Animated Videos in Assessment:

A Comparison Study of Validity Evidence from and Test-Takers' Reactions to an Animated and a Text-Based Version of a Situational Judgment Test

Anastasios Karakolidis, MSc

Thesis submitted to Dublin City University for the degree of Doctor of Philosophy

Supervisors:

Prof Michael O'Leary & Dr Darina Scully

Institute of Education

School of Policy and Practice

September 2019

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Anastasios Karakolidis

ID No.: 16212656

Date: 13th June 2019

Table of Contents

List of Tables

List of Figures

List of Acronyms and Abbreviations

AERA	American Educational Research Association
APA	American Psychological Association
DIF	differential item functioning
EFA	exploratory factor analysis
ETS	Educational Testing Service
GPA	grade point average
GRE	graduate record examinations
M	mean
NCME	National Council on Measurement in Education
OSCE	objective structured clinical examination
PK-SJT	situational judgment test of practical knowledge
SD	standard deviation
SE	standard error
SJT	situational judgment test

Acknowledgements

This work would not have been completed without the support of many people, to whom I offer my sincere thanks. I would like to start with my supervisors who helped me a lot with their insightful feedback and guidance. Michael O'Leary was so generous with me and my work. He gave me the opportunity to come to Ireland and start the most amazing journey of my life so far. A journey with many long days, late nights, busy weekends, but above all, a journey that allowed me to learn and expand my horizons, surrounded by beautiful people. Michael has managed to create not only a research centre, but a family, and I am proud to be a member of this family.

Darina Scully was my secondary supervisor (on paper). I am saying "on paper" because, in reality, her role was not secondary at all. She was always there for me; I mean literally beside me, as we shared the same office. I am the only person I know who shared an office with his supervisor and I have to say that it was fun and constructive. She was there to address all my weird questions, even in periods when she was extremely busy. I cannot describe how helpful she was even when she had to take care of someone much more important than this thesis, her new-born son.

I would like to acknowledge the contribution of my panel member, Zita Lysaght, who always gave me her thoughtful advice. Also, I am very grateful to Anastassios Emvalotis, who was one of the most incredible tutors during my bachelor studies in Greece. Since then, he has been a wonderful mentor. His knowledge, ethos, and continued support and encouragement keep me motivated to do purposeful and robust research.

I would like to thank my colleagues and friends from Dublin City University, National College of Ireland, and the University of Ioannina, Greece, who provided me with valuable support at different stages of my research: Ashling Bourke, Nicola Broderick, Deirdre Butler, Grainne Kent, Athina Koutsianou, Paula Lehane, James Lovatt, Kay Maunsell, Thomas McCloughlin, Paula Murphy, Bernadette Ní Áingléis, Sylvaine Ni Aogain, Caitriona Ni Cassaithe, Pia O'Farrell, Ellen Reynor, Katerina Sargioti, Joe Usher, and Peter Whelan.

Also, I would like to thank Steven Stemler, who was so generous and provided me with the original practical knowledge items that he developed along with his colleagues. Both the text-based and the animated test used in this study were adjusted version of these items. Additionally, I very much appreciate the time and effort all participants put into this study. I would have nothing to talk about if I did not have the support of these volunteers.

Yvonne Crotty, Therese Hopfenbeck, and Margaret Leahy were the examiners for the purposes of my progress and final viva voce examinations. Their feedback was very constructive and incredibly helpful, having a positive impact on the final thesis.

I would especially like to acknowledge the role played by Prometric during my studies over the past three years. The company's financial support made it possible for me to undertake the doctorate on a full-time basis. Thank you, Michael Brannick, Charles Kernan, Linda Waters, Garrett Sherry, Steve Williams, and Ian Clifford for all your varied support. I will always be grateful to Prometric for the wonderful opportunity the company created for me.

I would also like to formally thank my family: my parents, Konstantia and Nikos, my brothers, George and Stavros, my sister-in-law, Yiota, and my two little nephews, Nicolas and Dimitris. Being away from them was hard, but it motivated me to achieve this goal and make them proud. I love you all!

Last but not least, I would like to thank my partner, Vasiliki. She did absolutely everything to support me in accomplishing this goal, and I owe her a lot. I would not have been able to do any of this if I did not have her by my side. Her support, patience, love, and passion continuously inspire me and remind me of what is important in life.

Abstract

Animated Videos in Assessment:
A Comparison Study of Validity Evidence from and Test-Takers' Reactions to an Animated and a Text-Based Version of a Situational Judgment Test

Anastasios Karakolidis, MSc

The majority of tests in use today rely on static text to communicate information, ideas, and concepts and to pose questions. However, the overuse of text may have consequences for the validity of the inferences drawn from test-takers' scores. This may be true especially in the case of assessments taken by test-takers with poor reading comprehension skills or with low levels of proficiency in the language of the test. More specifically, linguistic complexity can be a source of construct-irrelevant variance as test-takers' performance can be negatively affected by factors that are beyond the focus of the assessment itself.

This study examined the extent to which the use of animated videos, as opposed to static text, can (i) reduce construct-irrelevant variance in test scores and (ii) have a positive impact on test-takers' reactions to the test. A true experiment was conducted with 129 native and non-native English speakers, using an animated-video and a text-based version of the same situational judgment test of practical knowledge.

The results indicated that, overall, the variance attributed to construct-irrelevant factors (i.e., native language, English proficiency, and reading comprehension in English) was lower by 9.4% in the animated-video versus the text-based version of the test. In addition, the animated-video test was perceived by participants to be more valid, fair and enjoyable. Study participants also found the language used in the animated-video test less difficult to process, but no significant differences between the two formats were found with respect to the perceived difficulty of the content. Finally, the use of animated videos did not result in participants investing a significantly greater effort in the test. The implications of these and other findings, as well as recommendations for policy, practice, and future research are discussed in the final chapter of the thesis.

Chapter 1: Introduction

1.1. Research Topic and Problem

This thesis explored the use of animated videos in assessment[1]. More specifically, it examined the extent to which animated videos are a useful alternative to text for testing[2] complex knowledge and skills. The majority of tests in use today rely heavily on static text. As outlined by Bialik, Martin, Mayo, and Trilling (2016), the most commonly used item format in large-scale assessments is the multiple-choice question, a format that relies predominantly on written text. In most instances, text is still the principal mode of presenting the stimuli and responses for both selection- (e.g., multiple-choice) and constructed-response tests (e.g., open-ended questions). In other words, test-takers usually receive all the information they need in the form of written text, and they provide their responses either by selecting a possible response option or by writing their response. Despite the ubiquity of text, it is not difficult to list some of the critical limitations inherent in text-based assessments.

- Tests that rely heavily on text require a threshold level of reading skills some test-takers may lack.
- Text-based tests are limited in terms of the complexity of what can be presented as stimuli and/or responses, especially in the context of highly sophisticated knowledge and skills.
- Long passages of text on electronic screens may present the reader with navigation and comprehension-related challenges.

[1] In this thesis, *assessment* is defined as the process of collecting, synthesising, interpreting and using data to answer questions, solve problems or facilitate decision making (Russell & Airasian, 2012).

[2] *Testing* constitutes an important aspect of assessment, but it is only one among the many different tools employed to collect information (R. J. Cohen & Swerdlik, 2009). More specifically, testing is a formal and systematic procedure, which involves the administration of a set of questions with the aim of obtaining some measure, usually numerical in nature, about individuals' or groups' performance (Russell & Airasian, 2012).

In order to deal with the provided written items[3] and fully comprehend the content of the given questions in text-based assessments, examinees are required to be able to read, comprehend and interpret texts, which in some cases can be very complex. This implies that test-takers should also have skills, such as verbal, reading comprehension and interpretation skills, that, in many cases, are irrelevant to the construct[4] that a given assessment purports to capture (Popp, Tuzinski, & Fetzer, 2016). Therefore, the extensive use of text in assessments might prove problematic, especially for tests that are intended to be administered to test-takers with varying levels of reading comprehension or poor proficiency in the language of the test (Kan, Bulut, & Cormier, 2018).

According to Abedi (2010), linguistic complexity may be a source of *construct-irrelevant variance* for such groups (e.g., non-native speakers), in that their performance can be negatively affected by a factor that is beyond the focus of the assessment they are taking. For example, if non-native speakers consistently perform more poorly on a science test, not because of their lack of scientific knowledge but due to their difficulty in understanding the language used in the items, then this suggests that the text has introduced test score variance that is irrelevant to the construct that the test purports to measure (i.e., knowledge in science). This is something that, according to the latest *Standards for Educational and Psychological Testing*, constitutes a significant threat to *validity* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

This problem linked to the use of text in assessment is exacerbated when it comes to the measurement of more sophisticated or higher-order skills, such as problem-solving, communication or practical knowledge. As Popp et al. (2016) argued, written language

[3] In the field of assessment, the term *item* is used to describe a specific task test-takers are asked to perform (Center for Advanced Research on Language Acquisition [CARLA], 2014). Each item includes a stem, which is typically the beginning part of the item and presents a problem, question or incomplete statement that the test taker has to solve, answer or complete. Apart from a stem, some types of items provide test-takers with a list of response options from which they should choose (i.e., selected-response items) (Russell & Airasian, 2012). In contrast to selected-response items, in constructed-response items, test-takers have to supply the response.

[4] In measurement, *constructs* are non-directly observable traits that are used to describe or explain behaviour (R. J. Cohen & Swerdlik, 2009). Anxiety, mathematics achievement, motivation, cognitive ability, job satisfaction and creativity are some examples of constructs. In the world of assessment, construct refers to a concept or characteristic that a test is designed to measure (AERA et al., 2014).

is often a good fit for measuring simple knowledge-based constructs that can be clearly communicated via text, such as knowledge of historical events. Nevertheless, when a test provides a great deal of sophisticated information to measure skills that are at a higher complexity level, text may not be suitable for facilitating this process (Popp et al., 2016).

As argued by Scully (2017), assessment of skills that go beyond the simple recall of knowledge usually involves test items that provide test-takers with complicated stimuli. This typically requires the use of longer, more complex pieces of text with difficult vocabulary and sentence structure. Therefore, the reading demands of tests that aim to assess complex skills using text can be particularly heavy. Examples of such tests are situational judgment tests (SJTs), where test-takers are provided with descriptions of challenging real-life situations as well as a number of possible alternative ways to deal with the given problem (Motowidlo, Dunnette, & Carter, 1990). These tests typically purport to assess complex knowledge and skills, such as leadership, interpersonal skills or emotional intelligence (Christian, Edwards, & Bradley, 2010). However, the "noise" caused by their heavy reading demands is likely to render them unsuitable for groups of people who may have the skills that the test intends to assess, but lack the required language fluency or reading comprehension levels (Popp et al., 2016). In other words, sophisticated text-based tests may simply favour participants with very good reading skills.

Another issue when it comes to the use of long passages of text in assessment has to do with the high demand for computerised tests, especially during the last number of years. In recent years, technological developments have affected the way that tests are administered, transforming testing from a paper-and-pencil to an automated, computerised process. More and more tests are administered via computers and other devices. For instance, in the Programme for International Student Assessment (PISA) 2015, for the first time, the primary mode of assessment was computer-based, with most countries opting for administering the entire survey via computers (Organisation for Economic Co-Operation and Development [OECD], 2017).

Computer-administered testing provides many practical advantages over paper-and-pencil testing, such as ease of administration and scoring, but there is some controversy over the potential impact that the administration mode (i.e., paper-and-pencil vs.

computer) may have on test-takers' performance (e.g., Jerrim, 2016; Kingston, 2008; Sangmeister, 2017). An aspect of this debate has to do with whether or not the text used in paper-and-pencil tests translates seamlessly to digital environments. Research evidence indicates that people tend to comprehend texts more easily when they read them on paper, as opposed to computer screens (Ackerman & Lauterman, 2012; Liu, 2005). This suggests that test-takers may have greater difficulty in comprehending text-based items, especially the more complex ones, when those are administered through computers. As a consequence, the aforementioned issues linked to the use of text in testing are likely to be exacerbated in computer-based administrations.

1.2. Significance of the Study

In the modern world, nothing is static. Individuals, companies and education systems are part of an ever-changing global community. Global migration has not left the fields of education and assessment unaffected (Spring, 2009). More and more people decide to study or work abroad, and, as assessment has become an integral part of everyday life, they often have to take tests in a language other than their mother tongue. The testing community is responsible for offering fair, valid and reliable test solutions. However, one of the main problems in text-based assessment relates to the fact that written text can introduce construct-irrelevant variance, especially for test-takers with weaker language skills (e.g., non-native speakers); this constitutes an important threat to validity (Abedi, 2010; AERA et al., 2014).

Generally, it seems that written descriptions in text-based tests are not necessarily the best way of communicating information to assess complicated skills. What makes this issue even more important nowadays is that the world is increasingly moving towards teaching, learning and assessment of more complex knowledge and skills that promote social transformation, such as critical thinking and problem-solving (United Nations Educational Scientific and Cultural Organization [UNESCO], 2014). As Griffin and Care (2015b) argued, traditional forms of assessment may not be suited to the measurement of many of these skills.

In recent decades, there has been much discussion about the types of knowledge and skills people should have to succeed in their personal, social and professional lives. Factual knowledge, which was what employers asked for in the previous century, has

lost its old glory as, nowadays, it can be instantly accessible by simply typing some keywords on any online search engine (He, von Davier, Greiff, Steinhauer, & Borysewicz, 2017). Instead, other skills are at the core of modern societies. The so-called *21st century skills* refer to a series of skills needed to navigate the 21st century (Griffin, Care, & McGaw, 2012), or in other words, a toolkit of knowledge and competencies that modern people should have to succeed at school or in their job, overcome their difficulties and solve personal and social problems. Problem-solving, critical thinking, creativity, communication and collaboration are just some examples of 21st century skills (Binkley et al., 2012).

The 21st century skills are a combination of cognitive and non-cognitive *soft skills*, with the latter attracting particular attention over the last number of years (Vandeweyer, 2016). Soft skills are associated with how people get along with each other, communicate, and work in teams, having a strong social dimension (Riggio, 2014), and as Griffin and Care (2015a) suggested, such skills will be even more critical for individuals as the world changes and becomes more complex. Some even argue that employers ask for soft skills far more often than for any other technical skills (Selingo, 2016).

In the context of the growing interest in 21st century skills, the World Economic Forum (2015) highlighted the need for more effective measurement of these skills. Policymakers are directing education systems towards assessments that can efficiently measure 21st century skills and that create incentives for these skills to be widely taught as a regular part of the curriculum (Adamson & Darling-Hammond, 2015). The field of assessment is, hence, challenged to adequately measure complicated knowledge, skills and competencies; and it seems that the reliance on text-based approaches with heavy reading demands is unlikely to facilitate the achievement of this goal. This is one of the principal reasons why this study aimed to inform policy and practice about the potential benefit of using animated videos, as an alternative to text, to facilitate the assessment of complex constructs for a range of test-taking groups.

1.3. The Power of Technology and Animations to Leverage Change in Assessment

Technology has greatly influenced many aspects of our lives. Among others, it has changed the way people communicate, work, learn and entertain themselves. Aiming to qualify young people with the higher-order skills needed in 21st century societies, education systems place much emphasis on technology and how it can facilitate learning (Trilling & Fadel, 2009). The importance of technology for learning is evidenced by many research studies undertaken across different age groups, countries and education systems that acknowledged the positive impact that technology-enriched educational approaches can have on students' performance in various disciplines such as geography, mathematics and language (Cheung & Slavin, 2013; Pitsia, Karakolidis, & Emvalotis, 2016; Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011). In relation to animations, more specifically, there is a number of research studies supporting their positive effects on learning outcomes. For instance, a meta-analysis supported the advantages of animations over static pictures for students' learning in various school subjects (Höffler & Leutner, 2007).

Assessment, either as an integral element of learning, namely assessment *for* learning[5] (Gardner, 2012) or as a process of evaluating the final learning outcomes, namely assessment *of* learning[6] (Russell & Airasian, 2012), has been also influenced by the technological developments of the last decades (Darling-Hammond, 2014). Technology can be used to facilitate assessment in different stages, such as test development, administration, item presentation and scoring, contributing to more reliable and fairer results (Norcini, Lipner, & Grosso, 2013).

As a first step in taking advantage of technology in testing, many institutes and organisations, not only in the field of education but also in certification, licensure and personnel selection, have moved from paper-and-pencil to computerised tests in order to facilitate the delivery and scoring of tests. However, as Ward (2002, p. 38) noted, "*this is a huge step, but only the first of several*". Indeed, technology is most frequently used simply to facilitate assessment administration, scoring and reporting, despite the

[5] Also known as formative assessment.
[6] Also known as summative assessment.

fact that a computer can be much more than just a device for delivering the questions from a paper-and-pencil test on a screen (Quellmalz & Silbertglitt, 2018).

Technology has the potential of greatly improving assessment by increasing the quantity and quality of the stimulus materials that test-takers are presented with, from simple text to audio support, still images, animated videos and highly complex interactive virtual environments (Agard & von Davier, 2018; Fetzer & Tuzinski, 2013). However, it is only in recent years that attempts have been made to integrate technology into assessment procedures, with international studies, such as PISA, paving the way for more complex testing techniques that are not dependent exclusively on verbal descriptions (OECD, 2016). In the context of this growing usage, there is an urgent need for research that evaluates the quality of technology-enabled test items (e.g., test items that use animated videos)[7].

In response to this need, there is a growing discussion around the use of animated videos in replacing written descriptions in tests (Tuzinski, 2013). Animation of complex representations, such as human interactions, natural phenomena or the function of body organs, can provide a high-fidelity (i.e., true-to-life) illustration of reality, something that, in many cases, cannot be achieved through oral or written descriptions. However, their use in testing has not been common. Despite their limited use, animations have the potential to bring significant benefits in assessment. As argued in the literature:

- Animations can present test-takers with a more realistic picture of a given situation (Bruk-Lee, Drew, & Hawkes, 2013).
- The depictive nature of animations, virtual objects, images and sound enables the visualisation of complex concepts and information that cannot be easily communicated via text (Popp et al., 2016).
- Animations can be manipulated and shaped to more accurately illustrate specific features of a situation, object or task (Hegarty, 2004).
- Test-takers find animated tests[8] to be more valid, engaging and enjoyable compared to text-based assessments (Bruk-Lee et al., 2016).

[7] *Technology-enabled items* are those that contain media that cannot be presented on paper.
[8] The term *animated tests* refer to tests that use animated videos to replace written text.

1.4. Origins of and Personal Rationale for the Study

The idea for this research originally came from the field of credentialing assessment. As Prometric, the funders of the research, highlighted, one of the main challenges facing credentialing testing organisations is the need to ensure that candidates' language and reading skills do not interfere with their test performance. This issue becomes more critical when large groups of candidates with low levels of education or limited proficiency in the language of the test are assessed. Such candidates may struggle to demonstrate the knowledge and skills that the assessment purports to measure, not because of a lack of competence, but because of difficulty in fully comprehending the language and, consequently, the content of the test items. Building on this idea, the original research plan was to examine the effectiveness of animated videos using testing instruments and data from actual certification and licensure candidates. However, this research plan proved too challenging to implement due to administration restrictions, data protection issues, and the limited number and complexity of the available animated items. Despite the initial difficulties, the researcher decided to persist with this idea, but in an alternative context (i.e., education rather than licensure and certification). He decided to examine the effectiveness of animated videos by conducting a research study in the field of educational assessment, which was his area of expertise. Being in charge of all the different aspects of the research design gave the researcher more flexibility in the implementation of the study and helped him to avoid the above-mentioned challenges.

1.5. Scope of this Study

This research study examined the use of animated videos as an alternative to written descriptions in the testing of complex knowledge and skills. An SJT that measures teachers' practical knowledge was selected as the "vehicle" for examining the effectiveness of the animations. SJTs usually present complex, realistic situations and require respondents to select the most appropriate response option or rate the suitability of the provided options (Motowidlo et al., 1990). The SJT that was used in this study focused on the measurement of pre-service primary teachers' practical knowledge of how to handle challenging social situations where students, parents, principals or other teachers are involved. The original SJT items were developed by Stemler, Elliott, Grigorenko, and Sternberg (2006) and were adapted for the purposes of this study. More

information about the instrument is provided in the Methodology Chapter of this thesis (Chapter 3).

The aim of this study was to ascertain the extent to which animations can contribute to the assessment of sophisticated knowledge and skills (i.e., practical knowledge) over and above a conventional text-based test. Therefore, a text-based test of practical knowledge and a parallel animated version of the same test were compared. In the text-based test, the SJT scenarios were presented through text, while in the animated version of the test, they were presented through animated videos. The impact of animations was explored with reference to two particular aspects of the test-taking process:

- unintended subgroup differences and construct-irrelevant variance in test scores and
- test-takers' reactions to the test.

With regard to construct-irrelevant variance, this research explored the impact of animated videos on test-taker performance across groups with different levels of proficiency and reading skills in English (i.e., native and non-native English speakers). In other words, this research aimed to examine whether animations can improve validity by allowing participants with lower language and reading skills in English to achieve their potential in the test. Regarding test-takers' reactions, on the other hand, data were collected pertaining to participants' perceptions of the validity, fairness and difficulty of the animated versus the text-based assessment, and on the extent to which test-takers enjoyed the test and invested effort in it. Figure 1.1 depicts the main topic and the focus of the thesis.

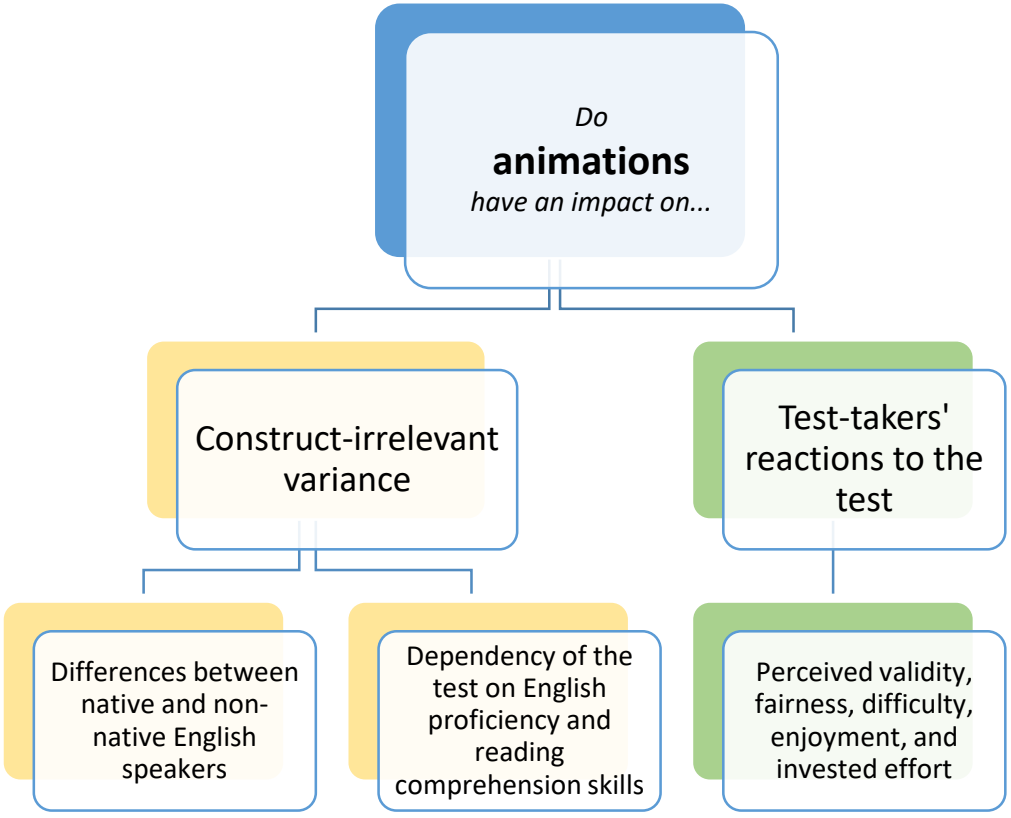


Figure 1.1. The focus of this research.

Another important aspect of this study was the documentation of the processes involved in the development of the animated version of the test. Taking into consideration that this was one of the first attempts to replace written descriptions in a test with animated ones, the documentation of the animation process aimed to provide (i) the rationale for the most important decisions made around different aspects and features of the animations and (ii) information about the steps followed and the challenges encountered in such a demanding venture. The main information regarding the animation of the text-based version of the test is presented in the Methodology (Chapter 3), while a detailed documentation of the process can be found in Appendix A.

It should be acknowledged early on that this study sought to address a series of very specific research questions related to the use of animated videos as an alternative to text in testing (the research questions are explicitly discussed in Chapter 2). As is the case in any research study, the scope was limited to these questions alone, and the study should not be viewed as an attempt to address other possible questions surrounding the use of animated technology in assessment. For example, the study did **not** focus on interactive virtual environments, where examinees are asked to perform a task in a simulated

setting. Rather, the discussion was focused on non-interactive animated videos that can also have a text-based parallel version. Thereby, the study aimed to explore the aspects of a text-based test that can be improved via the use of animations. Furthermore, it should be noted that the practical knowledge SJT was used in this study merely as a "vehicle" for facilitating the examination of the research problem, which relates to the advantages of animated over text-based descriptions in testing. This study should **not** be seen as an attempt to conceptualise and measure practical knowledge, as this has been already done by the developers of the instrument.

