross-Cultural Analysis of Image-based Assessments: Emerging Research and Opportunities*. IGI Global.

Koumi, J. (2006). *Designing Video and Multimedia for Open and Flexible Learning*. Routledge.

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *The Oxford Handbook of Personnel Assessment and Selection* (pp. 383–410). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.001.0001

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *The Journal of Applied Psychology*, *91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

Lingler, J. H., Schmidt, K. L., Gentry, A. L., Hu, L., & Terhorst, L. A. (2014). A new measure of research participant burden. *Journal of Empirical Research on Human Research Ethics*, *9*(4), 46–49. https://doi.org/10.1177/1556264614545037

MacDorman, K. F. (2006). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the Uncanny Valley. *ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science*.

MacDorman, K. F., Coram, J. A., Ho, C., & Patel, H. (2010). Gender differences in the impact of presentational factors in human character animation on decisions in ethical dilemmas. *Presence*, *19*(3), 213–229. https://doi.org/10.1162/pres.19.3.213

MacDorman, K. F., Green, R. D., Ho, C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior*, *25*(3), 695–710. https://doi.org/10.1016/j.chb.2008.12.026

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

Mayer, R. E., & Pilegar, C. (2014). Principles for managing essential processing in multimedia learning: Segmenting, pre-training, and modality principles. In R. E. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning* (pp. 316–344). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.016

Mislevy, R. J. (2011). *Evidence-Centred Design for Simulation-Based Assessment (CRESST Report 800)*. University of

California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Misselhorn, C. (2009). Empathy with inanimate objects and the Uncanny Valley. *Minds and Machines*, *19*(3), 345–359. https://doi.org/10.1007/s11023-009-9158-2

Mitchell, W. J., Szerszen, Sr, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A mismatch in the human realism of face and voice produces an Uncanny Valley. *I-Perception*, *2*(1), 10–12. https://doi.org/10.1068/i0415

Mori, M. (1970). The Uncanny Valley. *Energy*, *7*(4), 33–35.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education*, *53*(2), 160–175. https://doi.org/10.1111/ejed.12271

Popp, E. C., Tuzinski, K., & Fetzer, M. (2015). Actor or avatar? Considerations in selecting appropriate formats for assessment content. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 79–103). Routledge. https://doi.org/10.4324/9781315871493

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, *85*(6), 880–887. https://doi.org/10.1037//0021-9010.85.6.880

Riggio, R. (2014). The "hard" science of studying and developing leader "soft" skills. In R. Riggio & S. Tan (Eds.), *Leader Interpersonal and Influence Skills: The Soft Skills of Leadership* (pp. 1–8). Routledge.

Schneider, E., Wang, Y., & Yang, S. (2007). Exploring the Uncanny Valley with Japanese video game characters. *DiGRA 2007 Conference*, 546–549.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*, 1–13. https://doi.org/10.7275/swgt-rj52

Stemler, S. E., Elliott, J. G., Grigorenko, E. L., & Sternberg, R. J. (2006). There's more to teaching than instruction: Seven strategies for dealing with the practical side of teaching. *Educational Studies*, *32*(1), 101–118. https://doi.org/10.1080/03055690500416074

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. Springer.

Tinwell, A. (2014). *The Uncanny Valley in Games and Animation*. A K Peters/CRC Press. https://doi.org/10.1201/b17830

Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior*, *27*(2), 741–749. https://doi.org/10.1016/j.chb.2010.10.018

Tinwell, A., Grimshaw, M., & Williams, A. (2010). Uncanny behaviour in survival horror games. *Journal of Gaming & Virtual Worlds*, *2*(1), 3–25. https://doi.org/10.1386/jgvw.2.1.3_1

UNESCO. (2014). *Global Citizenship Education: Preparing Learners for the Challenges of the 21st Century*. UNESCO

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0

Vandeweyer, M. (2016). *Soft skills for the future*. OECD: Skills and Work. https://oecdskillsandwork.wordpress.com/2016/06/17/soft-skills-for-the-future/

White, G., McKay, L., & Pollick, F. (2007). Motion and the uncanny valley. *Journal of Vision*, *7*(9), 477. https://doi.org/10.1167/7.9.477

World Economic Forum. (2015). *New Vision for Education: Unlocking the Potential of Technology*. World Economic Forum.

Wouters, P., Paas, F., & van Merrienboer, J. J. G. (2008). How to optimize learning from animated models: A review of guidelines based on cognitive load. *Review of Educational Research*, *78*(3), 645–675. https://doi.org/10.3102/0034654308320320

.

Citation:
Karakolidis, A., Scully, D., & O'Leary, M. (2021). Eight issues to consider when developing animated videos for the assessment of complex constructs. *Practical Assessment, Research & Evaluation*, 26(15). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/15/

Corresponding Author

Anastasios Karakolidis
Educational Research Centre, Ireland

email: anastasios.karakolidis [at] erc.ie