1.6. Organisation of the Thesis

This introduction constitutes the first of the five chapters of this thesis. The next chapter focuses on literature and previous research in the area of video-based testing approaches. It provides an in-depth discussion of the difficulties resulting from the extensive use of text in testing, presents the alternatives to long written descriptions in tests, and critically examines the relevant research literature around the use of video-based media as an alternative to text in assessment, thereby informing the research questions. This is followed by Chapter 3 in which the methods employed to address the research questions and the main decisions regarding the development of the animated test are outlined. Chapter 4 presents the research findings, and finally, Chapter 5 offers a critical review of these findings and provides recommendations for policy, practice and future research.

Chapter 2: Literature Review

2.1. Introduction

This chapter explores the literature and previous research on alternatives to text-based assessment, with reference to the limitations of text-based approaches and the benefits that visual representations, such as animations, can bring to assessment. Its main purpose is to critically discuss the most relevant literature that led to the formulation of the research questions of this study. The chapter begins with an in-depth identification of the problem, namely the issues in text-based testing. Then, the alternatives to written descriptions are presented, with a particular emphasis on the different types of video-based approaches in testing, including animated videos. Following this, the most relevant research evidence in the field of video-based testing is discussed, with emphasis on three aspects: (i) test-takers' performance, (ii) validity, and (iii) test-takers' reactions to the test. The chapter ends with a summary of the main research findings of the literature, a discussion of the gap that this study aimed to address and an outline of the research questions.

Due to the limited relevant research literature in the field of educational assessment, studies from other areas, and mainly from the personnel selection literature, were also reviewed. Despite the fact that the literature search was expanded beyond the field of education, the relevant research was still very limited and, therefore, the studies presented in this literature review chapter were selected solely based on their relevance to the research topic and further quality criteria were not applied.. *Google Scholar* and *Scopus* were the main search engines used to source the relevant academic literature. Once the main articles were located, further relevant papers were found through the citations in these articles and/or by searching for more research by the same authors.

2.2. Validity Issues in Text-Based Assessment

2.2.1. The use of text in testing

A testing process can take different formats in terms of the types of items used (e.g., open-ended, selected-response), the way that the information in each item is presented to the test-takers (e.g., text, images, audio, video), and the way that these items are

administered (e.g., paper-and-pencil, computer). Regardless of item type and administration mode, written text is the most common way of communicating information in test items. Usually, simple selected-response items (e.g., multiple-choice and true/false) solely use text to present both the stem and the response options, and this seems to be the case even for more complex item types, such as situational judgment test (SJT) items. For instance, Christian et al. (2010), in their meta-analysis, found that the vast majority of SJTs did not use any kind of multimedia and that they tended to be exclusively based on written descriptions.

As mentioned in the previous chapter, text is considered to be a suitable format for assessing constructs related to reading ability or recall knowledge, which are not very complex and can be easily communicated in written language. However, when it comes to the presentation of sophisticated information for the measurement of complicated skills and knowledge, the use of text may be problematic (Popp et al., 2016). In text-based assessments, participants must meet the reading demands of the test in order to be able to fully comprehend the items and demonstrate their abilities. Indeed, such demands may be very high in some instances. This implies that apart from the construct that the test intends to measure, test-takers are also required to have skills (e.g., verbal, reading, comprehension and interpretation skills) that might be irrelevant to this construct (Popp et al., 2016).

This can be a substantial problem for test-takers with lower levels of reading comprehension, or poor proficiency in the language of the test, such as non-native speakers. Research evidence has indicated that there is a major performance gap between native and non-native speakers across various content areas and, particularly, in those areas with high language demands. More specifically, after reviewing the findings of the relevant research literature, Abedi (2006) concluded that the gap between native and non-native English speakers who completed the same test was much larger in verbal tests, while it decreased in mathematics, where the levels of language demands of the tests were lower. As Abedi (2006, 2010) highlighted, in many cases, such performance gaps are attributed to the use of linguistically complex passages of text in the assessments. Indeed, others also argued that abilities, such as language and reading skills, that are beyond the focus of an assessment may influence examinees' performance

and introduce adverse impact; namely, negative discrimination against particular groups of people (e.g., U.S. Office of Personnel Management, n.d.).

Research evidence has demonstrated that the amount of written text in an individual item can impact the performance of certain groups of test-takers. For example, Abedi, Lord, and Plummer's (1997) research indicated that longer test items were found to be more difficult for non-native speakers. This might suggest that the use of text can be particularly problematic in the assessment of complex skills, where, according to Scully (2017), test-takers are usually provided with extensive and linguistically complicated tests, which they have to fully comprehend before answering the questions.

In another research study conducted by Abedi (2004), the author examined whether linguistically complex, text-based tests accurately measured high school students' performance and progress in reading and mathematics. A group of about 14,000 students with limited English proficiency was followed for a period of seven semesters, from Grade 9 to Grade 12. Over time, some of these students came to be categorised as non-limited English proficiency speakers (i.e., fluent), whilst others remained in the limited English proficiency category. The performance of both of these groups in three different tests (reading, analytical mathematics, and mathematical calculations) was, then, compared to that of their native English-speaking counterparts. The results indicated that non-native students with limited English proficiency were significantly outperformed by their fluent non-native peers, with both groups performing well below their native English-speaking counterparts. The largest gap between the limited and fluent English proficiency groups was detected in the reading test (d = 0.21), which had the highest reading demands, while the smallest gap was detected in mathematical calculations (d = 0.08), which had the lowest reading demands. The effect size gap between these two groups in the analytical mathematics assessment was found to be in the middle of the other two effect sizes (d = 0.16). Based on these findings, the author concluded that students' performance even on tests that do not measure reading constructs (i.e., the analytical mathematics and mathematical test) can be affected by the reading demand of these tests. Hence, the poorer performance of students with limited English proficiency in mathematics, should not be attributed solely to their lack of mathematics knowledge, but rather to their difficulty in understanding the linguistic structure of the tests.

Ployhart and Holtz (2008) examined 16 assessment strategies for employment selection that were hypothesised to minimise any adverse impact on candidates' performance. They concluded that, in order to more accurately capture participants' competencies, the verbal and reading demands of a test should be kept to a minimum. In practice, this implies that test-takers should not be provided with complex and lengthy tests that require high levels of reading comprehension ability, especially in assessments measuring constructs that do not relate to verbal and language skills. However, this might not always be possible.

Such issues in text-based testing may be exacerbated when test items are presented via computers or other devices. Research studies that were conducted in the UK, the US, and Canada, with university students and academic staff, indicated that when someone has to read a text thoroughly, they prefer to read it from print rather than a digital display (Buzzetto-More, Sweet-Guy, & Elobaid, 2007; Jamali, Nicholas, & Rowlands, 2009). These findings imply that the extensive use of text may be even more problematic in computer-based assessments, where test-takers need to comprehend complex items by reading extended passages of text on a screen.

A number of research studies examined whether or not reading on a screen is, indeed, linked to poorer comprehension of a text. Ackerman and Lauterman (2012) and Mangen, Walgermo, and Brønnick (2013) employed samples of high school (n = 72) and university students (n = 30), respectively, to investigate whether reading on screen could impact readers' comprehension. Overall, the findings of both studies revealed that students who read the texts in print scored statistically significantly better in the comprehension test than those who read the same texts on screen, especially when they were under time pressure. Mangen et al. (2013) argued that this might be the case because factors, such as scrolling, can negatively impact reading comprehension. Ackerman and Lauterman (2012), on the other hand, argued that when facing time pressure, reading on screen can hinder the strategies applied to effectively learn and/or comprehend passages of text, something that is not the case when reading printed texts or in situations without time pressure. It should be acknowledged, though, that the above evidence refers to people's comprehension of long passages of text, and not specifically to test-takers' ability to comprehend the content of test items in non-reading tests.

There is a wide range of research studies examining the impact that test administration via computers may have on test-takers' performance. A synthesis of many of these studies was undertaken in a meta-analysis that compared K-12 students' performance in paper- and computer-administered multiple-choice questions (Kingston, 2008). Overall, the results of 81 studies revealed that the differences in test scores between students who took the paper-and-pencil tests and those who took the computer-administered version of the same tests were very small, with an effect size of $d = 0.01$ in favour of the former group. However, the effect sizes were found to fluctuate greatly from domain to domain. More specifically, computer-based tests were found to provide test-takers with a slight advantage in English Language Arts and Social Sciences ($d = 0.11$ and $d = 0.15$, respectively), while test-takers who took the paper-and-pencil tests were found to perform better in mathematics ($d = 0.06$). These findings suggest that the context in which the assessment is conducted may be a determinant factor, and thus, it is hard to draw a general conclusion about whether the problems linked to the use of text in testing are more pronounced when the test is administered via computer screens.

2.2.2. Issues linked to text-based testing as a validity problem

Validity is considered to be the most fundamental consideration in developing and evaluating tests (AERA et al., 2014). Despite the significant improvements in the field of assessment, there is no widespread consensus over the concept of validity. Newton and Shaw (2016) argued that this disagreement around validity is mainly focused on what the term should encompass and how it should be applied.

Validity has traditionally been defined as an estimate or judgment about whether a test measures what it purports to measure (Garrett, 1937). From that point of view, it is perceived as a technical characteristic that a test may or may not have. On the other hand, there are schools of thought that perceive validity from a different perspective, namely not solely as a test characteristic, but as a concept directly linked to the inferences drawn and decisions made based on test results (AERA, APA, & NCME, 1985). According to the latest *Standards for Educational and Psychological Testing*, validity is defined as *"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"* (AERA et al., 2014, p. 11).

The lack of widespread consensus over the meaning of validity in assessment has affected the way this term is conceptualised and used by scholars in empirical studies.

Indeed, many research studies in the field of technology-enabled assessment, especially earlier studies, adopted a more traditional approach in the way they present and explore validity, conceptualising it as a characteristic of the test. For the purposes of this study, validity is perceived as the extent to which evidence, as well as theory, support specific uses and interpretations of the test scores, according to the more recent *Standards* (AERA et al., 2014). Thus, whenever the term *validity* is used, this refers to *validity evidence regarding the use of a test in a specific context.*

To secure the validity of the inferences drawn from a test score, it is essential that test-takers' performance on the test is not affected by factors, such as reading ability, nationality or general knowledge, that are irrelevant to the construct the test aims to capture (AERA et al., 2014). Keeping this in mind, every effort should be made to minimise the negative impact that sources of construct-irrelevant variance may have on certain subgroups' performance. As illustrated by research evidence, when it comes to text-based tests, certain groups of test-takers performance can be significantly shaped by their reading and verbal skills. These issues in text-based assessments are likely to affect the quality of the inferences drawn from test-takers' scores about their knowledge and skills, especially for participants with less advanced language competencies. Thus, the extensive use of text in testing may constitute a critical threat to validity.

An illustrative example may clarify this argument. Consider, for instance, a situation whereby two people with the same levels of critical thinking take the same text-based critical thinking test. This test consists of a series of long passages of text that they have to read and comprehend before answering the questions. However, participant A does not have strong reading comprehension skills and this could lead them to perform less well than participant B, who has good reading comprehension skills. In other words, participant A may perform worse than participant B, not because they lack critical thinking skills but because they may have difficulty in comprehending the content of the passages of text. This implies that the inferences drawn from participant's A performance regarding their critical thinking skills would not be valid because their score would not be a clear reflection of their skills in the construct the test purports to measure.

These important validity issues can also be examined from a *fairness* point of view, considering that the two concepts (i.e., validity and fairness) are strongly linked, as

Zieky (2006) argued. One of the key challenges in assessment is to ensure that all participants are given the same opportunities to succeed and that their test performance is not affected by factors that are irrelevant to the construct that the assessment purports to measure (AERA et al., 2014). Providing all test-takers with the same opportunity to perform is a crucial aspect of a quality and fair assessment process. As the *Standards for Educational and Psychological Testing* highlight, fairness is a fundamental validity issue and requires attention throughout all the stages of test development and use (AERA et al., 2014).

If a test is designed to be administered to a broad range of test-takers, it should be sensitive to test-takers' individual characteristics, such as ethnicity, age, culture, disability, socio-economic status, in order to provide valid and indications of test-their performance. The key term linked to fairness is *accessibility* referring to the unobstructed opportunity that all participants should be given to demonstrate their standing on the construct being measured (AERA et al., 2014). For instance, a group of non-native English speakers, who take a mathematics test in English should be able to demonstrate their knowledge, abilities and skills related to mathematics, despite their potential limited English proficiency. Subgroup or individual differences that emerge from construct-irrelevant factors are indicators of lack of fairness, and consequently, undermine the validity of the inferences drawn from the test scores.

2.3. The Potential of Video Technology in Improving Validity

The technological developments of the last decades have significantly influenced assessment, improving test development, administration and scoring (Norcini et al., 2013). Additionally, through the use of technology, assessments can increase both the quantity and quality of the information that test-takers are presented with (Fetzer & Tuzinski, 2013). As argued by Boyce, Corbet, and Adler (2013), video representations seem to be an efficient way of decreasing test dependency on text, which, as outlined above, could eliminate a major source of construct-irrelevant variance and the associated adverse impact on certain groups of test-takers. This can ultimately yield improved validity.

The use of multimedia, such as videos, can also improve assessment fidelity, which refers to the degree to which an item accurately presents a situation and extracts test

takers' responses (Motowidlo et al., 1990). Test items can vary in the authenticity with which they replicate the physical and psychological fidelity of a given situation. At the bottom of the fidelity scale are tests that present a situation using written or oral formats (i.e., text-based SJTs). Tests employing audio-visual technologies to depict a situation have higher levels of fidelity. Finally, interactive assessments, whereby candidates perform tasks in live action are considered to be at the top of the fidelity scale (Tuzinski, 2013).

Some test developers have considered the use of videos in assessment in order to improve the fidelity of the test and overcome the difficulties associated with text-based assessments (e.g., Lievens & Sackett, 2006). The incorporation of videos in item stems allows for the visualisation of the situations or tasks that items aim to describe, leading to the optimal presentation of information. Generally speaking, videos can be categorised as either non-character representations or character-based interactions. Figure 2.1 presents an example of a non-character video used in a science test to represent a geographical phenomenon. In this example, the test developers replaced the complex text-based description with an illustrative animation that conveys the information that test-takers need in order to answer the question.

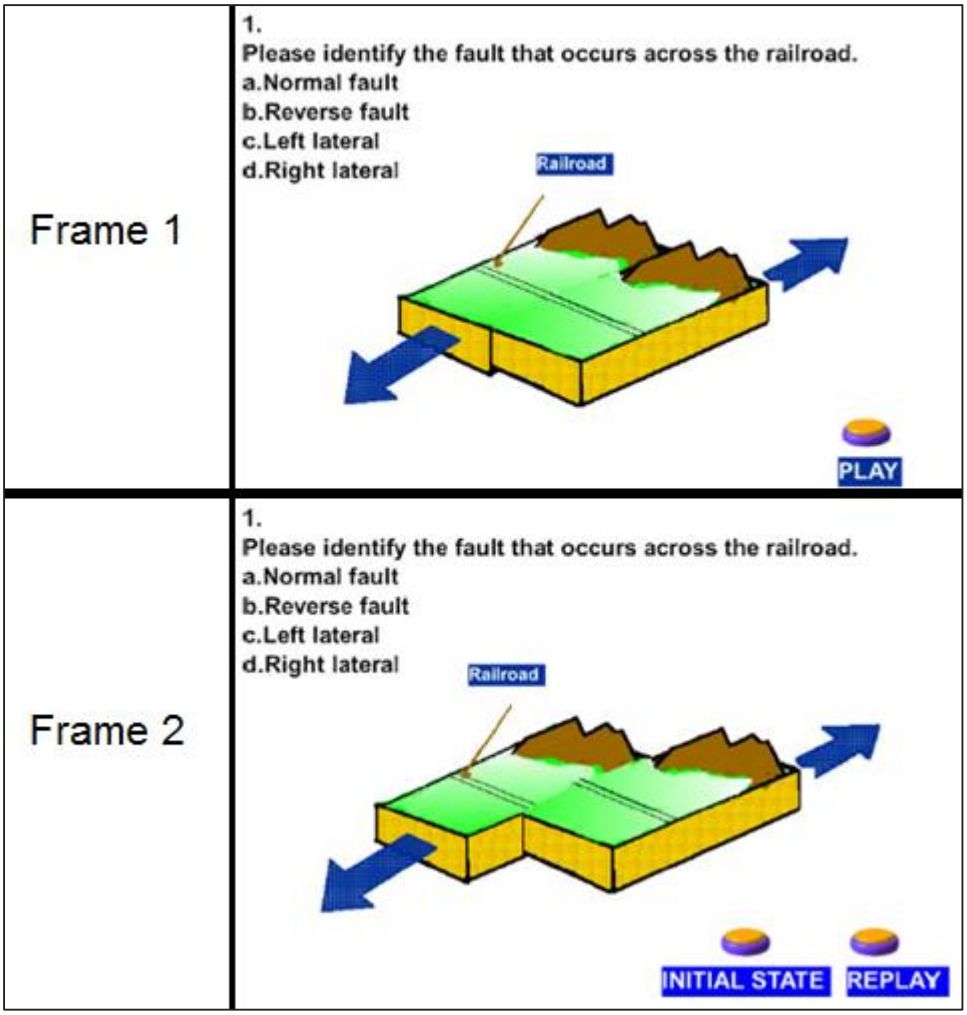


Figure 2.1. Example of a non-character video used in a science test. Reprinted from Wu, Chang, Chen, Yeh, and Liu (2010).

Regarding character-based descriptions, there are many different ways to take advantage of technology in order to replace text with videos. Early attempts to present these kinds of situations through video typically took the form of acted videos, with human actors performing scenarios, while lately, animated videos are becoming more popular (Popp et al., 2016). Animated characters can be presented in different formats that differ in their level of authenticity: (i) two dimensional (2D) animations, which have the lowest level of authenticity, (ii) 3D caricatured animations, which are more realistic than 2D animation but still not lifelike, and (iii) three dimensional (3D) realistic animations, which approach lifelike appearance (Popp et al., 2016). Figure 2.2 presents four different ways in which a character-based scenario can be presented with the use of videos in SJTs. It should be noted that, in this thesis, the term *video* or *video-based* is used as an umbrella term to refer both to *acted videos* and *animated videos*.

Figure 2.2. Images of a character-based scenario using multiple media formats. Reprinted from Bruk-Lee et al. (2013).
Note. Media formats: a. human actors, b. 2D characters, c. 3D caricatured characters, d. 3D realistic characters.

Although the use of videos in assessment is not very common, there have been attempts to incorporate them, mainly in the context of SJTs. An SJT item consists of a scenario and a number of alternative response options. Often, they present character-based situations. SJTs usually present the stimulus material (i.e., the scenario) and the response options in written format, however, this information can be replaced or accompanied by multimedia, such as images and videos. Early attempts to replace written descriptions in SJTs with videos took the form of acted videos, with human actors performing a scenario (for example, see Chan and Schmitt, 1997), while lately, some researchers, such as Bruk-Lee et al. (2016), have begun considering the potential of animated videos because of their attractiveness and the ease with which they can be modified and handled. The visualisation of situations via the use of either acted or animated videos has the potential to significantly enhance the fidelity and the quality of an assessment.

The limited number of research studies that used acted videos have provided promising findings about the benefits of video-based testing. A meta-analysis conducted by Christian et al. (2010), for example, indicated that acted-video SJTs tended to have

higher levels of criterion-related validity[9] compared to text-based SJTs. Although these findings provide a rationale in favour of the use of videos as opposed to text in assessment, they are based solely on studies where acted videos were used. 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. Thereby, it is possible to address mistakes or problems in the animated videos at any point and to keep the instruments up-to-date over time. Moreover, animations can be adjusted, in terms of language, character features, clothing and location in order to be used across countries and cultures, something that is not as easily achieved when human actors are involved (Popp et al., 2016).

An important issue to consider when comparing animated to acted videos is the time, effort and financial resources required for their development. Generally speaking, creating animated videos is a particularly time-consuming and expensive process. 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 character movements, facial expressions, environments, and objects (e.g., goanimate.com) 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).

Despite the advantages of animations, their potential in assessment has remained relatively unexplored. As it is shown later in this chapter, to date, very few research

[9] *Criterion-related validity* refers to the degree to which a test score can be used to infer a test-taker's standing on some external measure of interest; namely a criterion (AERA et al., 2014). There are two types of criterion-related validity: concurrent and predictive. In concurrent validity evidence, users are interested in the degree to which scores obtained from tests, or any other assessment tool, can be used to estimate a test-takers' standing on a criterion, which is obtained at about the same time with the test score (L. Cohen et al., 2011). In the collection of predictive validity evidence, on the other hand, test scores are used to predict individuals' standing with regards to criterion measures that will be obtained at a future time (L. Cohen et al., 2011).

studies have examined the use of animations as a useful alternative to text in assessment. Thus, there is insufficient empirical research in this area to support their effectiveness over and above text-based tests.

2.4. Research on the Use of Acted and Animated Videos in Testing

2.4.1. Early attempts to incorporate animations in testing and the impact on test-takers' performance

To date, a very small number of research studies have explored the possibility of incorporating animated videos in assessment. In order to examine the impact that animations may have on assessment, some of these research studies attempted to measure the potential influence on test-takers' performance. The available research literature is not limited to education, but it also extends to other fields, such as personnel selection. Moreover, some of these attempts examined the potential advantages of animations over text-based approached and/or static media such as still images.

Wu et al. (2010) conducted a study comparing test-takers' performance across an animated and a static version of the same test in earth science. In the static version, the test items were accompanied by static images and graphics, while in the animated test the graphics were enriched with motion. Each test consisted of 20 items, and the content was constant across the two formats. A total of 314 high school students from Taipei, who attended the respective science courses, participated in the study. Due to the restricted number of earth science teachers and learning materials, half of the students completed the earth science curriculum in the first semester (Group A), while the other half did so in the second semester (Group B). As the study was conducted at the beginning of the second semester, Group A had already completed the science course, while Group B was still attending. For this reason, it was hypothesised that participants would vary in their levels of prior knowledge in earth sciences. Students' prior knowledge was also measured by a series of earth science pre-tests. Based on their performance on these tests, students were categorised into three subgroups: high, moderate and low prior knowledge levels.

Overall, the analysis of the data did not show statistically significant differences between the animated and the static version of the test (Wu et al., 2010). However, further analysis revealed a statistically significant interaction between the format of the test and

the levels of prior knowledge. Specifically, while the performance of students with moderate and high prior knowledge was not affected by the animation of the static test, the use of animations positively influenced the performance of those with low prior knowledge (d = 0.3). Wu et al. (2010) concluded that although high- and moderate-knowledge learners were able to form mental representations from words or graphics alone, this was a difficult task for low-knowledge learners. Therefore, the use of animations was intended to help students with low prior knowledge to better understand the content of the items and to achieve their potential in the science test.

Similar conclusions were drawn by Malone and Brünken (2013), who examined whether the integration of animated traffic scenarios could enhance some quality aspects of selected multiple-choice items on the official German driving test. A sample of 57 novice and 63 experienced drivers were presented with 22 items, either in static (images and text) or in dynamic (animated) format. The results showed that only novice drivers benefited from the use of animations in the test (d = 0.55), with no significant differences emerging for experienced drivers. This finding corroborates the conclusions Wu et al. (2010) reached as well, namely, that animations may facilitate assessment, especially for groups that lack prior knowledge or experience.

Dindar, Yurdakul, and Dönmez (2013) conducted another application of animated assessment, in the field of English language learning. In contrast to the two aforementioned studies, where animations were used in the context of earth science problems and driving situations, Dindar et al. (2013) used animations to present character-based scenarios. The authors administered two versions of a 12-item multiple-choice test to 116 secondary school students in Turkey. The questions were presented either with static graphics and text or with animated graphics and text. The graphics were used to depict the context of a scene in test items and assist students in comprehending the situation presented in the question. However, the authors did not provide further details about the main features of the animated assessment versus the existing static test. Figure 2.3 presents a sample animated item used in the study.

The results of the study showed that, although students who took the animated test scored higher than those who took the static version, the differences were not statistically significant. However, in contrast to the studies described earlier, this study did not examine the effects that animations had on the performance of students with different

levels of prior knowledge, which could have given further insight into the potential impact of animations.

Figure 2.3. Sample animated item from an English language test [*sic*]. Reprinted from Dindar et al. (2013).

Summing up, based on the findings of the relevant research studies, it can be concluded that animations have a small but, in some cases, statistically significant impact on test-takers' performance, particularly when learners have low prior knowledge of the subject area. However, it should be taken into account that all of the above studies examined differences in test-takers' performance between animated and static tests, with the latter including both text and images or graphical representations. It could be expected that the magnitude of the gap may be different if animated tests were compared to purely text-based tests.

Research on the use of animations in assessment is at an early stage and this is evidenced by the small number, the fragmented nature and the depth of the available research studies. Although the investigation of the impact that animations may have on test-takers' performance is important, it should be appreciated that performance differences alone provide limited evidence regarding the value that animations may add to assessment. Along with performance, there are a number of other critical questions about validity, construct-irrelevant variance, unintended subgroup differences and test-takers' overall experience that should be answered. The following section discusses research studies that conducted more thorough investigations into the validity assets of video-based assessments compared to text-based ones, focusing, in particular, on construct-irrelevant variance in test scores that could be attributed to language skills.

2.4.2. Validity evidence for acted- and animated-video assessments

As outlined previously, validity is an essential component of any assessment. Consequently, it has been examined by research studies that have explored alternative assessment approaches, such as video-based testing. The vast majority of the research studies that attempted a video-based assessment approach used acted SJTs involving human actors in recorded videos, while the research on animated scenarios is much scarcer. Research evidence coming from the literature around acted videos is particularly valuable for informing studies that aim to take video-based assessments a step further via using animations.

The literature on video-based SJTs derives mainly from the field of personnel selection. Because of their ease in administration and scoring (Fetzer & Tuzinski, 2013) and their higher precision in predicting applicants' performance on the job, compared to other well-established selection instruments, such as cognitive ability tests (Lievens, Peeters, & Schollaert, 2008), SJTs are widely used in personnel selection assessments. Most of the research literature around SJTs focused on predictive validity, as the prediction of candidates' future performance is the main aim in most personnel selection assessments (AERA et al., 2014). Examples of SJTs used outside the personnel selection context are very limited but sufficient to demonstrate that they can be used to assess knowledge, skills and abilities in other contexts as well. Nevertheless, even for applications of SJTs in fields other than personnel selection, such as university admission exams (Lievens, 2013), the emphasis is typically placed on predictive validity, with respect to future performance in college, work placement and actual work settings.

One of the first validation studies in the area of video-based testing was conducted by Weekley and Jones (1997). These authors completed two research projects that explored the use of video-based SJTs in employment assessment for retailing and caregiver positions. In the first study, an SJT that presented the scenarios via acted videos, as opposed to written text, was administered to 787 employees in various store jobs in a discount retailer. In the second study, an acted-video SJT was administered to 148 employees in four entry-level caregiver positions (nursing assistants, dietary aides, laundry assistants, and housekeeping assistants).

Both studies examined the criterion-related evidence coming from the video-based SJT, using various criteria. Therefore, along with the SJT scores, Weekley and Jones (1997)

collected cognitive ability scores and experience data. For the same purpose, employees' performance was rated by their supervisors using a 47-item and a 58-item scale for the first and the second study, respectively. As the participants were already hired, these tests were not used as a basis of making hiring decisions, but rather, as a benchmark against which the efficacy of the video-based SJT in predicting job performance could be compared.

The analysis of the data for the two studies revealed similar results regarding the criterion validity of the inferences based on the video-based SJT test scores. In the first study, the video-based SJT scores were statistically significant predictors of test-takers' job performance, as rated by their supervisors, with a correlation of $r = .22$. Likewise, in the second study, participants' test scores in the video SJT were again statistically significantly correlated with their job performance, $r = .24$. Although the strength of these correlations is theoretically considered to be weak to moderate (J. Cohen, 1988), taking into account the predictive power of the other well-established instruments used as criteria, the authors concluded that the SJTs were good predictors of employees' performance at their jobs, providing evidence of incremental[10] validity over the other criterion measures.

Lievens, Buyse, and Sackett (2005) conducted another validation of a video-based SJT in the field of medical education. This was one of the few cases where SJTs were used in a context other than personnel selection. The authors used an acted-video SJT that measured interpersonal skills by examining doctor-patient interactions. This interpersonal skills SJT was administered to 7,197 candidates who attended for the Medical and Dental Studies Admission Exams in Belgium from 1997 to 2002. Lievens et al. (2005) hypothesised that, along with the conventional cognitive skill tests, the video-based SJT would be able to predict students' future grade point average (GPA) at the university, provided that GPA scores were not solely based on medical science courses but on interpersonal skill courses as well. Only participants who had passed the admission exam and started medical studies were included in this study. In total, the authors were able to obtain the first-year GPA of 1,768 students, the second-year GPA

[10] *Incremental validity* is the type of validity evidence that determines whether a particular test, or instrument in general, has increased criterion-related validity beyond that provided by another instrument or test (Sackett & Lievens, 2008).

of 1,087 students, the third-year GPA of 676 students, and the fourth-year GPA of 305 students.

The data supported Lievens et al.'s (2005) hypothesis. The acted-video SJT scores predicted medical students' GPA in curricula with strong interpersonal skills components, with correlations (*r*) of .12, .14, .40, and .55 in the first, second, third, and fourth year, respectively. However, and not surprisingly, this was not the case in curricula with minimal interpersonal skills elements. In addition, it was found that the video-based SJT exhibited incremental validity over the cognitive tests, only for interpersonal skills-oriented courses, but not for the overall GPA. Specifically, the video-based SJT accounted for 1% additional variance in GPA scores in the first year, 2% in the second year, 6% in the third year, and 7% in the fourth year. Especially for the last two years of the university, both the correlation coefficients and the incremental validity evidence suggested that the video-based SJT was a good predictor of students' future interpersonal skills, as measured by their GPA.

Lievens (2013) further expanded this large-scale research project by following medical students beyond the fourth year of their studies. Data were collected from 927 students who completed their seven-year studies, 261 of whom also opted for two extra years of general practice training. A job performance measure based on their supervisors' ratings was included as the main criterion validity evidence for the video-based SJT. Moreover, for students pursuing a career in general practice, additional outcome measures were taken via (i) an interpersonal skills assessment conducted using an objective structured clinical examination (OSCE), (ii) a knowledge test about general practice, and (iii) a case-based panel interview.

The analysis indicated that participants' scores in the video-based SJT were statistically significant predictors of their future job performance (r = .15), their OSCE score (r = .12) and their performance in the case-based interview (r = .19). Although the strength of these correlations is practically weak, SJTs added significant predictive value over and above cognitive tests for predicting job performance, OSCE score and participants' performance in the case-based interview. Overall, cognitive tests were found to be more accurate in predicting knowledge-based measures, such as the scores in the knowledge test about general practice (r = .21), but they failed to systematically provide predictive

validity evidence for employees' job performance, OSCE score and performance in the case-based interview.

Based on the results of both studies conducted by Lievens and his colleagues, it was concluded that even though the interpersonal skill video-based SJTs cannot replace the cognitive exams, they could further increase the validity of the inferences based on candidates' performance on the admission assessment. It is noteworthy that, in contrast to most of the available research studies in the field, these studies were conducted in the context of high-stakes exams, with test scores having consequences for the test-takers. This is something that adds weight to these findings.

It should be mentioned, though, that the reliability levels of Lievens' SJT were particularly low (Cronbach's alpha of .39). This was probably the case because, as Lievens and De Soete (2015) argued, the situations presented by most SJTs are inherently multidimensional, meaning that they typically measure a group of constructs, as opposed to one discrete, well-defined construct. The lack of construct validity evidence is a key issue in the SJT literature that has been receiving more and more attention in recent years, to the extent that many measurement experts (e.g., Guenole, Chernyshenko, & Weekly, 2017; Lievens, 2017) now recommend placing an explicit focus on the intended construct in the development stages of SJTs.

Another study in the field of medical admission exams was undertaken by Fröhlich, Kahmann, and Kadmon (2017). These authors conducted a validation of a video-based SJT assessing social competencies of medical school applicants in Germany. The sample consisted of 1,556 applicants who voluntarily took the test. Apart from participants' performance on the acted-video SJT, criterion data on social competency, personality and cognitive skills were also collected.

The results of the study showed that the acted-video SJT was statistically significantly correlated to most of the validation criteria, with weak to moderate correlations (r) ranging from .11 to .33. Moreover, in contrast to most SJTs, the test had satisfactory reliability, with a Cronbach's alpha value of .83. This was probably the case because the SJT was developed to measure a well-defined construct, namely social competency, instead of presenting a series of broad critical incidents.

Another recent example of the use of acted videos in SJTs is presented in Golubovich's et al. (2017) study, where more than 1,000 students from various American universities took part. In this study, the authors aimed to capture the relationships between behavioural perception skills and personality traits, critical thinking and mental ability. As the authors explained, they decided to use acted videos as opposed to text-based scenarios in order to improve the fidelity of the behavioural perception skills assessment, which they hoped would enable them to more adequately capture the complex construct that they were attempting to measure; namely, perceptions of others' interpersonal behaviour. According to the authors, higher fidelity can also be thought of as evidence of ecological validity, in that it refers to situations in which the stimuli in the testing method closely approximate participants' experiences in the real world. It should be mentioned, though, that the development and the use of the acted videos did not constitute the main focus of the study, hence, further validity evidence was not provided.

In the studies described earlier, test-takers were presented with acted videos describing various challenging situations and were asked to select one or more of the provided response options of handling those situations. Cucina, Su, Busciglio, Harris Thomas, and Thompson Peyton (2015), on the other hand, followed a different approach in taking advantage of multimedia in assessment. For the purposes of a Federal Law officer selection assessment, these authors created an acted-video SJT, where candidates were asked to orally give their open-ended responses to eight recorded scenarios that presented challenging situations. In total, data were collected from 4,725 applicants. Of the 4,725 applicants, 1,923 were eventually hired and successfully completed a training programme, providing data regarding their performance during the training. Finally, 439 of these individuals took part in a criterion-related validation study that included two job performance criteria (performance ratings and work sample scores). The data analysis revealed that candidates' performance on the acted-video assessment predicted their performance both in training and on the actual job, demonstrating statistically significant but low to moderate levels of criterion-related validity; the overall summary estimate of the video-based test criterion-related validity was $r = .18$ (Cucina et al., 2015).

Overall, from the limited examples where acted videos have been used in testing, it could be concluded that the incorporation of videos in assessment has been relatively successful. However, despite the fact that these studies provided some validity evidence

regarding the use of video-based tests, an important question is whether video-based tests facilitate more valid inferences compared to equivalent text-based tests. The non-experimental design of the aforementioned studies does not provide any comparative information among different SJT formats (i.e., video-based and written SJTs) to address this question. Comparisons of the video to the written formats of the same tests are needed to provide meaningful information regarding the additional benefits that videos can bring into assessment.

A seminal experiment that was conducted by Lievens and Sackett (2006) who compared validity evidence from a video-based and a text-based SJT. As mentioned above (i.e., Lievens, 2013; Lievens et al., 2005), the interpersonal skills acted-video SJT, which was developed for the Medical and Dental Studies Admission Exams in Belgium was administered from 1997 to 2002. However, due to cost and technological concerns, the Belgian government decided to explore the possibility of changing the SJT format from video-based to written. Towards this end, Lievens and Sackett created a text-based SJT version of the existing video-based SJT that was used in 2000. The sample of the study consisted of 2,909 students; 1,159 of them took the acted-video version in 2000 and the rest took the written version three years later. The demographic characteristics of the two groups confirmed that they were highly similar. Data on three measurement criteria were collected: (i) students' performance in interpersonal skills-oriented courses, (ii) students' overall GPA, and (iii) students' cognitive ability (verbal, numerical and figural).

The results of this research study indicated that the video-based SJT was a much stronger predictor of students' performance in medical courses focusing on interpersonal skills (r = .35) than the text-based version of the test (r = .09). Moreover, it was hypothesised that the acted-video SJT would have more incremental validity over the cognitive admission exams in predicting performance in interpersonal skills-oriented courses compared to the text-based SJT. This hypothesis was empirically supported as, in contrast to the text-based SJT, the acted-video SJT accounted for a substantial amount of variance in the interpersonal criterion (11%).

Finally, the results showed that the correlation between students' text-based SJT performance and their cognitive ability scores (r = .18) was statistically significantly higher than the correlation between the acted-video SJT scores and their cognitive ability

measures (r = .11). This finding led the authors to conclude that the increased power of the acted-video SJT in predicting students' performance in interpersonal skills-oriented courses was not attributed to its correlation with cognitive ability, but to the fact that the virtual representation of the situations increased fidelity and resemblance to the criterion.

The stronger criterion validity evidence for video-based SJTs over text-based ones was also supported by a meta-analysis undertaken by Christian et al. (2010). After analysing 134 independent effect sizes, 11 of which came from studies where acted-video SJTs were used, the authors concluded that the video-based SJTs tended to have stronger criterion-related validity than the paper-and-pencil SJTs. It should be highlighted, though, that most of the studies included in the meta-analysis used either a video or a text-based SJT and very few studies followed an experimental design administering the same test in both formats. Therefore, the lack of direct comparisons limits the conclusions that can be drawn from this evidence.

Furthermore, the same meta-analysis found that the majority of SJTs in employment selection focused on the assessment of leadership and/or interpersonal skills (Christian et al., 2010). Considerably fewer SJTs measured personal tendencies and teamwork skills, while 33% of the SJTs did not focus on a particular construct, rather, they measured heterogeneous composites. Based on the findings of their systematic review, the authors highlighted that a video-based format was more likely to facilitate the measurement of constructs that rely on complex contextual information, such as social skills. Christian et al. (2010) argued that this kind of information can be more successfully replicated and transmitted through video, compared to long passages of text. Their argument was supported, as the interpersonal skills SJTs, where scenarios were presented via acted videos, as opposed to written text, were stronger predictors of other criteria of interest (criterion validity evidence). This evidence informed the decision to use an SJT with a strong interpersonal orientation (i.e., practical knowledge of how to deal with others) for this doctoral research study.

Although this doctoral study focused on construct-irrelevant variance as a validity issue and it was beyond its scope to examine the predictive power of the practical knowledge SJT scores, it should be acknowledged that criterion, and particularly, predictive validity evidence has been the focus of the studies that examined the efficacy of video-based

testing. The findings of the validation studies discussed above are particularly promising regarding the potential benefits of the use of acted videos in testing. Given the limited use of animations in testing, however, similar conclusions regarding this format of assessment cannot be drawn.

2.4.3. The impact of acted and animated videos on reducing construct-irrelevant variance attributed to language and reading skills

Apart from the studies mentioned above, which focused mainly on criterion validity evidence, several attempts have been also made to compare construct validity evidence from video-based and text-based tests, particularly with regard to the reading demands the of tests. Videos can fully or partially replace written descriptions in tests and increase the quantity and quality of the stimulus materials that test-takers are presented with. With this in mind, a number of research studies explored whether video-based tests can overcome the issues linked to the excessive use of written text in assessment by mitigating the impact that construct-irrelevant factors, such as language and reading skills, may have on test-takers' performance. Most of these studies used video-based techniques in the form of acted videos, while the use of animated videos is much rarer.

Chan and Schmitt (1997) were among the first scholars to explore the benefits that videos can bring to assessment in terms of reducing construct-irrelevant variance. They presented 113 Black and 128 White undergraduate psychology students with either the acted-video or the paper-and-pencil version of a 12-item SJT, assessing work habits and interpersonal skills of candidates for human resources positions. A reading comprehension test was administered to both groups to provide evidence on the extent to which potential Black - White differences in performance on the paper-and-pencil test were due to the reading demands inherent in the test.

The results indicated that the ethnic subgroup differences were smaller in the video-based SJT (d = 0.21) compared to the text-based SJT (d = 0.95), in both cases favouring White participants. As the authors explained, the performance gap in the text-based SJT in favour of White students was not fully attributed to their more advanced construct knowledge, but rather to Black students' limited reading comprehension, which prevented them from fully comprehending the test items and, hence, performing to their potential. The use of acted videos, as opposed to written text, helped Black students comprehend the test items and perform better. More specifically, the results of the study

showed that there was a statistically significant correlation between students' test performance and their reading comprehension ability, only in the case of the text-based test, and not in the case of the video-based SJT[11]. This example illustrates that while a performance gap between two groups can be originally seen as a racial issue, it can actually be attributed to other construct-irrelevant factors (e.g., reading skills). The findings of Chan and Schmitt's (1997) study suggest that multimedia, and more specifically video-based tests, can actually eliminate the possible adverse impact that reading comprehension may have on assessment outcomes, improving the fairness and validity of the inferences from the test scores.

In contrast to the previous research findings, MacCann, Lievens, Libbrecht, and Roberts (2016) failed to find empirical support for the benefits of acted versus text-based SJTs. Using a well-established text-based measure of emotional management (MSCEIT; J. D. Mayer, Salovey, & Caruso, 2002), the authors developed a multimedia emotion management SJT that used acted videos, as opposed to written descriptions, to present both the scenarios and the response options (see Figure 2.4). The authors compared these two versions of the test (i.e., text-based vs. acted-video) by randomly assigning 427 US college students to one of the two assessment formats. They hypothesised that the text-based SJT might have been susceptible to construct-irrelevant variance because it required a threshold level of reading comprehension.

[11] Raw correlations (*r*) were not provided by the authors.

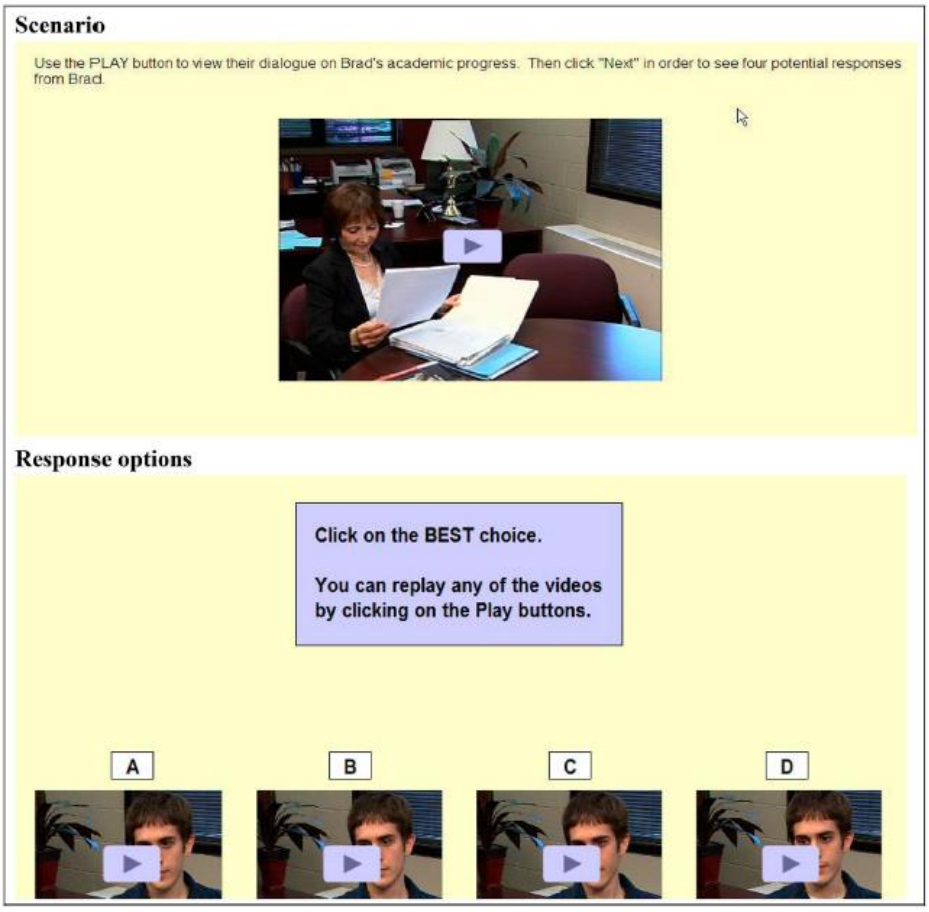


Figure 2.4. Sample scenario and response options from the multimedia emotion management SJT. Reprinted from MacCann et al. (2016).

To test this hypothesis, participants took a series of cognitive ability tests in addition to the SJT. MacCann et al.'s (2016) hypothesis was not supported, as the correlations between the cognitive measures and the SJT scores were not significantly lower for the acted (r = .39 – .55) compared to the text-based (r = .40 – .53) version of the SJT. Although the cognitive measures that the authors used had some verbal elements (i.e., sentence completion, vocabulary and verbal analogies), the study did not include any measure of reading comprehension. Therefore, robust conclusions about construct-irrelevant variance in the acted compared to text-based teat cannot be drawn based on this evidence. Additionally, it should be noted that the two versions of the test, although similar, they were not identical as when the original text-based scenarios were transformed into scripts for the filming of the acted SJTs, they took the form of dialogues as opposed to statements and, therefore, more details were included.

Although the above studies provide some insight into the limitations linked to the extensive use of text in assessment and highlight the alternatives that technology can

provide, they focused solely on examining the efficacy of acted videos. Dancy and Beichner (2006) were among the first researchers to date who addressed the issue of construct-irrelevant variance using animation technology. The aim of their research was to compare a static to an animated version of the same science test. Their argument was that science is a field that includes abstract and complex concepts, where still representations of phenomena or situations via text, images or graphs may not always be effective in providing students with a sound understanding of the intent of the questions or problems posed. In their study, a parallel version of the paper-and-pencil force concept inventory test was developed by replacing static pictures and descriptions of motion with animated representations for all 30 multiple-choice items of the test. Figures 2.5 and 2.6 present the same item in its static and animated format.

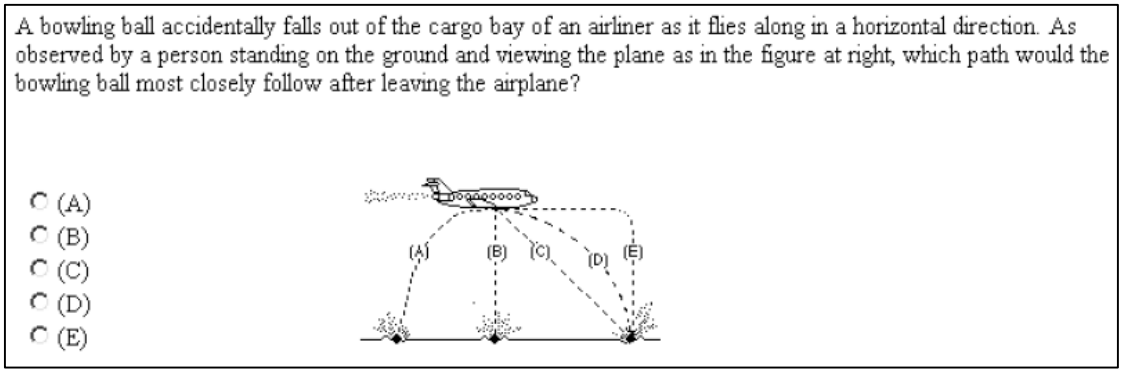


Figure 2.5. Static version of an item from the Force Concept Inventory test. Reprinted from Dancy and Beichner (2006).

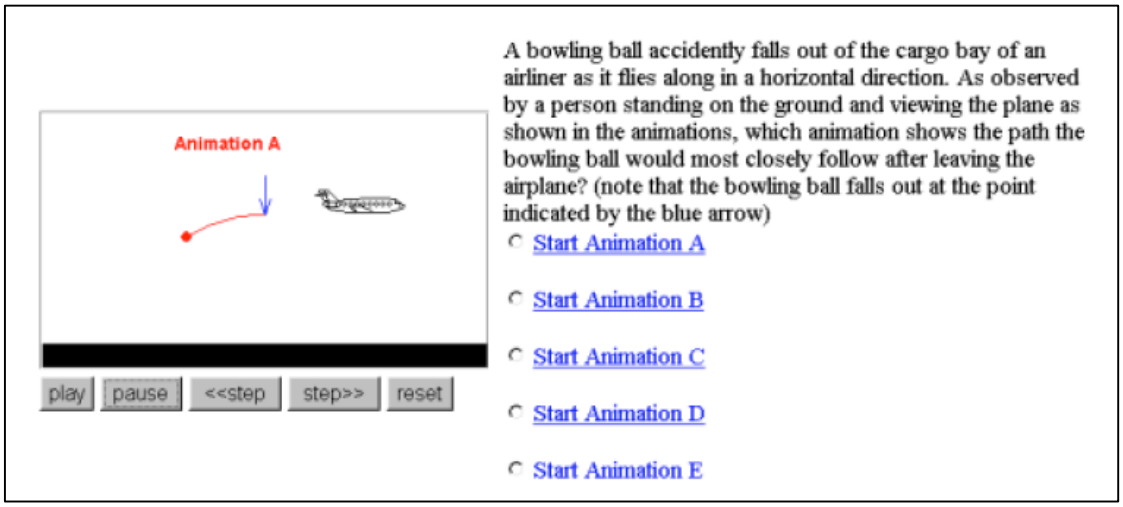


Figure 2.6. Animated version of an item from the Force Concept Inventory test. Reprinted from Dancy and Beichner (2006).

It should be noted that, according to the authors, only 14 of the 30 items of the test included animations that provided valuable information which could have assisted students in correctly understand the content of the question. For the rest of the questions, all the relevant information was adequately communicated through text and images, despite the fact that even these items were animated for the purposes of the study. Using random assignment, the animated and the static versions of the test were administered to a sample of 53 high school and 325 university students in the US.

Dancy and Beichner's (2006) preliminary analysis indicated that there were statistically significant performance differences between the experimental (i.e., animated) and the control (i.e., static) group, in six out of 30 items administered. There were three questions on which the experimental group performed significantly better and three on which the control group performed significantly better. Interestingly enough, the significant differences between those who took the static and those who took the animated version of the science test were found for the items that were hypothesised to contain information that could not be effectively communicated via static text and images. However, more than half of the items that were hypothesised to be affected by the use of animations were not answered differently by the experimental group, compared to the control group.

Although the findings provide some support for Dancy and Beichner's (2006) argument that animations should be an integral part of these questions to convey useful information, the fact that they did not consistently assist test-takers in selecting the correct response limits the conclusions that can be drawn from the study. Nevertheless, based on interviews conducted with 14 students following the testing, the authors argued that this was the case because the animations proved to be helpful only for students who had a good knowledge, assisting them in correctly answering the questions. On the other hand, it was argued that animations led those who did not have good knowledge, away from selecting the correct responses. According to Dancy and Beichner (2006), in both cases, the animated questions improved the precision of the assessment. However, this conclusion is open to debate.

To further explore the role of animations in enhancing assessment validity, Dancy and Beichner (2006) administered a verbal skills test in order to examine if animations would mitigate the adverse impact of lower verbal skills on students' performance in the static

version on the test. The results indicated that, although students' performance on the verbal skills test was significantly correlated to their performance on the static version of the science test, this was not the case for the animated version. In more detail, students with higher verbal ability tended to perform better in the static science test (r = .22), while this correlation was not statistically significant in the animated version of the test (r = .07). Moreover, student interviews in the same study confirmed that the static versions of the questions were more likely to be misread; even when students managed to correctly read the static questions, they often reported misunderstanding them (Dancy & Beichner, 2006).

Overall, Dancy and Beichner's (2006) results provided positive evidence regarding the use of animated as opposed to static representations in assessment. It should be acknowledged that the quality of the animations was quite basic and the fidelity they managed to offer was not very high. Despite the fact that these findings are relevant to the present thesis, they refer to the use of animations as an instrument of representing phenomena in science tests, and not as a way of representing character-based scenarios, which is something that this study purported to do.

2.4.4. The impact of acted and animated videos on test-takers' reactions to the test

Much of the discourse around assessment and the alternative testing formats has been focused on the valid and fair measurement of individuals' knowledge and skills – and rightly so. However, it is also valuable to know what test-takers think about the quality of the tests they take. A considerable number of research studies on assessment have examined test-takers' reactions to different assessment approaches, techniques and instruments. In the field of video-based assessments, the majority of studies focused on participants' perceptions of the validity and fairness of a test, as well as on the extent to which test-takers enjoy the assessment process. Research on how the use of multimedia, such as animated videos, may impact test-takers' perceptions of the difficulty of a test and their invested effort in it is much rarer.

As Chan and Schmitt (2004) argued, the examination of test-takers' perceptions of testing situations allows the subsequent investigation of the influence that these kinds of reactions may have on test-takers' behaviour during and after the assessment process, such as performance on the test and attitudes towards the body organising the

assessment. A central concept in the area of test-takers' perceptions is face validity. Face validity is defined as the extent to which a measurement tool appears effective and relevant in terms of its stated aim (Gravetter & Forzano, 2012). In assessment and testing, taking into account the latest *Standards for Educational and Psychological Testing*, according to which fairness constitutes a fundamental component of validity, it could be argued that face validity reflects perceptions of both the validity and fairness of the inferences drawn from a test. It should be noted, though, that most of the available studies perceived face validity and perceived fairness as two distinct constructs.

Popp et al. (2016) mentioned that a test should be face valid for three different groups of people: (i) those who make the final decision to use the test, (ii) those who administer the test, and (iii) those who take the test. It should be acknowledged that face validity evidence alone is not an acceptable substitute for other forms of validity evidence. That said, as Scott and Mead (2011) argued, the concept is still important because examinees who believe that they are being assessed on characteristics relevant to the purpose of the test are more likely to place credence on the measure and try their best. This has been supported by research studies, such as that conducted by Chan, Schmitt, DeShon, Clause, and Delbridge (1997). In this study, 210 university students in the US were given two parallel forms of a cognitive ability test used for personnel selection purposes in a large telecommunications company. Examinees were encouraged to treat the testing session as a practice opportunity for actual selection testing in which they might participate in the future. All participants were first given the original test, followed by a parallel version of the same test. Between the two tests, an instrument measuring their perceptions of the validity of the test (i.e., face validity) and their motivation to take the test was administered.

By measuring both face validity and motivation before and after test completion, Chan et al. (1997) managed to capture the complex nature of the relationships between these variables and test performance. The results revealed that participants' ratings of the validity of the initial test had a positive impact on their performance on the subsequent parallel version of the test. It was also found that these relationships were reciprocal as test-takers' performance was found to affect their perceptions of test validity as well. In other words, participants who performed well tended also to consider the test to be more valid, compared to those who did not perform so well at the same test. Nevertheless, this

relationship was indirect, as face validity affected test-takers' motivation to do well on that test, which in turn had an impact on performance.

As Popp et al. (2016) argued, in high-stakes tests, where the test results have short- or long-term consequences for the test-takers, participants are expected to be motivated and fully engaged in the process, independently of face validity. Nevertheless, face validity and motivation, as mediators for enhancing learners' engagement and invested effort in an assessment may be crucial in the case of more low-stakes and formative assessment procedures. Hopfenbeck and Kjærnsli (2016), for instance, highlighted the importance of further examining and promoting students' test motivation in low-stakes large-scale assessments, such as PISA, as a way to improve the validity of their results, given the influence that these results have on policy-making across the participating countries. Harlen (2012) contended that motivation and engagement should be principal characteristics in assessments for learning, implying that an engaging assessment that focuses on enhancing students' learning, is expected to further motivate them to achieve their learning goals.

Another study conducted by Edwards and Arthur (2007) with 455 psychology students in a US university also supported the important role of face validity in testing. In this study, participants were given mathematics and science reasoning tests, both in selection and constructive format. The results indicated the negative impact that low levels of face validity and motivation have, not only on test-takers' performance but also on their attitudes towards the fairness of a test. In addition, Popp et al. (2016) pointed out that another aspect that can be negatively influenced by lower levels of face validity is participants' attitudes towards the reliability of the results.

After reviewing the research literature on various selection assessments, Chan and Schmitt (2004) highlighted the positive impact that technology can have on improving test-takers' reactions to a test, such as face validity. Many of the studies that Chan and Schmitt (2004) reviewed examined the impact of video-based tests on participants' perceptions of an assessment. Evidence regarding test-takers' perceptions of video-based assessments derives mostly from studies that explored the use of acted videos, while research around the attitudes towards animated tests has only emerged in recent years.

One of the first research studies to compare a video- to a text-based test in terms of test-takers' perceptions was the Chan and Schmitt (1997) study (as discussed previously). After completing their SJTs, participants in this study were asked to give ratings regarding the degree to which the tests were relevant to the position for which they had hypothetically applied (the study was not conducted with real job candidates). The results indicated that the acted-video version of the test had higher face validity than the text-based one (d = 0.50). More specifically, the participants who took the video-based SJT considered it to be a more accurate measure of skills relevant to the posted job.

The role of videos in improving the levels of face validity has also been empirically explored by Richman-Hirsch et al. (2000). In this study, test-takers' reactions to three different versions of the same conflict resolution skills test (i. paper-and-pencil, ii. text-based administered on computer screens, and iii. acted videos) were examined. Instead of creating an overall face validity measure, the authors examined the different aspects of this construct separately; namely test-takers' perceptions of the content validity, predictive validity, and job relevance of the test. After randomly allocating 131 managers from several organisations to these three assessment formats, the authors measured their perceptions of the test they took. The results revealed that the participants who were assigned to take the videos-based version of the test perceived it as more content valid (d = 0.52), more predictively valid (d = 0.34) and more job relevant (d = 0.29) than those who took the text-based version of the test (administered via either paper-and-pencil or computer)[12]. Although the video-based test was perceived as slightly fairer than the text-based versions, this difference was not statistically significant.

Apart from face validity, innovative assessment formats that take advantage of technology may also improve test-takers' enjoyment of the testing process. Indisputably, providing test-takers with a more enjoyable assessment is likely to improve their overall testing experience. Lievens and De Soete (2012) argued that the use of technology-enabled test instruments, such as animated-video tests, could improve test-takers' perceptions of the assessment process, leading to a more enjoyable and engaging experience. In addition to investigating face validity, Richman-Hirsch et al. (2000)

[12] For the computations of the effect sizes, the means and standards deviations for the video-based test were compared to the pooled average of the other two administration formats.

provided evidence to support the argument that video-based tests can improve the extent to which test-takers enjoy the test they complete. More specifically, the analysis of 131 participants' attitudes towards the three different formats of the same personnel selection test indicated that the acted-video test was perceived as a more enjoyable (d = 0.44) and satisfactory experience (d = 0.48) in comparison to the two text-based formats.

Lievens and Sackett (2006) reached a different conclusion to Richman-Hirsch et al. (2000) regarding the relationship between test format and participants' perceptions of test validity. These authors compared acted-video to text-based SJTs for medical admission exams in Belgium, which aimed to measure interpersonal skills (as discussed previously). Analysis of more than 3,000 students' responses revealed no statistically significant differences in the levels of face validity between the video- and the text-based versions of the test. Some contextual factors need to be taken into account when interpreting such conflicting findings. For example, although the two studies conceptualised face validity in a similar way, it should be acknowledged that Richman-Hirsch et al. captured the overall construct more comprehensively by using a more extensive list of items that specifically captured perceptions of different types of validity evidence. On the other hand, the fact that in Lievens's and Sackett's (2006) study participants were real candidates taking high-stakes admission exams adds more weight to their findings, something that was not the case in Richman-Hirsch et al.'s (2000) study.

It should be noted that all of the above studies adopted the same approach, whereby the scenarios were presented via acted videos, but the response options were kept in written format. Kanning, Grewe, Hollenberg, and Hadouch (2006), on the other hand, attempted to create a video-based assessment where both the stems and the response options of the items were presented with the use of acted videos. The purpose of their study, conducted with 284 police officers in Germany, was to examine whether such an approach would elicit more favourable responses regarding the experience that test-takers had. After administering a selection of SJT items in three different formats, the authors measured test-takers' attitudes towards each test. In the first format, the SJT items were presented via text, in the second format the scenarios were presented via acted videos and the response options were presented via text, and in the third format, both the scenarios and the response options were presented via acted videos.

The results of Kanning et al.'s (2006) study supported the main findings of previous research; the text-based items were perceived to have the lowest levels of job-relatedness, an indicator of face validity. However, their findings showed that there were no statistically significant differences in test-takers' perceptions of the test fairness and validity between the acted-video SJTs with and without recorded response options. This finding suggested that the most important step in improving test-takers' attitudes towards quality aspects of SJTs may be the visual representations of the stimulus materials (i.e., scenarios), as the use of acted videos in the response options did not have an impact on test-takers' experience. This evidence informed the decision-making about the development of the animations for this doctoral research (additional information is provided in Chapter 3 and Appendix A).

In the last decade, the potential advantages of animated over acted videos, as described earlier, have led to some research projects specifically investigating the use of animations in testing. A seminal research study conducted by Bruk-Lee et al. (2016) is one of the few sources of evidence regarding test-takers' perceptions of animated SJTs. In this study, test-takers were presented with either an animated or a text-based SJT for a sales position. In the animated version of the SJT, the original text-based scenarios (i.e., stimuli) describing the situations were replaced by animated videos, while the response options kept their original text-based format. A total of 440 students from a large American university were given the same SJT in one of the two formats and were asked to respond as if they were actual job candidates.

The analysis of participants' responses to the survey questions following the assessment indicated that the animated SJT was perceived as significantly more job-relevant than the text-based version of the test. It should be mentioned, though, that the effect size of the difference between the two formats was particularly small ($d = 0.08$) and significant only in the one-tailed Independent-Samples T-test. No statistically significant differences were found in test-takers' perceptions of opportunity to perform, which was used as an indicator of perceived fairness, between the animated and the text-based test.

Another research project (unpublished master's thesis; Halabi, 2012) provided evidence regarding test-takers' reactions to an animated test. A total of 148 undergraduate psychology students from a US university voluntarily took both an animated and a written format of the retail sales and services simulation SJT. At the end of each testing

process, applicants were provided with a list of adjectives and were asked to select all those that applied to the SJT they had completed most recently. They were also asked to mention which of the two administration formats they preferred. The vast majority of participants (81%) reported that they preferred the animated SJT. Moreover, a statistically significantly greater percentage of students found the animated format of the test to be enjoyable (animated SJT: 82%, text-based SJT: 18%), modern (animated SJT: 73%, text-based SJT: 27%), and engaging (animated SJT: 76%, text-based SJT: 26%) than the text-based one. However, the non-experimental design of the study restricts the conclusions that can be reached.

An attempt to incorporate animated scenarios in assessment and measure test-takers' perceptions was undertaken by Wu et al. (2010). Here, the authors compared animated and static versions (text and images) of the same science test and measured 314 high school students' perceptions of solely the animated test via the Attitude toward Animation Assessment Scale questionnaire. The results demonstrated that over 60% of the students who took the animated test were satisfied with the assessment process and, overall, the animated version was characterised by wide acceptance and positive perceptions among the participants. However, the fact that participants' perceptions towards the static version of the test were not measured significantly limits the conclusions that could be drawn about the greater acceptance of the animated over the static representation of natural phenomena.

As a final point, it should be reiterated that, with some exceptions (i.e., Fröhlich et al., 2017), most of the SJTs in the aforementioned studies had low levels of reliability (i.e., Cronbach's alpha values below 0.7). As Lievens and De Soete (2015) explained, this can be attributed to the multidimensional nature of SJTs. From a measurement point of view, this is a concerning issue that affects the robustness of the results of these research studies, but it can also have significant implications for legal defensibility of high-stakes tests used for selection, certification, and licensure purposes.

2.5. Are Animations a Panacea?

All video-based tests, including animations, are expected to have higher levels of fidelity than their equivalent text-based tests. However, they should not be considered to be a panacea. Animated videos may not always be the optimal way of presenting complex

information. As Wouters, Paas, and van Merrienboer (2008) argued following a review of the relevant literature, the fact that animations can present distinct elements of a situation simultaneously may not always render them better than static representations, where learners can digest information at their own pace. In animations, many different sources of information such as objects and human representations interact and simultaneously convey sophisticated messages. This can create substantial extraneous cognitive load, which can place excessive demands on learners' working memories and may affect learners' ability to comprehend the material (Sweller, Ayres, & Kalyuga, 2011). This theory around cognitive load, although formulated in the field of learning, may also apply to the field of assessment, where animations can be used to present complex scenarios that test-takers must fully understand in order to answer the question. Indeed, a number of scholars argued that this might be the case in testing as, when multimedia is used, it may add irrelevant contextual information and inadvertently introduce error into the measurement of participants' knowledge and skills (Dicerbo, 2017; Kirschner, Park, Malone, & Jarodzka, 2017; Weekley & Jones, 1997). However, the discussion around the demands that multimedia can input on test-takers does not necessarily suggest that the presentation of information using multiple media should be avoided, rather, that it should be done carefully to avoid the introduction of measurement error.

In contrast to the *cognitive load theory*, the majority of the research studies, as presented in this literature review, suggest that, when it comes to stimulus information, more is better. It was demonstrated that media-rich assessments and, more specifically, video-based tests, were more efficient than text-based equivalents, in terms of enhancing assessment validity. On the other hand, there is no evidence favouring text-based assessments over video-based ones.

These results support the *additive theory*, where it is postulated that accuracy accumulates as a linear function of available information (Archer & Akert, 1980). This implies that video representations of a situation are expected to be more authentic than written text transcripts. It should be noted, though, that careful and purposeful design is a key requirement for achieving positive results. Even though there is some evidence to support the advantages of incorporating acted videos in assessment, research literature

regarding the efficacy of animations is much scarcer. Given the limited evidence, robust conclusions regarding the incremental quality of animated tests cannot be drawn.

2.6. Conclusions

On the whole, the research literature exploring alternatives to text-based tests is quite limited. Although there is an abundance of research on the use of technology to enhance teaching, learning, and instructional materials, far less attention has been afforded to assessment and testing contexts. This doctoral study aimed to address this gap by providing important information about the potential of technology that is specific to the field of assessment.

Despite the fact that technology provides great opportunities for improving the fidelity and complexity of tests, text is still the main way of presenting information in the vast majority of assessments. However, as the research discussed in this chapter suggests, in certain contexts and for certain groups of test-takers, a text-centric approach is likely to negatively impact the quality of an assessment in many ways. First of all, tests with heavy reading demands require a threshold level of reading comprehension and verbal skills, competencies that are often irrelevant to the construct of interest. This can affect test-takers' performance and, as a consequence, the quality of the inferences drawn from their scores. This upshot has a significant impact on validity. Moreover, text-based tests do not always meet the current demands for the measurement of more sophisticated knowledge and skills, due to the restricted complexity of what can be presented as stimuli and/or responses through this medium. Finally, in the context of computer-based assessments, reading long texts on computer screens, as opposed to printed passages, may impinge on perceivers' ability to fully comprehend the content of the test items. In recognition of these issues, this study focused specifically on a computer-based assessment of a complex construct that is unrelated to reading ability.

A small number of studies have examined the use of videos as an alternative to written text, with the first attempts involving acted, rather than animated videos. The results of some of these studies were promising regarding the role of videos in (i) reducing construct-irrelevant variance and unintended subgroup differences attributed to language and reading skills and (ii) improving test-takers' perceptions of the test. Based on limited evidence, it could be argued that, in most of the cases, video representations,

both animated and acted, may be a more effective way of presenting complex information than text-based approaches. The advantages that videos have over text-based tests seemed to improve the validity of the inferences drawn from test scores by mitigating the adverse impact of construct-irrelevant factors, such as reading comprehension. However, the research literature is quite limited, with the highest quality studies being conducted more than a decade ago. Additionally, there have been studies that found no difference between acted and text-based tests both in terms of construct-irrelevant variance. This study, thus, sought to provide some up-to-date information and greater clarity regarding this issue. Finally, the role of animations in reducing construct-irrelevant variance attributed to language and reading skills has not yet been explored in the context of character-based SJTs – as such, this study was the first to provide information specific to this context.

Despite the important role that test-takers' attitudes may play in assessment, there is not an abundance of research studies comparing participants' perceptions of video- and text-based assessment formats. In most of the cases, video-based approaches were perceived as more valid, fair and enjoyable than their text-based equivalents. Again, it should be noted that there have also been examples where no differences between acted and text-based tests were found in terms of test-takers' perceptions. However, conclusions are not definitive due to the limited quantity and questionable quality of the available studies. Additionally, other important test-takers' reactions (e.g., perceived difficulty of different aspects of an assessment or invested effort in a test) require further investigation, as they have not been given sufficient attention by the available research literature. This is another gap that this doctoral study aimed to address by examining a range of test-takers' reactions to the test.

It should be highlighted that, in the context of high-stakes exams, it is expected that participants are highly motivated and engaged in the assessment process and that they put much effort to perform well, regardless of what they 'think' of the test (Penk & Richter, 2017). However, in studies where the results do not have any short- or long-term consequences, test-takers can be expected to vary in terms of their motivation. Although certain designs (e.g., true experiments) can control for participants' different levels of motivation to take an assessment, they cannot control for the effort that participants put in during an assessment, as this is a motivational aspect that can be

affected by the test format. In other words, the format of a test may affect not only test-takers' perceptions, but also the effort they put in the test, and consequently, their performance. This is an important aspect that has been taken into account in this study.

Overall, based on the available research literature as described above, it can be argued that despite the advantages of animations over acted videos, their potential in assessment has remained relatively unexplored. The efficiency of animations, as another alternative to text-based assessment, has been investigated in a handful of studies only. Some of these have been undertaken in the field of science testing, where images and written descriptions of motion or natural phenomena were replaced by animated representations (e.g., Dancy & Beichner, 2006; Wu et al., 2010). A very small number of studies have used animations to replace written descriptions in character-based SJTs (e.g., Bruk-Lee et al., 2016; Halabi, 2012). These studies attempted to explore the impact of animations on test-takers' perceptions of the test, putting less emphasis on other important quality aspects of the assessment, such as construct-irrelevant variance. This was a significant gap that this research attempted to address.

In summary, this literature review has revealed significant gaps in knowledge which now provide the justification for why the study described in Chapter 3 was undertaken. This doctoral study was the first-of-its-kind in setting out to investigate in-depth what animations can contribute over and above conventional text-based SJTs. On top of that, this was one of the few studies that investigated the effectiveness of animations in a context other than personnel selection (i.e., education). The study was designed with two broad aims in mind: (i) to examine whether animated videos had an impact on construct-irrelevant variance and unintended subgroup differences attributed to language and reading comprehension skills and (ii) to evaluate whether test-takers' reactions to the test were affected by the use of animated videos. The specific research questions pertaining to these two aims are outlined in the following chapter.

Chapter 3: Methodology

3.1. Introduction

This study explored the benefits that animations can bring to assessment and testing. In light of knowledge gaps identified in the literature, it sought to discover whether animated videos, as an alternative to text-based descriptions in situational judgment tests (SJTs), have the potential to reduce construct-irrelevant variance and positively affect test-takers' reactions to the test. For the purposes of the study, two versions (i.e., animated and text-based) of the same practical knowledge SJT (PK-SJT) were compared.

The following sections describe the methodology used to address the research problem of the study. A conceptual framework for the study is first presented and discussed. Following that, the research questions underpinning the study are stated. The design, the participants and the measures of the study are then described. Particular emphasis is placed on providing a detailed description of the test of practical knowledge used in this study (i.e., PK-SJT). Finally, the processes followed in the pilot and the main studies, as well as the ethical considerations are outlined.

3.2. Conceptual Framework

The research problem, as discussed in Chapters 1 and 2, underpinned the conceptual framework for this study (Figure 3.1). The majority of tests today are highly dependent on written text and, attempts to reduce the reliance of a test on written text via the use of videos are rare. As depicted on the following page, along the left-hand side of the conceptual framework, the issues associated with the use of text-based tests are presented, while, along the right-hand side, the hypotheses linked to the use of animated videos in tests are outlined.

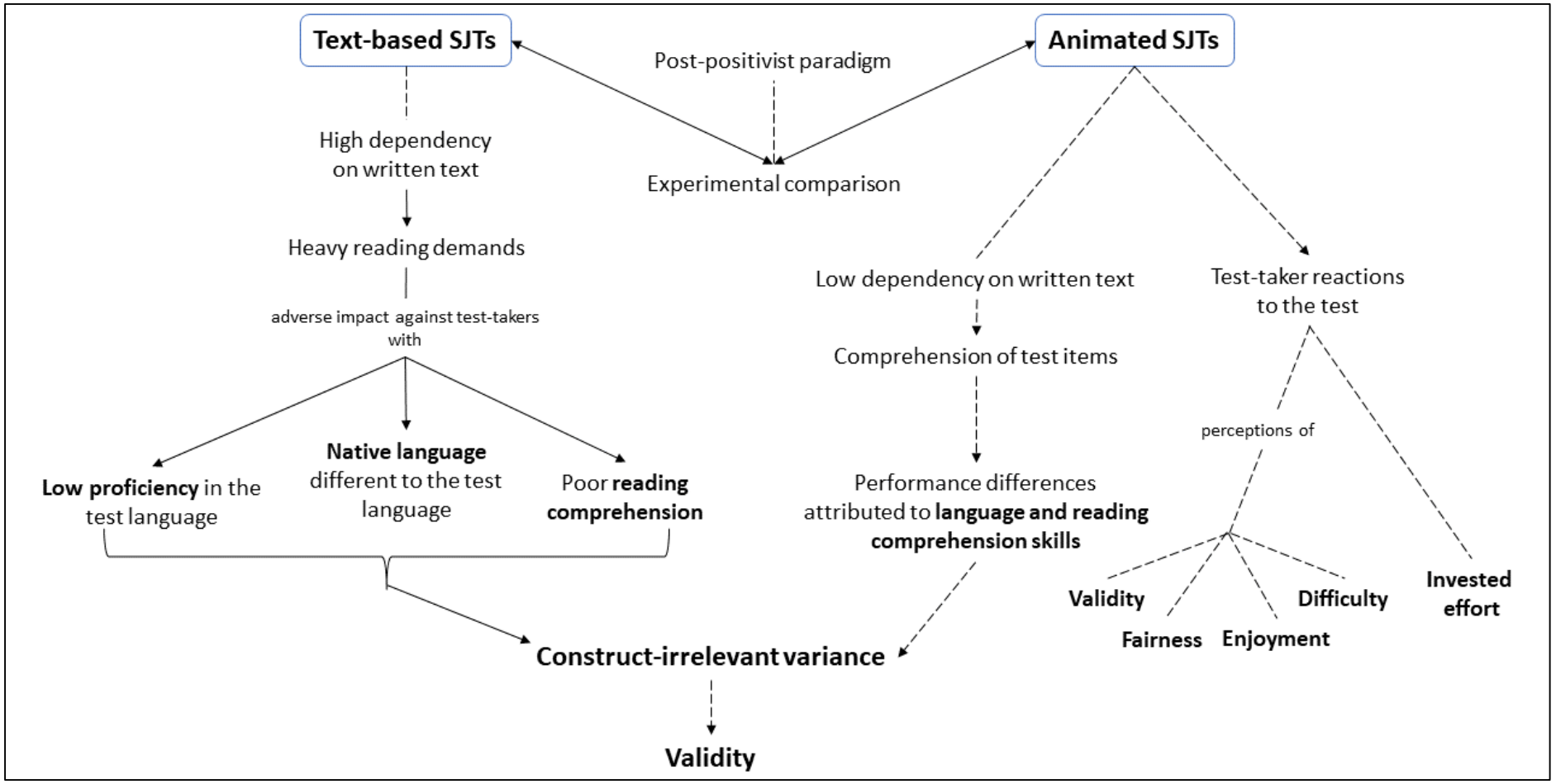


Figure 3.1. The conceptual framework for the study.

Note. A ---- B, indicates that A includes B or that B is a characteristic of A
A ⟶ B, indicates that A leads to B
A ⇢ B, indicates that A is expected to have an impact on B
A ⟷ B, indicates that A is examined through B

SJTs were included in the conceptual framework because they constitute a particular type of selected-response test that aims to measure more complex knowledge and skills, such as practical knowledge. To achieve this, SJTs present test-takers with extended descriptions of challenging situations, usually provided via long passages of text. As illustrated along the left-hand side of the conceptual framework, the heavy reading demands of such tests may discriminate against not only test-takers who are non-native speakers or have limited proficiency in the language of the test, but also native speakers who are not highly competent at processing text, such as those who have reading comprehension difficulties. Figure 3.1 illustrates that test-takers' performance in text-based tests might be affected by factors that are irrelevant to the construct that the test aims to measure. The construct-irrelevant variance introduced by such factors constitutes a severe threat to validity as it can negatively affect the quality of the inferences drawn from test-takers' scores.

The conceptual framework was designed to convey the idea that the use of animated videos, as opposed to written text, could reduce the reading demands of text-based tests and the reliance of test-takers' scores on their language and reading skills. As illustrated along the right-hand side of Figure 3.1, it was hypothesised that by animating the information provided through written descriptions in complex tests (e.g., SJTs), the content of the test items could be more easily communicated to and comprehended by test-takers. Such an approach could lead to assessments that are not so strongly affected by construct-irrelevant factors, such as language and reading skills. If this is the case, the resultant test scores should provide a more accurate measure of test-takers' knowledge and/or skills, ultimately leading to more valid judgments.

Apart from the validity issues related to construct-irrelevant variance, as outlined along the left-hand side in Figure 3.1, the potential impact of animations on test-takers' reactions to the test is another important aspect of every test-taking process that should be taken into account. More specifically, the use of animated videos was expected to have an impact on test-takers' perceptions of the validity, fairness, enjoyment and difficulty of the test, as well as on the effort they invested in it.

As outlined in the conceptual framework, this study compared two different formats of an SJT (i.e., animated vs. text-based) to investigate the impact of animated videos on (i) construct-irrelevant variance and the validity of the inferences drawn from test scores

and (ii) test-takers' reactions to the test. To undertake a robust comparison between a text-based and an animated SJT, a true experiment was conducted. Experimental designs in social sciences reflect post-positivist philosophical assumptions, in which causes determine the possible outcomes in a reality that does exist, but that can be known only imperfectly (Creswell, 2014; Mertens, 2014). Post-positivism represents thinking after positivism, challenging the traditional notion of the absolute truth of knowledge. The knowledge that emerges via a post-positivistic approach is based on attempts to measure as accurately as possible the objective reality that exists in the world, but the researchers modify their claims to understandings of truth based on probability, rather than certainty. The purpose of studies conducted from a post-positivistic point of view is to identify and access the causes that influence outcomes by reducing the "problem ideas" into a parsimonious and discrete set of variables comprising the research questions (Creswell, 2014). Following such a paradigm, the variables of interest were carefully measured to provide evidence of causal relationships and address the research questions of this study.

3.3. Research Questions

Based on the gaps in knowledge identified from reviewing the relevant research literature, two overarching research questions were developed to guide the study. Each of these research questions is presented below, in addition to a number of linked sub-questions.

Research Question 1: What impact does the use of animated videos have on construct-irrelevant variance attributed to language and reading skills?

> ***Research question 1.1.a:*** *Is there a PK-SJT performance gap between native and non-native English speakers?* ***b:*** *If so, is this gap smaller in the case of the animated PK-SJT?*
>
> ***Research question 1.2.a:*** *Is the performance of non-native English speakers on the PK-SJT related to their level of proficiency in English?* ***b:*** *If so, is this relationship weaker in the case of the animated PK-SJT?*
>
> ***Research question 1.3.a:*** *Is test-takers' PK-SJT performance related to their performance on the reading comprehension test?* ***b:*** *If so, is this relationship weaker in the case of the animated PK-SJT?*

Research question 1.4.a: *What proportion of the total variance in PK-SJT scores can be attributed to construct-irrelevant factors?* ***b:*** *Does the use of animations reduce the proportion of variance attributed to construct-irrelevant factors in PK-SJT scores?*

Research Question 2: What impact does the use of animated videos have on test-takers' reactions to the test?

Research question 2.1: *What impact does the use of animated videos have on the face validity of the PK-SJT?*

Research question 2.2: *What impact does the use of animated videos have on test-takers' enjoyment of the PK-SJT?*

Research question 2.3: *What impact does the use of animated videos have on test-takers' perceptions of the difficulty of the PK-SJT?*

Research question 2.4: *What impact does the use of animated videos have on test-takers' invested effort in the PK-SJT?*

3.4. Research Design

This study aimed to explore whether the use of animations in testing can improve assessments by (i) reducing construct-irrelevant variance and (ii) enhancing test-takers' perceptions of and invested effort in the test. To examine these relationships, an experimental design was utilised. Experimental designs are an appropriate way of testing the extent to which an independent variable has an impact on a dependent variable (Creswell, 2012). The implementation of a "true experiment" facilitated the investigation of causal relationships between the independent (test format: animated vs. text-based PK-SJT) and the dependent variables (test-takers' performance on and perceptions of the PK- SJT assessment), whilst controlling for other variables, such as participants' native language and reading comprehension skills.

According to Creswell (2012), the random assignment of research participants to the different conditions of "treatment" is what makes true experiments the most rigorous and strong of all experimental designs[13]. Random assignment is a powerful tool in

[13] Other types of experiments include: pre-experimental, quasi-experimental, single-subject and factorial designs (Creswell, 2014).

creating two statistically equivalent groups that are similar in every way, on both known and unknown variables. In this study, participants were randomly assigned to take either the animated or the text-based version of the PK-SJT, constituting the experimental and the control groups, respectively. Hence, the two groups were expected to be equivalent except for the fact that one group got the intervention (animated PK-SJT) and the other did not. Evidence on the equivalence of the two groups is provided in Chapter 4.

Such a design ensures that any differences between the groups will be due to the incorporation of the animations in the test, as any unknown factors will be similarly distributed between the two groups. Figure 3.2 presents the specific experimental design of this study, following the classic notation system provided by Campbell and Stanley (1963).

Group A (Experimental)	R------X------O
Group B (Control)	R----------O

Figure 3.2. True experimental design (post-test only). Adapted from J. W. Creswell (2014).

Note. R indicates random assignment
X represents an exposure of a group to an experimental variable
O represents the outcome measurements recorded on instruments

Because of the random assignment of individuals, in true experiments, most of the threats to internal validity[14] can be eliminated. Table 3.1, on the following page, presents the different threats to internal validity and how this study controlled for them.

[14] According to Creswell (2012), *internal validity* refers to "the validity of inferences drawn about the cause and effect relationship between the independent and dependent variables" (p. 303).

Table 3.1

Controlling for threats to validity

Type of threat	Description of threat	True experiment
History	A noteworthy event might happen just prior to the experiment that could theoretically influence the results.	Controlled via random assignment.
Maturity	Participants in an experiment may mature or change during the experiment, thus influencing the results.	Controlled via the cross-sectional nature of the study.
Regression	Participants with extreme scores are selected for the experiment. Naturally, their scores will probably change during the experiment.	Controlled via random assignment.
Selection	Participants with certain characteristics that predispose them to have certain outcomes (e.g., they are brighter) may be selected.	Controlled via random assignment.
Mortality	Participants drop out during an experiment due to many possible reasons. The outcomes are, thus, unknown for these individuals.	Controlled via random assignment.
Diffusion of treatment	Participants in the control and experimental groups communicate with each other. This communication can influence how both groups score on the outcomes.	Controlled via prohibiting communication between the participants during the testing process.
Resentful demoralisation	The benefits of an experiment may be unequal or resented when only the experimental group receives the treatment (e.g., the experimental group receives therapy and the control group receives nothing).	The PK-SJT scores did not have any consequences for the participants. Thus, in case the experimental group performed better on the PK-SJT, they would not have any particular gains against the control group of the study.
Compensatory rivalry	Participants in the control group feel that they are being devalued, as compared to the experimental group, because they do not experience the treatment.	This is a potential threat that may lead participants in the experimental group to be more motivated and put more effort into the test. This is something that this study controlled for by measuring the effort that participants put in the assessment (self-report measure).

Note. Information taken from Creswell (2012, 2014).

3.5. Research Participants and Sampling

To explore the use of animations in assessment, this study used a PK-SJT measuring pre-service primary teachers' practical knowledge of how to handle challenging social situations at school (additional information about the measure used is provided in section 3.6.2.). The sample of this study consisted of both pre-service and experienced teachers. Pre-service teachers constituted the main sample to address the research questions of the study, while experienced teachers served as the "expert" group and their responses were only used to inform the subsequent scoring strategy for the PK-SJT (additional information about the scoring of the pre-service teacher responses is provided in section 3.6.2.6.).

3.5.1. Pre-service teachers

The pre-service teacher sample consisted of 129 third-year Bachelor of Education students in the field of primary education. Given that one of the main aspects of the study was to explore the impact of animated testing across both native and non-native English speakers, the study was conducted in institutions from two countries; Ireland (Dublin City University) and Greece (University of Ioannina). This ensured adequate numbers of non-native English speakers and a higher variance in participants' proficiency and reading comprehension skills in English. These two universities were selected for convenience reasons. The cohort of third-year students in both institutions was specifically approached for two main reasons: (i) students at this stage of their studies had some practical experience in teaching due to their school placements, something that enabled them to effectively deal with the PK-SJT and (ii) the fourth-year cohort was not available at the time of data collection due to students' final school placements in both universities. Students in both universities were informed about the research study during their lectures and were invited to volunteer their participation. In total, 51 Irish (84% females) and 78 Greek (85% females) students took part in the research study. All participants from Ireland were native English speakers, while all participants from Greece were non-native English speakers.

Given that the available research literature in the field was very limited, it did not provide robust and reliable indications regarding the expected effect sizes, which could have

further informed about the required sample size for obtaining statistical power at the recommended .80 level (J. Cohen, 1988).

3.5.2. Experienced teachers

The experienced teacher sample was selected following the purposive[15] sampling approach, as teachers were expected to have certain characteristics to participate in the study. To be eligible to participate, teachers were required to have enough teaching experience, so that their responses could be used to inform the subsequent scoring of pre-service teachers' responses. According to Darling-Hammond (2000), the benefits of experience for teachers appear at a level of about five years. Therefore, only primary teachers with five or more years of experience, both from Ireland and Greece, were invited to complete the PK-SJT. The same experience criteria were used by Elliott, Stemler, Sternberg, Grigorenko, and Hoffman (2011) to inform the scoring of the original practical knowledge test in the US.

Teachers from various schools were contacted and informed about the study. The experienced teachers took only the PK-SJT and their ratings on the suitability of every SJT practice for dealing with the problem that was described in the scenario were used to categorise the practices into *Good*, *Bad* and *Neutral*, as explained in section 3.5.2.6. In total, 45 teachers from Ireland and 39 teachers from Greece voluntarily agreed to complete the PK-SJT.

3.6. Measures and Variables

3.6.1. Demographics and level of proficiency in English

Prior to the assessment process, participants were asked to report their biological sex. Additionally, in the pre-service teacher sample, non-native English speakers reported their level of proficiency in English, according to the Common European Framework of Reference for Languages, selecting from a range of categories (i.e., lower than B2, B2, C1, C2). The Common European Framework of Reference for Languages is an

[15] A *purposive sample* is different to a *convenient sample* in that researchers do not simply select whoever is available, but they purposively study cases with certain characteristics that based on previous research or theory, they are expected to provide the data that the researcher needs (Fraenkel, Wallen, & Hyun, 2012). As a non-probability sample, a purposive sample does not represent any group apart from itself and, therefore, findings cannot be generalised to the wider population (Cohen, Manion, & Morrison, 2011).

international standard for describing language ability. In this study, it was used because the majority of Greek people receive English language certification within this framework and would, therefore, be quite familiar with it. Finally, experienced teachers were asked to report the number of years they had been teaching, to ensure that responses from teachers with less than five years of experience could be excluded from the analysis.

3.6.2. The main outcome measure: An SJT measuring practical knowledge

The aim of this study was to explore the use of animated videos, as an alternative to written text, in assessment. To examine this research problem, a practical knowledge test based on items developed by Stemler et al. (2006) was used. As outlined previously, this test was used as an example to facilitate this research. The following sections discuss the reasons why such a test was chosen, and describe in detail the characteristics of this test. It should be appreciated that any other test with similar characteristics could have been used.

3.6.2.1. The selection of a suitable testing instrument for animation

The primary aim of this study was to examine the use of animations as an alternative to text-based testing of complex knowledge and skills. It should be highlighted, though, that 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 added value to be gained from the use of technological applications, in this case, animated videos. In the context of this study, the emphasis was primarily placed on validity and the potential added value of animations was conceptualised in terms of the reduction of construct-irrelevant variance and unintended subgroup differences caused by the heavy reading demands of text-based SJTs. Additionally, given their importance, test-takers' reactions to the animated versus the text-based PK-SJT were carefully examined. Before starting such a time-consuming and expensive process, it was important to identify a text-based test that stood to benefit from the use of animation technology. Such a test should have certain characteristics.

To begin with, the test should go beyond the sole measurement of recall knowledge. The hypothesis that animations can improve assessment is primarily applied to the measurement of more complex knowledge and skills that cannot be adequately captured

by conventional text-based instruments. The incorporation of animations 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, may benefit strongly from the incorporation of animated videos that can facilitate the communication of multi-faceted messages.

Assessment of more 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. One test that uses excessive amounts of written text is the Educational Testing Service (ETS) critical thinking test (https://www.ets.org/s/heighten/pdf/critical-thinking-sample-questions.pdf). However, such practices may introduce construct-irrelevant variance into the assessment, as other factors, such as reading skills, that are not related to the construct that the test aims to capture, may impact on test-takers' performance (AERA et al., 2014). It follows that such tests can potentially be improved by the use of animated videos. Animations could be used to fully or partially replace the long passages of text in the former assessments, with the aim of reducing the heavy reading demands and enhance the accessibility of the assessment.

After exploring the relevant research literature, SJTs were identified as a potentially suitable "vehicle" for examining the effectiveness of animated videos in assessment. SJTs provide test-takers with detailed descriptions of challenging real-life situations accompanied by possible alternative ways of dealing with the given problem (Motowidlo et al., 1990). Therefore, they can potentially benefit from the use of multimedia, as much of the information provided in SJTs could be presented with the use of alternative means, such as videos. Indeed, many of the studies that have attempted to incorporate video technologies in assessment have used SJTs (e.g., Bruk-Lee et al., 2016; Chan & Schmitt, 1997; Lievens & Sackett, 2006). The majority of these studies put their emphasis on the assessment of so-called *soft skills*, such as communication and interpersonal skills (Christian et al., 2010). This seems sensible because the benefits of using multimedia and particularly videos, either acted or animated, in testing can be maximised when they represent human interactions, which are usually hard to describe in texts. Animating these scenarios can also enhance the fidelity of the stimulus, as the way in which these scenarios are encountered in real life is more closely approximated.

Taking all of the above into account, an SJT developed by Stemler et al. (2006) to measure teachers' practical knowledge of how to deal with others was selected as the "vehicle" for answering the research questions of this study. There were a number of reasons for using this particular SJT. First, as a text-heavy instrument measuring practical knowledge, it met the above requirements and was, thus, deemed suitable for animation. Second, Stemler and his colleagues generously provided the researcher with full access to all of the test items. Due to the work involved in developing SJTs, it is very challenging to gain access to a complete instrument. However, Stemler was interested in the research idea and wanted to see how this instrument could further evolve. Furthermore, he was willing to provide support and advice throughout the project.

An additional factor influencing the selection of this instrument was the context to which the scenarios related (i.e., primary school teaching). The researcher had formal training in this area, and this content knowledge facilitated informed decision-making regarding the required adaptations and the animation of the text-based scenarios, as explained in detail later in this chapter and in Appendix A. Finally, the researcher was based in the Institute of Education in Dublin City University, and, thus, had the opportunity to approach large numbers of potential participants to whom the instrument would be relevant (i.e., pre-service teachers).

3.6.2.2. The concept of practical knowledge

Practical knowledge constitutes the tacit aspect of practical intelligence, which in turn is one of the three elements of successful intelligence, as conceptualised by Robert Sternberg. According to the theory of successful intelligence (Sternberg, 1997, 1999), people's intelligence comprises of analytical, creative and practical skills (see Figure 3.3). In practice, successfully intelligent people are those who have developed the skills they need to achieve their own goals within their sociocultural context. People who possess successful intelligence manage to capitalise their strengths and correct their weaknesses using a combination of analytical, creative and practical skills. It is important to mention, though, that successfully intelligent people are not necessarily those with the highest intelligence in any of its three forms (Sternberg, 1997). This implies that, for example, someone may have high levels of creative and analytical intelligence but lack practical intelligence and vice versa.

Practical intelligence, as a component of successful intelligence, is defined as *"the ability to find a more optimal fit between the individual and the demands of the environment through adapting to the environment, shaping or changing it, or selecting a new environment in the pursuit of personally valued goal"* (Sternberg & Grigorenko, 2001, p. 2). It refers to performance in real-world pursuits and helps people cope successfully with problems, constraints and realities of day-to-day life (Sternberg, 1997). Practical intelligence consists of a cognitive and a behavioural component. The cognitive component requires knowledge (both explicit and tacit) about how to deal efficiently with a situation. Explicit knowledge is an outcome of formal training. Tacit knowledge, on the other hand, refers to that kind of knowledge that is not easily articulated as people did not acquire it through formal training but rather through their own experiences. However, knowledge of what to do is only one part of practical intelligence. The other element of practical intelligent is behavioural and is linked to whether someone is actually able to do the right thing in a given situation.

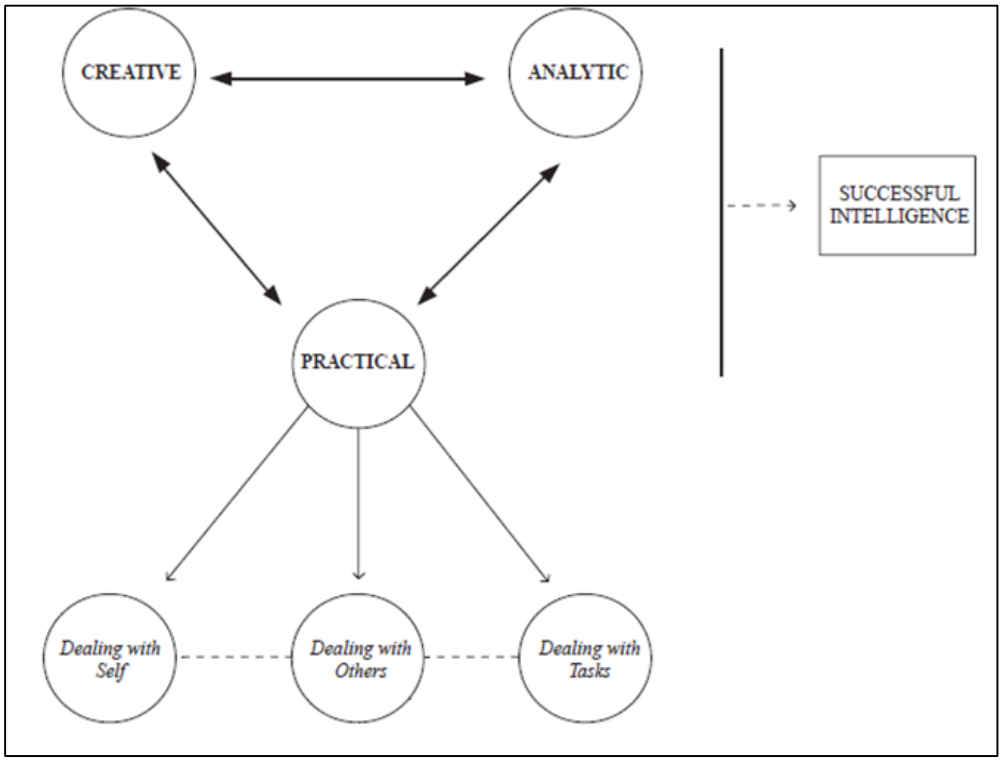


Figure 3.3. Illustration of the framework of successful intelligence. Adapted from Stemler et al. (2006).

According to Sternberg's theory, the everyday life situations that someone might have to address applying their practical knowledge and skills are divided into three categories; (i) dealing with self, (ii) dealing with tasks, and (iii) dealing with others. In situations

dealing with self, such as attempting to reduce personal stress levels prior to an important presentation, self-management skills are required. Dealing with tasks refers to situations where the actions taken to resolve the problem are concentrated around a specific task, for example how to manage the family budget. Finally, in situations dealing with others, such as deciding how to respond to a racist comment made by a colleague, strong interpersonal skills are required (Stemler et al., 2006).

Practical knowledge, as measured in this study, constitutes the tacit element of practical intelligence applied in situations dealing with others, where interpersonal skills are necessary. One of the reasons Stemler and his colleagues decided to explore only the knowledge aspect of practical intelligence was because, as they explained, conducting behavioural assessment is very challenging due to the resources required (Stemler & Sternberg, 2006).

Within the context of teaching, the social aspect of practical knowledge is very important (UNESCO, 2014). Recent research has revealed that social orientation is one of the two major dimensions of effective teaching, with pedagogical orientation being the other (Stemler, Elliott, McNeish, Grigorenko, & Sternberg, 2012), because of the range of challenging social situations that teachers may face in their professional lives. Indeed, teachers are required to deal with three major categories of people in the school context, namely (i) supervisors, (ii) peers, and (iii) subordinates, each of whom may have different demands and needs.

It is critical for teachers to know the appropriate actions they should take once a challenging situation involving one or more of these groups emerges. As Ingersoll (2003) argued, lacking knowledge of how to deal with others is likely to lead to a problematic teaching and learning environment and to quicker professional burnout for teachers. Finally, it has been empirically supported that the social environment that teachers form in their classrooms shapes students' motivation and engagement in the learning process (Ryan & Patrick, 2001). Therefore, it can be concluded that despite the importance of pedagogical knowledge for effective teaching, the possession of practical knowledge of how to deal with others can be considered to be equally important, both in and outside the classroom.

3.6.2.3. Development of the original SJT items measuring teachers' practical knowledge

Although SJTs are particularly popular and efficient, there is not a clear understanding of what exactly they measure, while little emphasis has been placed on developing a theory of why they are so efficient (Lievens & Motowidlo, 2016). Most scholars have concluded that SJTs simultaneously measure a multitude of constructs (Campion, Ployhart, & MacKenzie, 2014). According to Chan and Schmitt (2002), the heterogeneity of SJTs is what makes them such strong predictors of performance, which in nature is a diverse concept. However, the value of discovering the construct(s) that SJTs measure was acknowledged by Motowidlo et al. (1990). In their seminal paper, Motowidlo et al. (1990) argued that, even though most SJTs are not designed to measure any particular psychological construct, it is important to examine what construct(s) are associated with the behaviours measured by these assessment instruments.

As explained in Chapter 2, the lack of firm knowledge of the construct(s) measured by an instrument can negatively affect the quality and the transparency of the assessment. For instance, it may be challenging to legally and professionally defend SJTs that do not explicitly measure a clearly-defined construct (Ployart & Ryan, 2000, as cited in Cabrera & Nguyen, 2001). Furthermore, as Christian et al. (2010) argued, without any knowledge about the construct measured, it is difficult to generalise the use of an SJT beyond the context for which it was developed. Additionally, according to Lievens and De Soete (2015), the heterogeneous nature of most SJTs and the measurement of multiple constructs is responsible for the low internal consistency of these tests.

On the basis of these issues linked to the lack of construct knowledge, Stemler and Sternberg (2006) argued that theoretically-grounded approaches to the development of SJTs are critical in order to more precisely target and assess the constructs of interest. Towards this goal, they highlighted that the theory of successful intelligence could facilitate this process. Thus, by focusing on the practical elements of successful intelligence, Stemler et al. (2006) sought to detect different strategic decisions that teachers make about how to handle challenging situations when dealing with others, having as an ultimate goal to develop an SJT. To do so, they contacted the principals of

243 schools in the US, nominated as National Blue Ribbon Schools[16] for the 2000/2001 school year, inviting them to participate in the project. Stemler et al. (2006) asked the principals who assented to participate in the project to nominate three teachers in their school that they considered to be particularly excellent. Finally, 20 of the teachers who were nominated by their principals agreed to take part in the interviews. Those teachers came from elementary, middle and high schools. Each interview lasted from 60 to 90 minutes.

During the interviews, teachers were asked to describe specific incidents in which they had to deal with challenging situations involving principals, teachers, students and parents, and to describe how they handled these situations. Given the fact that the instruments were focused on the measurement of the tacit aspects of practical skills, teachers described situations for which they had not received formal training. Then, they were asked to think of alternative ways of dealing with these problematic situations. Teachers' responses were analysed to provide information about potential response trends across different situations. As outlined on the following page (Table 3.2), the content analysis led Stemler et al. (2006) to seven practical strategies that applied across a variety of social situations; namely, *comply*, *consult*, *confer*, *avoid*, *delegate*, *legislate and retaliate*.

[16] The National Blue Ribbon Schools Program is a programme under the US Department of Education that recognises outstanding public and non-public schools in the country.

Table 3.2

Key characteristics of the seven strategies for dealing with others

Strategy	Defining characteristics and behaviours	Appropriate use/potential advantages	Inappropriate use/potential disadvantages
Comply	• Actor does whatever is asked of him/her, regardless of who is asking • Actor takes action that can be interpreted as actively condoning behaviours of others in the situation	• Actor agrees with what he/ she is being asked to do • Short-term compliance has long-term benefits (e.g., choose your battles)	• Actor fears emotional consequences of non-compliance • Short-term compliance leads to negative long-term consequences
Consult	• Actor appeals to an external source for advice ● Actor asks people to work together to solve the problem	• Actor wishes to capitalize on other people's expertise	• Actor will be perceived as incapable of solving his/her own problems
Confer	• Actor engages in verbal discussion with source of interaction. Conversation takes place in a private, one-on/to-one setting and is characterized by rational explanation of the actor's point of view	• Actor wishes to increase awareness and communication • People are more apt to change when reasons for requests are revealed	• Revealing too much leaves actor vulnerable to being used as a pawn by others • Rational discussion of each decision takes too much time to be practical
Avoid	• Actor avoids, delays or puts off dealing with a situation or problem • No action is taken at all, or actions that are taken do not deal directly with the situation	• Actor believes that the situation or problem could resolve itself	• Actor avoids action in order to put off emotionally difficult decisions
Delegate	• Actor either implicitly or explicitly delegates responsibility for taking action to someone else • Actor absolves him/herself of responsibility for action	• Actor recognizes his/her own lack of expertise for dealing with situation	• Actor is capable of dealing with situation him/herself
Legislate	• Actor explicates rules governing future actions of self and others	• Actor is interested in procedural justice • A certain class of situations comes up frequently	• Actor creates too many policies • Policies are too situation-specific • Impossible to remember all policies
Retaliate	• Actor reacts physically or verbally in direct response to a situation. Direct response is often like-for-like in nature or involves punishment	• Other strategies have failed • Antagonist does not respond to rational discussion	• Actor retaliates as an instinctive reaction • Actor retaliates as an act of revenge without a strategy for changing antagonist's behaviour

Note. Reprinted from Stemler et al. (2006).

Stemler et al. (2006) argued that, even though in real life each social interaction is unique, these seven strategies for dealing with others provide a framework for evaluating potential courses of action that a teacher could pursue to handle various situations. It should be highlighted that each one of these seven strategies has advantages and disadvantages within the unique context of any interpersonal situation. This implies that there is not a single strategy that is suitable for all the different situations that someone could face. Finally, Elliott et al. (2011) supported that the strategy that teachers select to deal with the given social situations is primarily driven by their professional understanding and expertise, rather than by their personality traits, such as introversion/extroversion, mentioning, though, that further research in this area is necessary. Figure 3.4 illustrates how these seven strategies fit within the framework of successful intelligence.

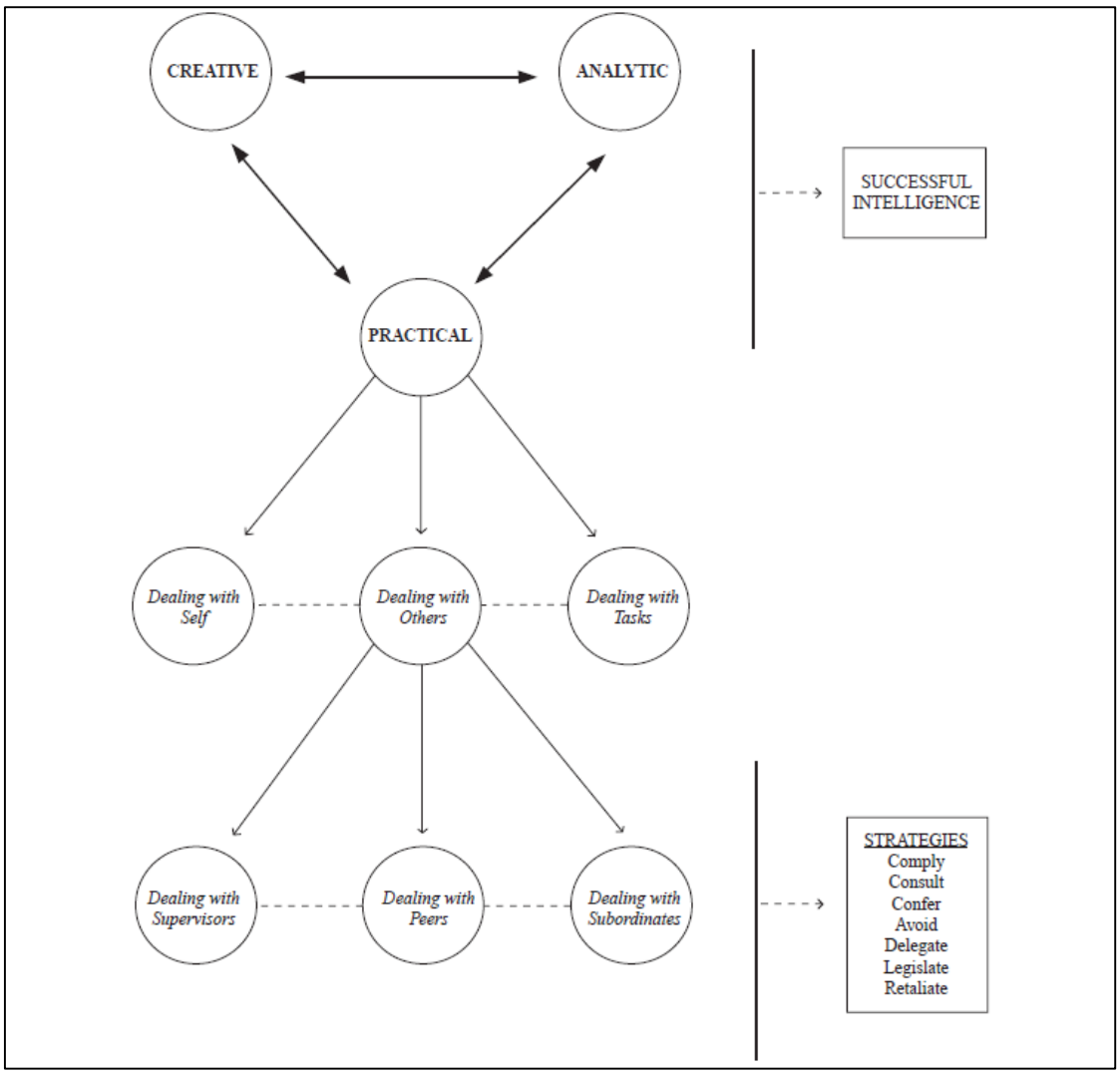


Figure 3.4. The seven strategies for dealing with others within the framework of successful intelligence. Reprinted from Stemler et al. (2006).

Based on the outcomes of the teacher interviews and the resultant seven strategies, Stemler et al. (2006) developed a series of scenarios accompanied by a number of different approaches for dealing with the provided situation (i.e., practice statements). In addition to the teachers' responses, the research team developed additional practice statements to ensure that the options provided in each scenario would cover all seven strategies. In order to decide which strategy is represented by each practice statement, Stemler's research team independently categorised them into one of the seven strategies. Then, the team reviewed the items, resolved areas of disagreement, refined the statements and made the final decisions regarding the strategy that each practice statement represented.

The next step in that process was to ask two independent raters (experienced teachers) who were not involved in the project to categorise each of the practice statements into one of the seven strategies for dealing with others. The results showed good levels of agreement between the project team and each of the raters[17] (Stemler et al., 2006). These results provided evidence to support the hypothesis that the seven response strategies were empirically distinguishable from one another.

The outcome of this process led to the development of three practical knowledge tests (one for elementary, one for middle, and one for high school teachers) in which each scenario was accompanied by seven practice statements representing the seven strategies. In total, 29 unique scenarios were developed with some of these being used across different school levels. Each SJT required test-takers to rate the extent to which they agree or disagree with each of the available practice statements. Each practice statement represented one of the seven strategies. As each one of these seven strategies has advantages and disadvantages within the unique context of any scenario, teachers were asked to rate the practice statements without knowing the strategy it represented. On the basis of these ratings, Stemler and his colleagues developed a number of methods for scoring these SJTs, where experienced teachers' responses served as the basis for informing the subsequent scoring strategy.

[17] The percentage agreement between each of the raters and the development team was consistent across instruments, with a median percentage agreement of 73% for the elementary school SJT, 82% for the middle school SJT and 71% for the high school SJT (Stemler et al., 2006).

These SJTs have been used in a series of research studies with teachers across different countries indicating: (i) criterion validity for predicting teachers' effectiveness (as rated by their principals), (ii) incremental validity over and above age and years of experience, and (iii) moderate to reasonable levels of reliability, which is valuable bearing in mind the low reliability of most SJTs (Elliott et al., 2011; Grigorenko, Sternberg, & Strauss, 2006). It should be mentioned that, although this instrument was initially designed for teachers, it can also be used to examine pre-service teachers' practical knowledge (e.g., Elliott, Stemler, Sternberg, Grigorenko, & Hoffman, 2011).

3.6.2.4. The practical knowledge SJT (PK-SJT) used in this study

The scenarios from each of Stemler and colleagues' SJTs formed the basis for the development of the instrument used in this study. Each scenario was independently reviewed by four people (the PhD candidate, the supervisors and an associate researcher) with several criteria in mind, namely:

- The information in the original scenarios could be communicated with the use of animated videos. Given the fact that SJTs provide test-takers with plenty of information describing real-life situations and interpersonal interactions, most of the scenarios used in the original SJTs could be animated. Therefore, the research team prioritised the items that provided the most complex information and, therefore, could benefit from the use of animations to a greater extent.
- The scenarios should be relevant to a primary school context. This was essential, given that the PK-SJT would be administered to pre-service primary teachers. Any scenarios that were suitable but originally designed for a non-primary context, were adjusted accordingly.
- The final set of scenarios selected should include all of the different groups that primary school teachers may deal with (i.e., students, parents and colleagues). This was important to ensure content validity.

This process led to the creation of the 15-scenario PK-SJT that was used in the current study. The 15 selected scenarios were culturally adapted to suit the cultural contexts of the study. Specifically, the terminology was adapted with terms such as *grade*, which are primarily used in American English, being replaced by terms used in Hiberno and British English, such as *class*. Moreover, the scenarios and characters' names were

edited to achieve a balance between the male and female characters starring in the scenarios and the status of males and females was balanced. Additionally, it was ensured that any subjects referenced in the scenarios were representative of those typically taught in both Greece and Ireland, and gender stereotyping in terms of student performance on those subjects was avoided (e.g., having boys as the being strong in mathematics and girls in reading). All changes that could improve the quality of the PK-SJT scenarios and practice statements were made after obtaining Stemler's consent.

An example of the original and the adapted version of an SJT scenario, along with the seven practices for addressing the situation, are presented in Figures 3.5 and 3.6.

Original scenario

Katie, one of Mrs. White's 3rd grade students, struggles with school. She is always behind the class, no matter how much time Mrs. White spends tutoring her after class. These days Mrs. White is trying to teach Katie multiplication tables. Neither in class, nor during the tutorials, has Katie made any progress. Yesterday Mrs. White received a note from Katie's mother stating that her daughter could not learn the tables, that the experience frustrated Katie, and that Katie cried when she came home from school. The mother added that she did not see any reason for Katie to learn the multiplication tables. Yesterday evening, Mrs. White called Katie's mother and invited her in for a conference. Katie's mother came and they talked, but unfortunately the conversation did not go well at all. Katie's mother persistently asserted that the learning of multiplication tables was damaging to Katie, and therefore Mrs. White should stop forcing Katie to learn them.

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

1	2	3	4	5
Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**

1. Mrs. White should talk to Katie's mother again, explain the importance of mastering multiplication tables, and try to convince her that with help Katie will be able to master them and succeed.
2. Mrs. White should organize a meeting with Katie's parents and the administration of the school.
3. Mrs. White should increase her efforts to ensure that Katie learns her multiplication tables.
4. Mrs. White should ignore Katie's mother's opinions and keep doing what she has been doing.
5. Mrs. White should bring this situation to the principal's attention and ask her to contact Katie's mother to resolve this conflict.
6. Mrs. White should stop tutoring Katie.
7. Mrs. White should institute a policy that requires students to get extra help during recess or after school if their skills in any subject matter are sub-par

Figure 3.5. The original version of a sample scenario along with its response practices.

Adapted scenario (as used in this study)

Ms. Green is the 3rd class teacher, and Katie is one of her students. Katie struggles with learning her multiplication tables and Ms. Green helps Katie by working with her after class. Neither in class, nor during the after school sessions has Katie made any progress. Yesterday, Ms. Green received a note from Katie's mother stating that her daughter could not learn the tables, that the experience frustrated Katie, and that Katie cried when she came home from school. The mother added that she did not see any reason for Katie to learn the multiplication tables. That evening, Ms. Green called Katie's mother and invited her in for a meeting. Katie's mother came and they talked, but unfortunately the conversation did not go well at all. Katie's mother insisted that the learning of multiplication tables was damaging to Katie, and therefore Ms. Green should stop forcing Katie to learn them.

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**

Ms. Green should…

1. talk to Katie's mother again, explain the importance of mastering multiplication tables, and try to convince her that, with help, Katie will be able to master them and succeed.
2. organize a meeting with Katie's parents and the principal of the school.
3. increase her efforts to ensure that Katie learns her multiplication tables.
4. keep doing what she has been doing.
5. bring the situation to the principal's attention and ask him to contact Katie's mother to resolve this conflict.
6. stop helping Katie after class.
7. suggest to the principal that a new rule be introduced that requires students to get extra help during break or after school if their skills in any subject matter are below average.

Figure 3.6. The adapted version of a sample scenario along with its response practices.

Participants were required to rate the extent to which they agreed or disagreed with each of the practice statements using a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). Although both 5-point and 7-point scales have been used in the past by the designers of the instruments, the smaller scale was selected, as research in the field indicates that an increase in the number of options could bias respondents against answers containing the strongest expressions (Wakita, Ueshima, & Noguchi, 2012).

All the 15 PK-SJT scenarios used in the study, each accompanied by its seven response practices, are included in Appendix B.

3.6.2.5. The animated and the text-based version of the PK-SJT

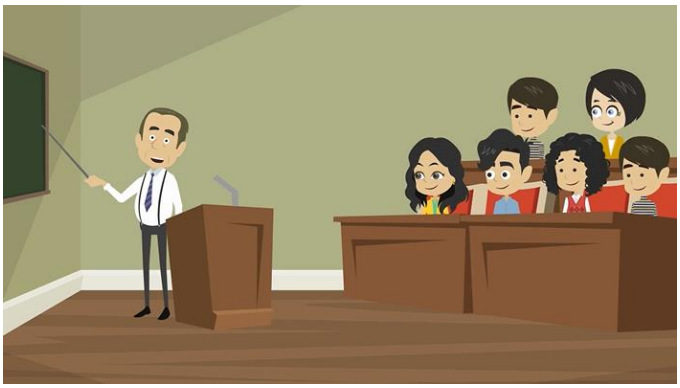
As outlined previously, two versions (i.e., text-based and animated) of the final PK-SJT were compared. The animated version of the test was developed based on the text-based PK-SJT. For the development of the animated videos, the research team worked with an UK-based animation company. After selecting a test that could potentially benefit from the use of animations, a number of decisions regarding the main features of the animated videos had to be made. These decisions concerned the appearance, expressions, voice and movement of the animated characters as well as the characteristics of the animated environments where the scenarios took place (e.g., classroom). A particularly challenging aspect of this process was the animation of elements that were not described in the text-based PK-SJT. Specifically, consideration was given to the following matters:

- gender balance amongst background characters,
- ethnic diversity,
- age diversity of adult characters,
- contemporary classroom environment, and
- characters' reactions and facial expressions.

To ensure the quality of the animations, each time an animated scenario was prepared and submitted by the animation company to the PhD candidate, it was independently reviewed by at least four different people[18] with expertise in educational research. This process, although time-consuming, provided the animators with meaningful and diverse feedback that allowed them to improve the quality of subsequent animations and, thus, the assessment. In total, 45 draft animated scenarios were developed, leading to the final 15 scenarios that met the requirements of the project. To illustrate how important this process was, Figure 3.7 shows the first, second, and final draft of one of the animated scenarios. It is clear that the first two examples are not representative of a modern classroom environment, while the final scenario represents a contemporary classroom, where students from various backgrounds work in groups.

[18] Those people were: the doctoral student, the two supervisors of the project, and an independent research assistant.

First draft



Second draft

Final draft

Figure 3.7. The first, second, and final draft of an animated scenario.

It should be acknowledged that the animation of text-based tests is a relatively new venture in the research literature. Significantly more, but still limited, attempts have been made to incorporate acted videos in scenario-based testing. However, the few studies that have examined the use of videos in assessment have primarily focused on the research problem, providing inadequate information and rationale about the steps followed, the decisions made and the challenges faced during the transformation of a

text-based test into an animated one. Therefore, the development of the animated version of the original text-based PK-SJT was particularly challenging, as no previous literature had discussed the intricacies involved in each step of this process in sufficient detail. Particular attention should be paid to Appendix A which provides a more detailed description of the process of animating the original text-based PK-SJT, the critical decisions that had to be made, the justifications for these decisions based on relevant literature from other fields, the challenges faced and the time required to complete a project of this scale. This documentation of the animation process is a unique aspect of this study that makes a significant contribution to the literature in this area.

As a final point, it could be argued that the animation of written text may actually further standardise an assessment as all test-takers are presented with exactly the same stimuli. In other words, by presenting a given scenario through animation, assumptions and thus, interpretations on behalf of test-takers about the characters and settings involved in the scenario are avoided.

3.6.2.6. Scoring of the PK-SJT

Following Elliott et al.'s (2011) approach (i.e., expert-judgment scoring), the PK-SJT, both in its text-based and its animated version, was initially administered to a number of experienced teachers with five or more years of experience in order to develop an appropriate scoring system. Teachers were asked to rate the extent to which they agreed or disagreed with each of the 105 provided practice statements (i.e., items) using a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree)[19].

In total, 39 Greek and 45 Irish primary teachers completed the PK-SJT. However, only responses provided by experienced teachers (five or more years of experience) who took 15 minutes or more to complete the test were taken into account[20]. This yielded a final sample of 36 teachers from Greece (67% females) and 38 from Ireland (82% females). The vast majority of the teachers constituting the final sample (87%) had at least 11 years of teaching experience.

[19] The PK-SJT consisted of 15 scenarios and each scenario was accompanied by seven practices statements. So, the final assessment consisted of 105 practice statements, which constituted the items of the test.

[20] The minimum time required to properly complete the 15 scenarios was set to 15 minutes, which translates into one minute for each scenario.

Based on experts' responses, the mean for each item was computed. A practice statement was considered as *Good* if it had a mean greater than or equal to four, as four represents the "Agree" option in the 5-point scale. Similarly, a practice was categorised as *Bad* if it had a mean smaller than or equal to two, with two representing the "Disagree" option in the 5-point scale. The rest of the items were categorised as neutral. Table C1 (Appendix C) presents the mean values and the standard deviations for all 105 items.

The scoring of pre-service teachers' responses was further complicated by the cross-cultural element of this study (i.e., the fact that pre-service and practice teachers from two different countries participated). As previous research that used a similar version of the PK-SJT showed, the way that teachers react to some of the provided situations can vary across different cultures (Stemler et al., 2018). In other words, although some approaches to resolve the provided situations can be universally considered *Bad* or *Good*, there are practices that in some contexts may be considered *Good*, but in other contexts not and vice versa. This is something that was taken into account when between-country comparisons, in terms of participants' performance, were conducted.

More specifically, given that native and non-native speakers were treated as cases of the same sample, their responses to the PK-SJT should be evaluated under the same criteria. To achieve this, only those items for which there was not a statistically significant difference between Greek and Irish teachers' ratings were taken into account for the scoring of the PK-SJT. This was achieved by running Mann-Whitney U tests for each of the 105 practices. Based on the statistical analysis results, the responses that Greek and Irish teachers provided were not statistically significantly different for 48 out of the 105 items, ten of which were categorised as *Bad* and six as *Good* practices (Table C1, Appendix C).

In line with this scoring approach, subsequent test-takers' (i.e., pre-service teachers') overall performance on the PK-SJT was comprised of two subscales: (i) their ability to detect the *Good* practices in order to deal with the provided challenging social situations and (ii) their ability to detect and avoid the *Bad* practices in order to effectively resolve the problem. The *Good* subscale for each test-taker was computed by summing the ratings they had provided for each of the items that had been categorised as *Good*, as per the experienced teachers' rating(s). Higher values in the subscale indicated higher ability to detect *Good* practices. A similar approach was followed for the *Bad* practices

subscale. This subscale was computed by summing the ratings that test-takers gave for each of the responses categorised as being *Bad* by the experts. Test-takers' ratings were reversed so that, again, higher values in the subscale indicated higher ability in detecting the *Bad* practices. Test-takers' responses to the items categorised by the experts as *neutral* did not contribute to their overall score.

In order to clarify this scoring approach, an illustrative example is presented below:

- For a practice statement categorised as *Good*, based on experienced teachers' judgments, participant A rated the extent to which they agreed or disagreed with this item, choosing the fourth option "Agree" in the 5-point scale. Thus, participant A would get four score-points (with a maximum of 5) for this response.
- For another practice categorised as *Bad*, the same participant rated the extent to which they agreed or disagreed with the proposed action, choosing the first option "Strongly Disagree" in the 5-point scale. After reversing their response, they got five score-points for this response. In other words, they would get the highest points possible in this scale, because they strongly disagreed that the given practice is a *Good* option, which aligned with experts' ratings.

It should be acknowledged that, in most of the cases, the designers of the original SJT used these two subscales as separate indicators of participants' practical knowledge. However, in this study, these two subscales were combined to create an overall practical knowledge scale. This decision was made for two reasons: (i) to improve the reliability of the practical knowledge measure given that the reliabilities of the two subscales were particularly low[21] due to limited items and (ii) to facilitate a more comprehensive presentation of the results. For the computation of the final score for each test-taker, their scores from the two scales were summed. The final 16-item scale indicated pre-service teachers' overall level of practical knowledge of how to deal with challenging social situations at school. The reliability of this scale was satisfactory, Cronbach's α = .743, with values greater than .70 being considered as acceptable (L. Cohen, Manion, & Morrison, 2011). Possible scores in the scale ranged from 16 to 80.

[21] Cronbach's alphas for the G*ood* and the *Bad* scales were .452 and .705, respectively.

Given the complex nature of this assessment and the fact that there are no objectively correct responses to the scenarios, it should be acknowledged that there are many alternative methods that could have been used to score participants' responses. Some of these approaches are described in Appendix D.

3.6.3. Measures of test-takers' perceptions of and invested effort in the PK-SJT

At the end of the assessment, test-takers were asked to complete a short questionnaire about their perceptions of the PK-SJT; specifically, their attitudes towards the validity, fairness, "enjoyableness" and difficulty of the test. In addition, participants were asked to report the extent to which they had tried hard to do well in the assessment (i.e., their invested effort). The items were rated on a 5-point scale, ranging from one (strongly disagree) to five (strongly agree). A detailed search of the relevant literature was undertaken to find examples of items that had been used to measure these constructs. A decision was made to use items that had been extensively used by other studies in the field of video-based assessment (e.g., Bruk-Lee et al., 2016; Lievens & Sackett, 2006; Richman-Hirsch et al., 2000).

3.6.3.1. Validity and Fairness

The following items were used to measure the perceived validity and fairness of the assessment (i.e., face validity) (adapted from Bauer et al., 2001; Smither et al., 1993):

1. *The content of this assessment was clearly related to the job of a primary school teacher.*
2. *It was clear to me that this assessment is related to someone's ability to deal with challenging social situations that may be encountered in the teaching profession.*
3. *A person who can successfully tackle challenging social situations would do well on this assessment.*
4. *There was **NO** real connection between this assessment and the job of a primary school teacher.*
5. *A person's overall performance on this assessment would predict how well they deal with challenging social situations at school.*
6. *This assessment was a fair indicator of someone's knowledge of how to deal with challenging social situations that may be encountered in the teaching profession.*

7. *This assessment gave me the opportunity to demonstrate my knowledge of how to deal with challenging social situations in teaching.*
8. *This assessment would **NOT** afford everyone the same opportunity to demonstrate their knowledge of how to deal with challenging social situations in teaching.*
9. *This assessment was biased against test-takers who do not have strong language skills in English.*
10. *Overall, the assessment was fair.*

3.6.3.2. Enjoyment

The following items were used to measure participants' enjoyment of the test (adapted from Macan, Avedon, Paese, & Smith, 1994; Smither et al., 1993):

1. *Participation in this assessment was a positive experience.*
2. *This assessment was interesting.*
3. *I did **NOT** enjoy taking this assessment.*

3.6.3.3. Perceived difficulty

The following items were used to measure the perceived difficulty of the test (the items were specifically developed for this study):

1. *I found the content of the assessment difficult.*
2. *I found the language used in the assessment difficult to understand.*

3.6.3.4. Invested effort

The following items were used to measure the effort that test-takers considered they put in the PK-SJT (adapted from Eklöf, 2010):

1. *I gave my best effort on this assessment.*
2. *I worked on each item in the assessment.*
3. *I did **NOT** give this assessment my full attention.*
4. *I could have worked harder on this assessment.*
5. *I did **NOT** try as hard on this assessment as I normally would when taking an assessment at university.*

Finally, two open-ended questions were developed to give test-takers the opportunity to provide some feedback on the assessment. Participants' responses to the open-ended questions were not used to address any of the research questions of this study. Instead, they were used to inform future administrations of the PK-SJT or other animated tests. A summary of participants' feedback based on their response to the open-ended questions can be found in Appendix E.

1. *In general, what do you feel worked well in this assessment, if anything?*
2. *What could be improved in this assessment, if anything?*

3.6.3.5. Exploratory factor analysis of the perception and invested effort statements

To examine whether the above statements could be attributed to a set of underlying factors, an exploratory factor analysis (EFA) was carried out. One of the main questions was whether an overall factor of face validity would emerge and if this would solely consist of statements regarding validity or whether statements about the perceived fairness of the assessment would also be part of such a scale. Although it could be perceived as a distinct characteristic of an assessment process, according to the latest *Standards for Educational and Psychological Testing*, fairness constitutes a fundamental component of validity (AERA et al., 2014). Therefore, test-takers' perceptions of the validity and fairness of the assessment could be perceived as the two main elements of the face validity of the PK-SJT. The EFA was applied to explore this assumption as well.

All variables were included in the EFA, apart from the two difficulty-related statements that were designed to be used as two independent ordinal variables in the analysis[22]. Based on the correlation coefficients among the variables, all 18 variables were statistically significantly correlated to at least one other variable. Having met this assumption, the EFA was applied using all 18 variables.

There are many "rules of thumb" regarding the minimum sample size required for EFA. A common rule is that researchers need at least five to 20 cases per variable included in the analysis (Field, 2017; Stevens, 2009). On the other hand, empirical studies in the

[22] These two statements examining the perceived difficulty of the (i) content of and (ii) language used in the assessment were expected to be independent from each other and not necessarily related.

area suggested that the differences in participant-variable ratio alone have little impact on the stability of the factors (Arrindell & Van der Ende, 1985) and that the adequacy of the sample size should be evaluated on the basis of the study design, communalities and factor loadings (Guadagnoli & Velicer, 1988; MacCallum, Zhang, Preacher, & Rucker, 2002). The sample size of the study was $N = 129$, with a ratio of 7.2 cases per variable. From a statistical point of view, the Kaiser-Meyer-Olkin measures of sampling adequacy (KMO) values was .697, with values greater than .5 being acceptable.

The principal axis factoring estimation approach and the oblimin rotation method were used to extract the factors. In contrast to other popular factor extraction approaches, such as the maximum likelihood approach, principal axis factoring does not require multivariate normality (Fabrigar, Wegener, MacCallum, & Strahan, 1999), an assumption that was not met by the ordinal data of this study. The direct oblimin method was selected as some of the resulting factors were expected to correlate with each other, given that they all measured different aspects of test-takers' reactions to the test (Field, 2017). Bartlett's Test of Sphericity was statistically significant ($p < .001$), providing evidence to reject the null hypothesis of no correlation among the variables and thus, continue with the interpretation of the results.

Five extracted factors had eigenvalues over Kaiser's criterion of 1 and in combination explained 43.6 % of the total variance. As expected, the items measuring invested effort and enjoyment clearly loaded on two separate factors; factor 1 and factor 3, respectively. The variables loading on these factors were used to create the final invested effort and enjoyment scales. The pattern matrix showing the factor loadings after rotation is presented in Table F1 (Appendix F).

The findings regarding the items measuring participants' perceptions of the validity and fairness of the assessment were less clear cut. These items did not clearly load on two separate factors, namely attitudes towards test validity and fairness, respectively. Instead, there were both validity- and fairness-related statements that loaded on more than one common factor. The interplay between validity and fairness statements has been corroborated by the latest *Standards,* according to which fairness is an integral part of validity and is probably an indication that these two constructs should not be treated separately (AERA et al., 2014).

To explore the face validity factor structure in more depth, only the ten validity and fairness statements were included in an EFA. In this case, the unrotated results were interpreted, as the aim was to extract the simple solution (Table F2, Appendix F). Apart from item number one, the rest of the fairness and validity items loaded satisfactorily on one common factor, rather than two different ones. On this basis, it was decided that item one (i.e., The content of this assessment was clearly related to the job of a primary school teacher) be excluded from the face validity scale because, in addition to the low loading on the common factor, it also had particularly low variance. Item one was very similar to item four, therefore, its exclusion from the scale was unlikely to have any consequences on content validity. The remaining nine items contributed to one overall scale, representing the face validity of the PK-SJT.

The final scales for each factor were computed by taking test-takers' average scores on the relevant items. Scores in the scales could range from one to five. Table 3.3 presents the reliability levels for each scale (Cronbach's alpha values).

Table 3.3

Reliability levels of the perceptions and invested effort scales

Scale	Cronbach's alpha
Face validity	0.627
Enjoyment	0.724
Invested effort	0.772

3.6.4. Reading comprehension

An English language reading comprehension test was used to explore the extent to which the animation of the PK-SJT may mitigate the dependency of test-takers' scores on their reading skills. Specifically, retired items of the Graduate Record Examinations (GRE) general test, developed by the ETS, were used. The GRE General Test is a graduate-level admissions test that has been designed to be taken by prospective graduate school applicants from around the world. As this study involved both native and non-native English speakers studying in higher education contexts, the GRE reading comprehension test was deemed to be a suitable choice of instrument.

The GRE General Test measures verbal reasoning, quantitative reasoning and analytical writing skills. The verbal section of the test includes antonyms, analogies, sentence

completions and reading comprehension questions. As participants' level of reading comprehension in English was the main variable of interest, for the purposes of this research, only items measuring reading comprehension were used. The reading comprehension items constitute a discrete section in the GRE test measuring test-takers' ability to (i) read with understanding, insight and discrimination and (ii) analyse a written passage from several perspectives. Passages are drawn from many different disciplines and sources. The final test that was used in this study consisted of three passages of text with 11 reading comprehension items, in total. The passages used in this study were selected to offer questions with a wide range of difficulties, as reported by the ETS. The test was administered on paper and participants had 25 minutes to read the passages and answer the multiple-choice questions. It should be acknowledged that the internal consistency of the reading comprehension scale was low, Cronbach's alpha of .52, something that may be attributed to the restricted number of items. The reading comprehension test used in this study is presented in Appendix G.

3.7. Design of a New Testing Platform

After finalising the two versions of the PK-SJT, the next step was to find an online platform that could best facilitate the administration of the assessment and the perceptions survey. The following were identified as key requirements for creating an assessment environment that would be engaging and take full advantage of the animation technology:

- The platform should support the high definition animated videos and the audio files that would accompany the practice statements.
- Test-takers should be able to give their responses one by one in each scenario after watching the animated scenario, reading the practice statements and listening to the relevant voiceovers.
- Test-takers should have the option to watch the video again at any point while dealing with the given scenario.
- The videos and the statements should be easily accessible by the test-takers with minimum scrolling, as recommended by the relevant research literature (Bridgeman, Lennon, & Jackenthal, 2003; Sanchez & Goolsbee, 2010).

- The assessment should be available only to the selected sample of the study and not to anyone who has access to the URL link of the research, which is usually the case with online survey platforms. The aim was to create a simulation of a testing experience and for test-takers to be given individual usernames and passwords to access the test such that their responses could be linked to their performance on the reading test.

It was challenging to find a tailor-made platform for this assessment as off-the-shelf versions of commercial platforms, such as eSurvey Creator and SurveyMonkey, did not meet these requirements. Ultimately, a fortuitous encounter with a representative from a software company (*Psycholate*) at a testing conference led to the development of the bespoke platform that was eventually used in this study. *Psycholate* were commissioned to work on a trial basis in order to produce a platform for the purposes of the project – a task completed over a period of two months, informed by ongoing consultations with the researcher. The final product met all the requirements outlined above. The platform was used to host both the text-based and animated versions of the test and provided test-takers with a user-friendly environment to complete the assessments. Figure 3.8 provides four screenshots of the platform presenting sample of both the animated and the text-based version of the PK-SJT.

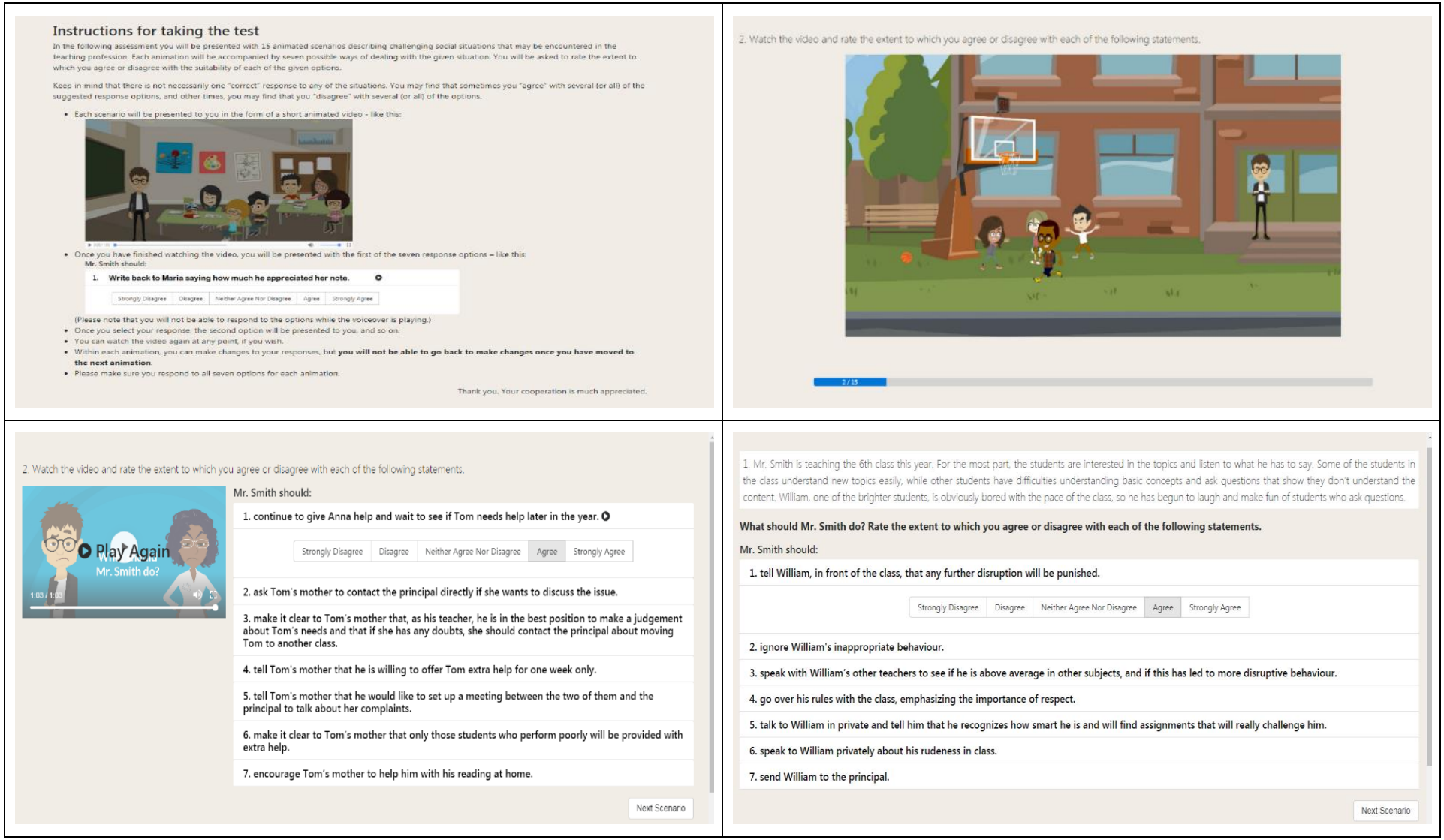


Figure 3.8. Examples of the Psycholate platform environment.

3.8. The Pilot Studies

Prior to the main study, two small-scale pilots were conducted; these are described in detail below.

3.8.1. Pilot one: The animated videos

The first pilot took place right after the completion of the animation process, in November 2017. The aim of this pre-pilot was to make sure that the quality of the animations both in terms of content and technical characteristics was sufficient and that none of the elements of the animations would jar with potential test-takers. The 15 scenarios, along with their practice statements, were administered to 11 educators and researchers in both countries. The feedback from these 11 experts was constructive and provided many useful ideas to improve the animated videos. However, there were not any issues severe enough to warrant further editing or exclusion of any of the scenarios. This was a necessary step to ensure the quality of the assessment and to reduce or eliminate any potential sources of unintended noise or perceived bias, before administering it to a group of test-takers for the pilot study.

3.8.2. Pilot two: Full pilot

After ensuring the quality of the animated videos, the main pilot study took place in both countries during the months of January and February in 2018. A sample of four experienced and five pre-service teachers from Ireland and Greece agreed to participate in the pilot study. The purpose of the pilot study was to:

- identify potential issues in the computer-based administration of the PK-SJT and the survey,
- identify any problems with the items that were developed or adapted to measure test-takers' perception,
- confirm the length of the testing process, and
- collect feedback about the PK-SJT.

Due to the limited number of participants in the pilot study, statistical analysis of the collected data was not conducted. Instead, the emphasis was placed on the functionality of the instruments and participants' responses to the open-ended questions regarding

aspects of the assessments that worked well or could be improved. Participants were also approached to individually share their experience regarding the assessment. According to participants' feedback, the instructions and the assessment tasks were very clear. They reported enjoying the content of the assessment and the technology used. Issues regarding the online administration of the test were not reported. The main issue that concerned the pre-service teachers was the nature of the reading comprehension test. They reported that they did not enjoy completing the test and they could not see its relevance to the research project. Non-native speakers, in particular, reported that the reading test was quite challenging. To address this issue, in the main study, participants were informed about the important role that the reading comprehension test played in the research as potential predictor of the PK-SJT scores and were encouraged to do their best. Finally, some typos detected by the participants were corrected.

3.9. The Main Study

3.9.1. Pre-service teachers

The study took place in one Irish and one Greek University and, as such, two separate data collections were organised. The students who agreed to participate in the study completed the tests and the survey in computer labs that served as testing centres. Upon their arrival, students were provided with login details that randomly assigned them to either the animated or the text-based PK-SJT. Random assignment led to roughly equal numbers of participants across the two groups. Participants were required to log in to the assessment using their username and password. Both the PK-SJT and the post-test survey were administered via computers.

First, participants were administered the PK-SJT and the post-test survey measuring their test experience (i.e., invested effort and perceptions of validity, fairness and enjoyment). As explained above, the aim was to examine the extent to which animations could facilitate assessments for both native and non-native speakers. Therefore, both native and non-native English speakers took the PK-SJT in English. However, participants were expected to better express their opinion about the test experience in their mother tongue and, therefore, the post-test survey was translated into Greek for the non-native English speakers. This process was expected to take approximately 50 to 60 minutes (45 minutes for the PK-SJT and 10 minutes for the post-test survey).

Following the PK-SJT and the survey, students were administered the paper-and-pencil reading comprehension test. Students had 25 minutes to read the three passages of text and answer all 11 questions. In total, the whole process lasted approximately one and a half hour.

Recruiting participants for this study was particularly challenging. The number of eligible participants was restricted due to the specific characteristics required (as explained above). Furthermore, it is likely that the significant amount of time required to complete the two assessments and the questionnaire discouraged some pre-service teachers from participating. The relatively low numbers of participants led to an extension of the data collection period in Ireland, with pre-service teachers being invited to participate in the project until the very end of their semester. The main data collection with pre-service teachers was conducted during the months of March, April, and May in 2018.

3.9.2. Experienced teachers

Teachers from Ireland and Greece, who agreed to participate in the research study, had to complete the PK-SJT only; they did not have to take the survey and the reading comprehension test. Pre-service teachers completed the PK-SJT in their mother tongue so as to be able to fully comprehend the situations and provide their informed responses. As explained earlier, their responses were used to inform the scoring of student teachers' responses. The PK-SJT that these experienced teachers completed incorporated both the animated and the text-based version of the test so that teachers would not favour any of the two formats with their responses. Additionally, the aim was to provide teachers with as much information about the situations as possible before rating the suitability of each practice statement.

The main data collection with experienced teachers was conducted during the months of March, April, and May in 2018.

3.10. Ethical Considerations

This was a low-risk research project and ethical approval was obtained from the Dublin City University Ethics Committee. There were no expected risks for participants from taking part in the study greater than those encountered in everyday life. None of the measures used was expected to pose any great difficulty to the participants and personal

or sensitive information was not requested. Participation in the study was voluntary and confidential and each participant signed an informed consent form prior to their participation. Participants were free to withdraw from participation at any time during the study, without providing any justification. After the completion of the data collection, participants were verbally debriefed and informed about how they could access the results of the study when these would be available. A copy of the ethics approval letter is included in Appendix H.

3.11. Summary

This chapter presented the methodology of this study. It provided information about the conceptual framework and how it informed the experimental design, through which the animated and the text-based PK-SJT were compared. The measures used, the sampling and the procedures followed in the study were discussed in detail. In the next chapter, the results of this research study are presented.

Chapter 4: Results

4.1. Introduction

In this chapter, the results of the study are presented. Firstly, the sample demographics, including evidence pertaining to the equivalence of the control and experimental groups are provided. Following this, information about students' performance on the PK-SJT is given. Finally, the main results of the study are reviewed, structured according to the research questions. The first set of research questions focused on the role of animations in reducing construct-irrelevant variance and unintended subgroup differences attributed to language and reading skills. The second set of research questions explored test-takers' reactions to the two test formats.

4.2. Demographics, Descriptive Statistics, and Performance on the Reading Comprehension Test

A total of 129 pre-service teachers from Ireland and Greece participated in the study. The participants were randomly assigned to take either the animated or text-based version of the same PK-SJT, creating two groups of approximately equal size. These two groups formed the experimental and the control groups of this research study, respectively. Table 4.1 summarises the demographic information about both groups.

Table 4.1

Demographic information about the sample of the study

	Experimental Group - Animated PK-SJT (n = 66)		Control Group - Text-based PK-SJT (n = 63)	
	Native English speakers (n = 26)	Non-native English speakers (n = 40)	Native English speakers (n = 25)	Non-native English speakers (n = 38)
Females	22	32	21	34
Males	4	8	4	4

About 60% of the participants were from Greece (i.e., non-native English speakers). The remaining 40% were Irish (i.e., native English speakers). The native and non-native

English speakers were almost evenly distributed across the control and the experimental group. The majority of participants in both Ireland and Greece were females. Taking into account the over-representation of females in the teaching profession in Europe, especially at the primary level (Eurostat, 2016), this was to be expected.

Prior to the assessment, non-native English speakers were asked to report their level of proficiency in English. As Figure 4.1 shows, the majority of participants categorised themselves as independent users of English at level B2 (56.4%), while 30.8% of people reported that they were proficient in English (level C2). For the purposes of the bivariate and multivariate analyses, two English proficiency groups were formed. The first group comprised of non-native English speakers who reported having an English level of C1 or above (advanced English speakers - 35.9%), while the second group comprised of non-native speakers who reported having an English level of B2 or below (non-advanced English speakers - 64.1%). In the experimental group, 32.5% of non-native English speakers belonged to the advanced English group, while this percentage was 39.5% for the control group.

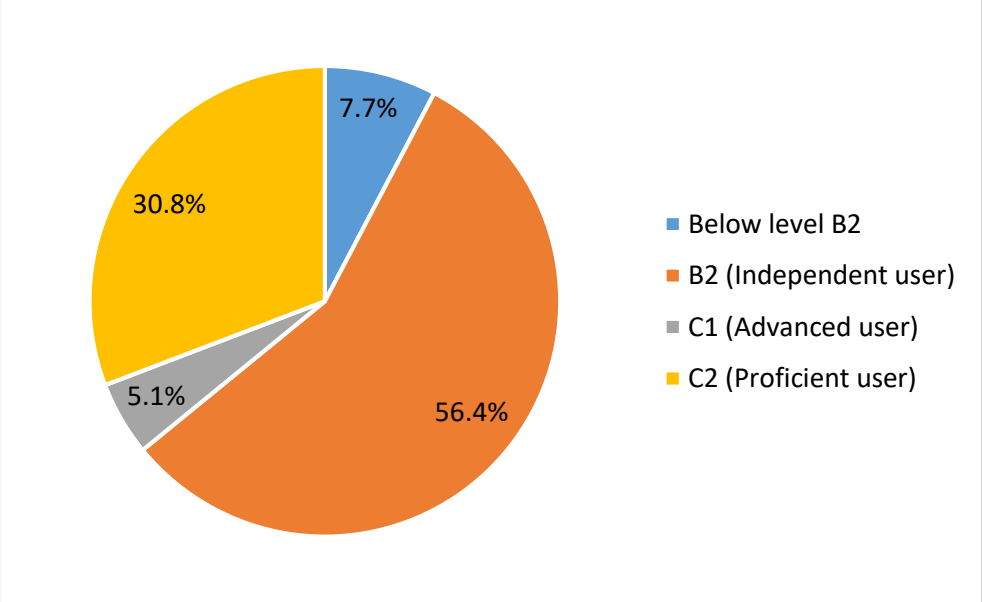


Figure 4.1. Non-native English speakers' level of proficiency in English.

Table 4.2 summarises participants' performance on the reading comprehension test. Overall, participants performed relatively poorly on the reading comprehension test (M = 3.84, SD = 2.01)[23]. As expected, native English speakers performed statistically

[23] The reading comprehension scale ranges from 0 to 11.

significantly better than their non-native peers, $t(127) = -6.137$, $p < .001$. The effect size of this difference ($d = 1.11$) exceeds J. Cohen's (1988) convention for a large effect ($d = 0.80$). Additionally, results of the Spearman's Rho correlation indicated that there was a statistically significant positive association between non-native English speakers' reading comprehension ability and their level of proficiency in English, $r_s(78) = .31$, $p = .006$.

Table 4.2

Performance on the English reading comprehension test

	Minimum	*Maximum*	*M*	*SD*
Overall	0	9	3.84	2.01
Subgroups				
Native English speakers	1	7	5.02	1.73
Non-native English speakers	0	7	3.06	1.80

4.3. Equivalence Between the Experimental and the Control Groups

Despite the random assignment, it is acknowledged that minor differences between the control and the experimental groups in terms of participants' background were likely to exist. With this in mind, a series of statistical tests were conducted to investigate the extent to which this was the case. Bivariate analyses indicated that there were no statistically significant differences in participants' native language, $\chi^2(1, N = 129) = 0.001$, $p = .973$, gender, $\chi^2(1, N = 129) = 0.740$, $p = .390$, proficiency in English, $U = 745.00$, $p = .866$, $\chi^2(1, N = 129) = 0.421$, $p = .638$, and reading comprehension ability, $t(127) = -.240$, $p = .811$, between the experimental and the control groups. Consequently, it could be argued that the participants who completed the animated version of the test were not different from those who completed the text-based version, at least in terms of the background characteristics that were of interest to this study.

4.4. Test-Takers' Performance on the PK-SJT

Pre-service teachers' average performance on the PK-SJT ($M = 65.64$, $SD = 6.02$) indicated that, overall, participants performed well on the practical knowledge assessment, scoring at the upper end of the scale[24], on average. By comparing the two

[24] The practical knowledge scale ranged from 16 to 80.

versions of the test, it was found that participants who completed the animated PK-SJT performed significantly better (M = 66.67, SD = 5.41) than those who took the text-based version (M = 64.57, SD = 6.47), $t(127) = 2$, $p = .048$. The magnitude of the difference between the two means was small to medium, $d = 0.35$. Table 4.3 presents these results along with the minimum and maximum values achieved by the overall sample and each treatment group.

Table 4.3

Performance on the PK-SJT

	Minimum	*Maximum*	*M*	*SD*
Overall	46	77	65.64	6.02
Treatment groups				
Animated PK-SJT	49	77	66.67	5.41
Text-based PK-SJT	46	76	64.57	6.47

It should be acknowledged that the existence of statistically significant performance differences in favour of those who completed the animated PK-SJT does not, on its own, necessarily indicate that the use of animations enhanced the quality and validity of the inferences from test-takers' performance. To explore these aspects in more depth, the following research questions were addressed.

4.5. Research Question 1: What Impact Does the Use of Animated Videos Have on Construct-Irrelevant Variance Attributed to Language and Reading Skills?

The first set of research questions examined whether animations managed to reduce construct-irrelevant variance and improve assessment validity by mitigating potential subgroup differences arising from test-takers' language and reading comprehension skills.

Research question 1.1.a: *Is there a PK-SJT performance gap between native and non-native English speakers?*

Research question 1.1. consisted of two parts (i.e., 1.1.a and 1.1.b). The first part (1.1.a) sought to investigate the potential performance gap on the PK-SJT between native and non-native English speakers that may be attributed to factors other than participants'

practical knowledge (e.g., language and reading comprehension skills), independently of the test format that they completed. To examine this, an Independent-Samples T-test was applied. The results indicated that native English speakers (M = 68.84, SD = 3.97) statistically significantly outperformed non-native speakers (M = 63.55, SD = 6.22) on the PK-SJT, $t(127) = 5.894$, $p < .001$. The effect size of this performance gap was particularly large, $d = 1.01$. To further examine the nature of this gap and the potential role of animations in reducing it, the research question 1.2 was addressed.

Research question 1.1.b: *Is this performance gap between native and non-native English speakers smaller in the case of the animated PK-SJT?*

To investigate the effectiveness of animations in reducing the performance gap between native and non-native English speakers on the PK-SJT, two Independent-Samples T-tests, one for the control and one for the experimental group were conducted. The results indicated that native English speakers significantly outperformed their non-native peers both in the text-based and the animated PK-SJT; text-based PK-SJT: $t(61) = -3.703$, $p < .001$, animated PK-SJT: $t(63) = -4.555$, $p < .001$. However, in the case of the animated assessment, the average performance gap between Irish and Greek participants was smaller (mean difference of 4.99 score-points) compared to the performance gap in the text-based PK-SJT (mean difference of 5.62 score-points). These differences are illustrated on the next page (Figure 4.2).

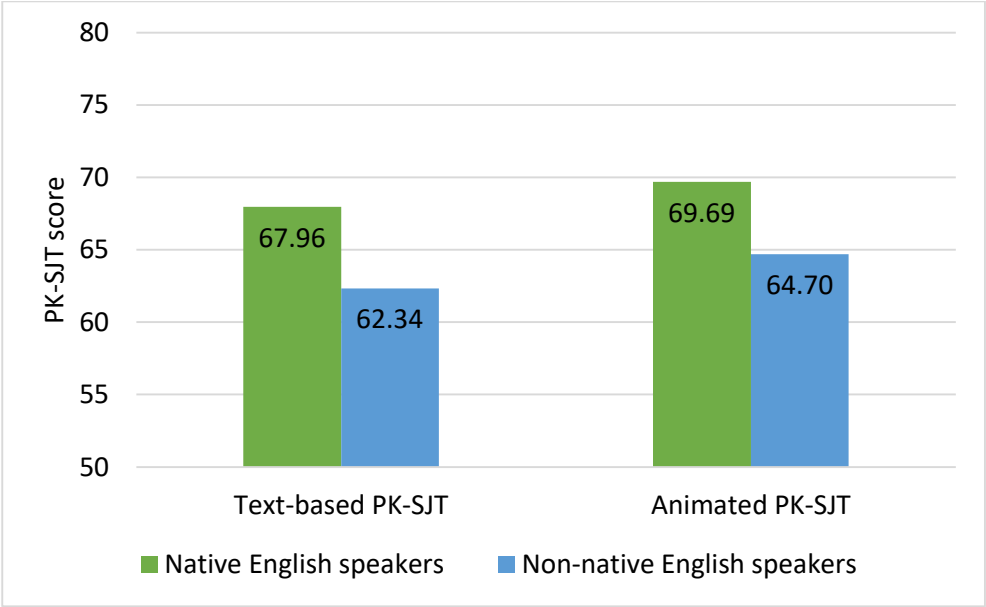


Figure 4.2. Native and non-native English speakers' performance across test formats.

Although this could be regarded as evidence supporting the idea that the use of animations can reduce unintended subgroup differences, the statistical significance of the impact of animations on reducing construct-irrelevant variance had to be tested. To examine whether the performance gap was significantly smaller in the case of the animated assessment, a hierarchical multiple linear regression with participants' PK-SJT scores as the outcome variable was conducted. *Test format* (animated/text-based PK-SJT) and *Native language* (native/non-native speaker) were entered as predictor variables in step one, whilst the *Test format*Native language* interaction term was added in the model in step two, accounting for the original variables.

The regression results indicated that, although test format and test-takers' native language were statistically significant predictors of their PK-SJT performance, explaining 21.7% of the total variance (R^2) in the outcome variable, the interaction between the two variables was not statistically significant and it did not further contribute to the variance explained. Based on the standardised coefficients (β) of the models (as summarised in Table 4.4) native language was a stronger predictor of PK-SJT performance than test format.

Table 4.4

Test format - Native language regression model

Predictors	*B*	*SE B*	β	R^2
Step 1				.217
Test format	2.111*	0.945	0.176	
Native language	5.298**	0.966	0.432	
Step 2				.218
Test format*Native language	-0.626	1.940	-0.041	

Note. The analysis was based on a sample of 129 test-takers.
*$p < .05$. **$p < .01$.

According to the unstandardised coefficients (*B*) of the models, the participants who took the animated version of the test were predicted to perform better in the PK-SJT by 2.11 score-points compared to those completing the text-based version of the test. Similarly, with other variables held constant, native English speakers were expected to perform 5.30 score-points higher than non-native speakers. The interaction of the two variables was negative, indicating that the performance gap in favour of native English speakers was expected to be smaller rather than larger in the case of the animated

assessment, however, this effect was not large enough to be statistically significant. It should be acknowledged that, due to the limited sample size, the power of the regression models might not be adequate for detecting small effect sizes. More details on this topic are provided in the limitations section (5.4).

Research question 1.2.a: *Is the performance of non-native English speakers on the PK-SJT related to their level of proficiency in English?*

To further examine the nature of non-native English speakers' lower PK-SJT performance, additional analysis using a different measure of English proficiency was conducted. As mentioned above, based on non-native English speakers' responses regarding their level of proficiency in English, two groups of English proficiency were formed: (i) the advanced English speakers (i.e., participants who had English level of C1 or above) and (ii) the non-advanced English speakers (i.e., participants who had English level of B2 or below) (see Figure 4.1). To answer this research question, an Independent-Samples T-test comparing these two groups in terms of their PK-SJT performance was applied. The results showed that advanced English speakers (M = 66.54, SD = 5.01) performed statistically significantly better than non-advanced English speakers (M = 61.88, SD = 6.25), $t(76) = -3.376$, $p = .001$. The effect size of this gap was large, d = 0.80. This finding provided some further support for the existence of construct-irrelevant variance favouring those with stronger English language skills.

Research question 1.2.b: *Is the relationship between non-native English speakers' performance and their level of proficiency in English weaker in the case of the animated PK-SJT?*

Running separate analysis for those who took the animated and the text-based version of the test, it was found that the average performance difference between advanced and non-advanced non-native English speakers, although statistically significant in both formats, was smaller in the case of the animated PK-SJT (mean difference of 4.20 score-points) versus the text-based PK-SJT (mean difference of 5.49 score-points); animation PK-SJT: $t(38) = -2.325$, $p = .026$, text-based PK-SJT: $t(36) = -2.698$, $p = .011$. These differences are illustrated in Figure 4.3.

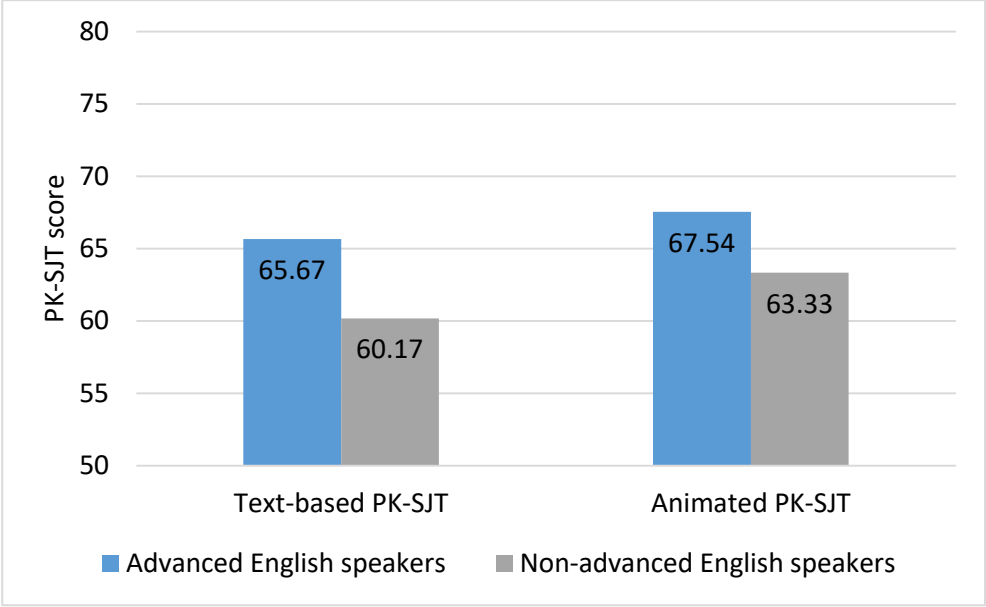


Figure 4.3. Advanced and non-advanced non-native English speakers' performance across test formats.

To examine whether the performance gap between advanced and non-advanced non-native English speakers was significantly smaller in the case of the animated assessment, a hierarchical multiple linear regression with participants' PK-SJT as the outcome variable was conducted. *Test format* and *English proficiency* (advanced/non-advanced English speaker) were entered as predictor variables in step one, whilst the *Test format*English proficiency* interaction term was added in the model in step two, while accounting for the original variables.

The regression showed that test format and proficiency in English were statistically significant predictors of non-native English speakers' PK-SJT scores. However, the interaction between these two predictors was not statistically significant (Table 4.5).

Table 4.5

Test format - Proficiency in English regression model

Predictors	*B*	*SE B*	*β*	R^2
Step 1				.178
Test format	2.697*	1.299	0.218	
Proficiency in English (Advanced)	4.860**	1.354	0.377	
Step 2				.180
Test format*Proficiency in English	-1.288	2.722	-0.078	

Note. The analysis was based on a sample of 78 test-takers.
*$p < .05$. **$p < .01$.

On average, the non-native pre-service teachers who completed the animated version of the test were expected to perform 2.70 score-points higher than those completing the text-based assessment. Additionally, advanced non-native English speakers were predicted to outperform their non-advanced English-speaking peers by 4.86 score-points on average. Based on the standardised coefficients (β), proficiency in English was a stronger predictor of PK-SJT performance than test format. Even though the animation of the text-based test managed to reduce the performance gap between advanced and non-advanced English speakers, the interaction effect was not large enough to be statistically significant.[25]

Research question 1.3.a: *Is test-takers' PK-SJT performance related to their performance on the reading comprehension test?*

As outlined in Chapter 2, high reading demands may cause difficulties not only for test-takers who are non-native speakers, but also, generally, for people who are not highly competent at processing text. Research question 1.3. (1.3.a and 1.3.b), thus, aimed to provide evidence of whether the potential construct-irrelevant variance that may be attributed to test-takers' reading comprehension can be mitigated through the use of animated videos.

For the first part of this research question (1.3.a) and without taking into account the test format, Pearson correlation analysis showed that there was a statistically significant, moderate, and positive correlation between participants' performance on the PK-SJT and their reading comprehension score, $r(129) = .482$, $p < .001$. Test-takers with stronger reading comprehension skills tended to perform much better on the practical knowledge assessment. Pre-service teachers' reading comprehension skills were a good predictor of practical knowledge performance, explaining 23.2% of the variance in their PK-SJT scores.

By investigating this correlation within the two subgroups of the study (i.e., native and non-native English speakers), it was found that the positive correlation between reading comprehension and PK-SJT performance was statistically significant only in the case of non-native English speakers, $r(78) = .422$, $p < .001$, while native pre-service teachers'

[25] Subgroup analysis results should be interpreted with more caution because of the smaller number of cases in each group, which may lead to reduced power. A detailed discussion of the statistical power of this study is presented in Chapter 5.

performance was not affected by their reading comprehension ability, $r(51) = .181$, $p = .203$[26].

Research question 1.3.b: *Is the relationship between test-takers' PK-SJT performance and their performance on the reading comprehension test weaker in the case of the animated PK-SJT?*

To examine whether the animations managed to reduce the reading demands of the original text-based assessment and decrease the impact of reading comprehension skills on test-takers' PK-SJT performance, the correlation between performance and reading comprehension was examined separately for each test format group. The results of the statistical analysis showed that the correlation between performance and reading comprehension was weaker in the case of the animated PK-SJT, but it remained statistically significant in both cases, explaining a significant amount of the variance in test-takers' performance (R^2); animated PK-SJT: $r(66) = .469$, $p < .001$, text-based PK-SJT: $r(63) = .499$, $p < .001$. Figure 4.4 illustrates the difference between the two correlations, showing a slightly steeper slope in the case of the text-based PK-SJT.

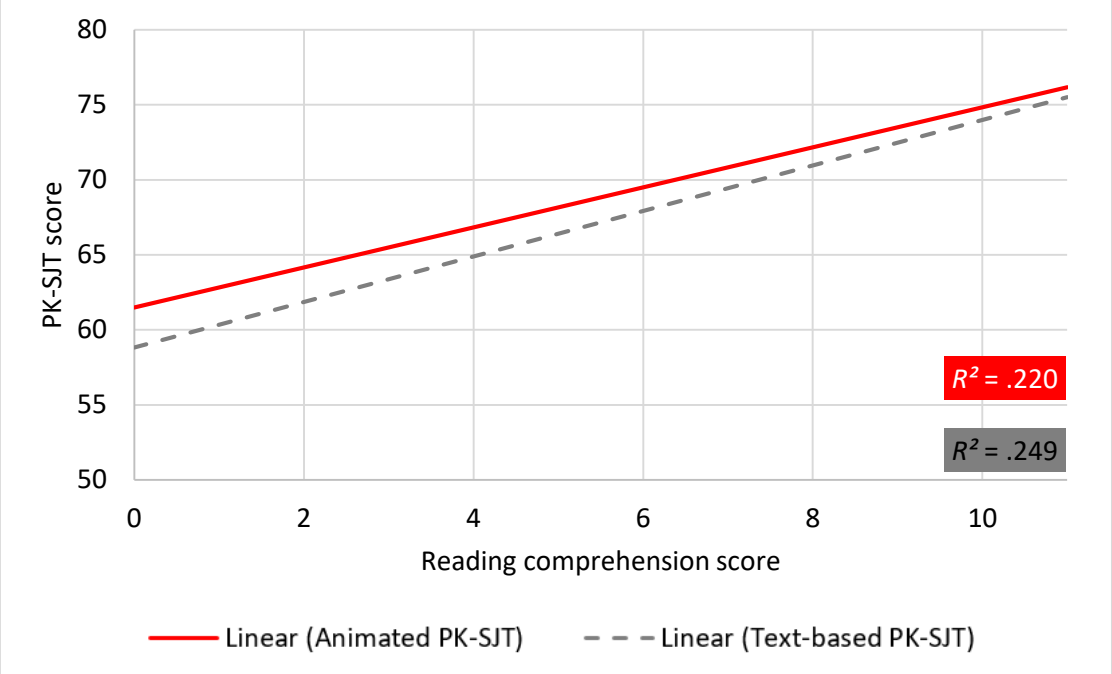


Figure 4.4. Correlation between reading comprehension and PK-SJT performance.

[26] Subgroup analysis results should be interpreted with more caution because of the smaller number of cases in each group, which may lead to reduced power.

This indicates that the effect of reading comprehension ability on pre-service teachers' practical knowledge performance was less intense when the animations replaced the text-based scenarios. Reading comprehension explained 22% and 24.9% of the variance in PK-SJT performance, in the animated and the text-based version of the text, respectively. In other words, the use of animations led to a reduction of the variance attributed to this construct-irrelevant factor by 2.9%.

These differences were more pronounced when the analysis was conducted separately for native and non-native English speakers[27]. As illustrated in Figures 4.5 and 4.6, the impact of reading comprehension skills on performance was weaker in the case of the animated PK-SJT, both for the non-native, animated PK-SJT: $r(40) = .395$, $p = .012$, text-based PK-SJT: $r(38) = .457$, $p = .004$, and the native English speaking group, animated PK-SJT: $r(26) = -.028$, $p = .890$, text-based PK-SJT: $r(25) = .270$, $p = .191$.

It should be mentioned, though, that in the Irish sample, the correlations between reading comprehension ability and PK-SJT performance were not statistically significant. This implies that native English speakers' PK-SJT performance was not statistically significantly shaped by their reading comprehension skills, independently of the test format they completed. In the case of non-native English speakers, animations managed to reduce the variance attributed to their reading comprehension ability in English (R^2) by 5.5 percentage points, which is equivalent to a 26% decrease of the variance explained by this factor.

[27] Subgroup analysis results should be interpreted with more caution because of the smaller number of cases in each group, which may lead to non-significant due to reduced power.

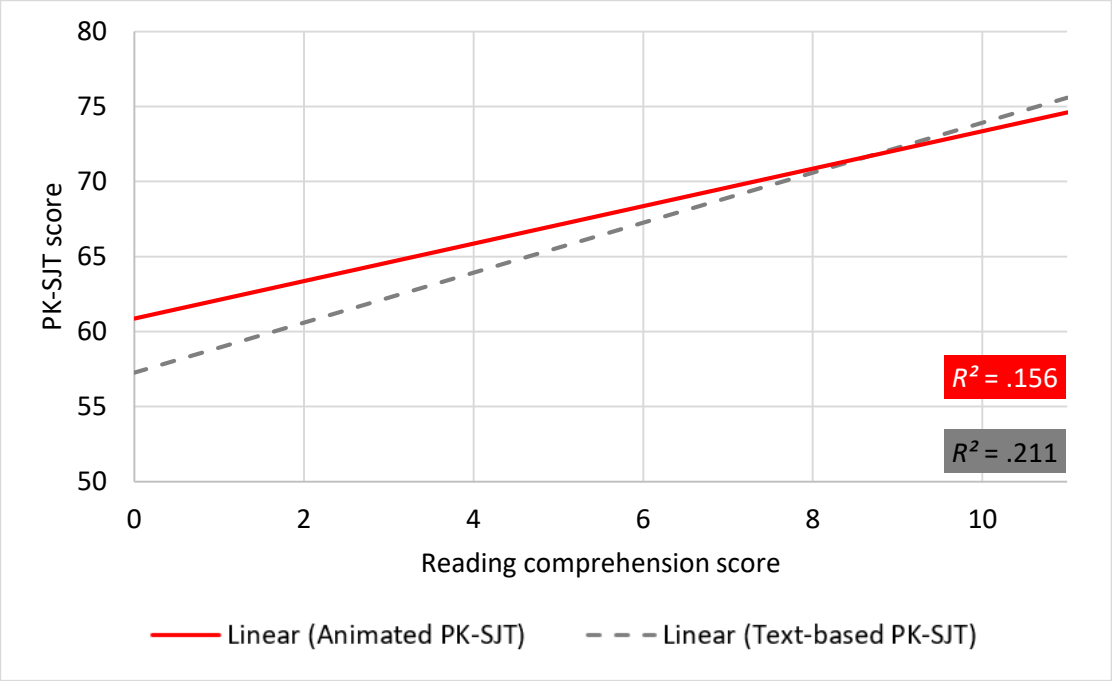


Figure 4.5. Correlation between reading comprehension and PK-SJT performance (non-native English speakers).

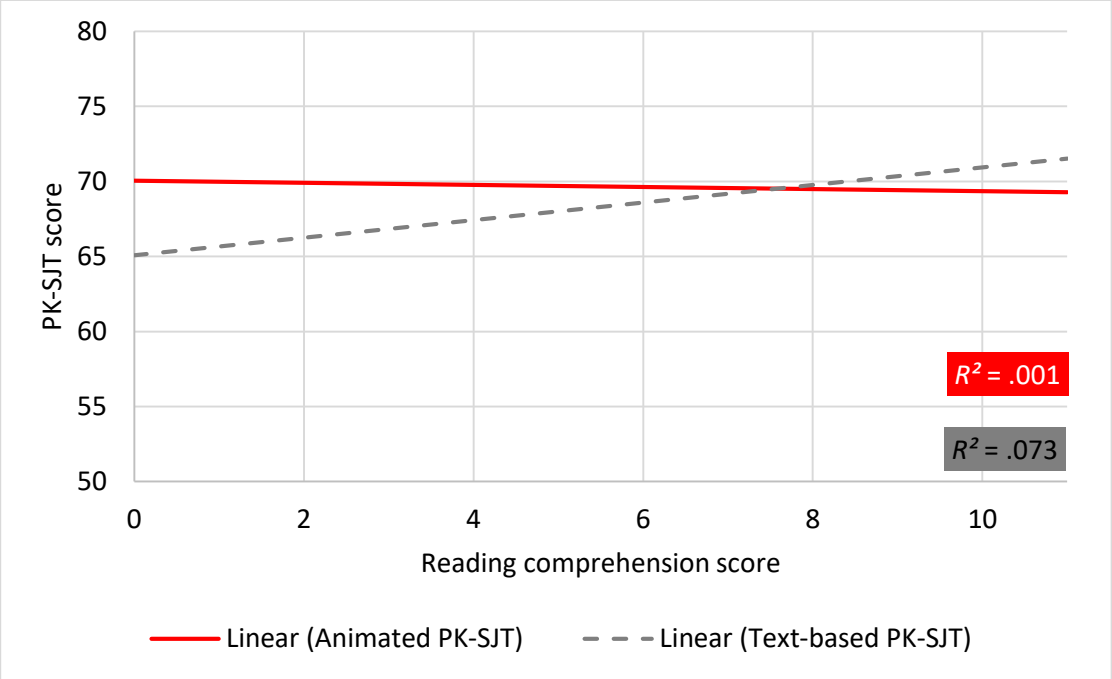


Figure 4.6. Correlation between reading comprehension and PK-SJT performance (native English speakers).

In order to further explore the overall effectiveness of animations in reducing the impact of reading comprehension ability on PK-SJT performance, the interaction between test format and reading comprehension ability (i.e., *Test format*Reading comprehension*) was tested in a hierarchical multiple linear regression. In step one, the regression model

included PK-SJT performance as the outcome variable, and *Reading comprehension* and *Test format* as predictors, while the interaction term of the two variables was added in step two as the third predictor of the outcome variable.

Based on the unstandardised coefficients *(B)* of the regression model, as presented in Table 4.6, with reading comprehension ability held constant, the experimental group was expected to perform 1.97 score-points higher than the control group on the PK-SJT, a difference that was statistically significant. Reading comprehension was also a statistically significant predictor of PK-SJT performance, with every extra unit[28] in the reading comprehension scale leading, on average, to a 1.43 score-point increase in the practical knowledge scale. Reading comprehension was a stronger predictor of PK-SJT performance than test format, according to their standardised coefficients (β). The negative interaction between the two predictors indicated that, overall, the impact of reading skills on the PK-SJT performance was expected to be reduced via the use of animations, however, this effect was not large enough to be statistically significant.

Table 4.6

Test format - Reading comprehension regression model

Predictors	*B*	*SE B*	β	R^2
Step 1				.259
Test format	1.973*	0.920	0.165	
Reading comprehension	1.434**	0.230	0.478	
Step 2				.260
Test format*Reading comprehension	-0.185	0.463	-0.073	

Note. The analysis was based on a sample of 129 test-takers.
*$p < .05$. **$p < .01$.

Another hierarchical multiple linear regression was applied only for the group of non-native English speakers, where the impact of reading skills on PK-SJT performance was statistically significant. The model revealed similar results regarding the interaction between the test format and participants' reading comprehension score. Again, the interaction between the two variables was negative, indicating a reduced adverse impact when animations were used, albeit not a statistically significant one. Table 4.7 summarises the results of the regression models.

[28] A *unit* refers to a score-point in the relevant scale.

Table 4.7

Test format - Reading comprehension regression model (non-native English Speakers)

Predictors	*B*	*SE B*	*β*	R^2
Step 1				.213
Test format	2.325	1.267	0.188	
Reading comprehension	1.457**	0.355	0.481	
Step 2				.217
Test format*Reading comprehension	-0.656	0.713	-0.134	

Note. The analysis was based on a sample of 78 test-takers.
**p < .05. **p < .01.*

Research question 1.4.a: *What proportion of the total variance in PK-SJT scores can be attributed to construct-irrelevant factors?*

Multiple regression analysis indicated that, overall, construct-irrelevant factors (i.e., native language, English proficiency[29] and reading comprehension skills) accounted for a large proportion of variance in participants' PK-SJT scores (R^2 = 33.9%). As Table 4.8 shows, English proficiency and reading comprehension were both statistically significant predictors of pre-service teachers' SJT performance. After accounting for these two factors, participants' native language was not a significant predictor of their performance, indicating that the gap between native and non-native English speakers can be attributed to their language and reading comprehension competencies.

Table 4.8

Regression model with construct-irrelevant factors (I)

Predictors	*B*	*SE B*	*β*	R^2
				.339
Native language	1.050	1.215	0.086	
English proficiency (Advanced)	0.914**	0.253	0.305	
Reading comprehension	3.831**	1.191	0.311	

Note. The analysis was based on a sample of 129 test-takers.
**p < .05. **p < .01.*

[29] *Native language* (native/non-native English speaker) was used as a proxy for proficiency in English. As explained above, non-native speakers were asked to report their English proficiency. To address these research questions (i.e., 1.4.a and 1.4.b), it was assumed that all native English speakers were either advanced or proficient in English, and thus, classified under the *Advanced* English proficiency group.

Research question 1.4.b: *Does the use of animations reduce the proportion of variance attributed to construct-irrelevant factors in PK-SJT scores?*

As Table 4.9 shows, the English proficiency and reading comprehension variables were statistically significant predictors of participants' PK-SJT performance, both in the animated and the text-based format of the test. Similar to the previous regression model (Table 4.8), after accounting for English proficiency and reading comprehension ability, there was not a statistically significant gap between native and non-native English speakers in either the animated or in the text-based PK-SJT.

Table 4.9

Regression models with construct-irrelevant factors (II)

Predictors	*B*	*SE B*	*β*	R^2
Text-based PK-SJT				.372
Native language	1.050	1.773	0.080	
English proficiency (Advanced)	0.992**	0.356	0.326	
Reading comprehension	4.486*	1.781	0.337	
Animated PK-SJT				.337
Native language	1.025	1.620	0.093	
English proficiency (Advanced)	0.753*	0.352	0.265	
Reading comprehension	3.603*	1.548	0.330	

Note. The analyses were based on a sample of 63 and 66 test-takers for the control and the experimental groups, respectively.
*$p < .05$. **$p < .01$.

In order to compare the impact of participants' reading comprehension and language proficiency on their PK-SJT scores in the animated versus the text-based version of the test, it is necessary to compare the proportion of the variance in PK-SJT performance explained by these construct-irrelevant factors in both test formats. In the text-based PK-SJT, the amount of variance (R^2) explained by construct-irrelevant factors was equal to 37.2% of the total variance. In the case of the animated PK-SJT, though, the percentage of the construct-irrelevant variance was lower (R^2 = 33.7%) – a decrease of 3.5 percentage points. This finding indicates that the use of animations led to a 9.4%

decrease of the variance in PK-SJT performance attributed to construct-irrelevant factors[30].

To explain these findings in practical terms, the unstandardised coefficients (*B*), as shown in Tables 4.8 and 4.9 were used. Overall, with other variables held constant, advanced English speakers were expected to outperform their non-advanced counterparts in the PK-SJT by 0.91 score-points, regardless of the test format they completed. The gap was expected to be smaller in the cases of the animated (gap of 0.75 score-points) compared to the text-based PK-SJT (gap of 0.99 score-points). Similarly, with other variables held constant, for every one-unit (score-point) increase in their reading comprehension score, pre-service teachers' PK-SJT scores were expected to increase by 3.83 score-points on average. For example, someone who had a score of six out of 11 in the reading comprehension test was expected to score 3.83 score-points higher on the PK-ST than someone who had a score of five out of 11 on the reading test. Comparing the experimental to the control group, it was anticipated that this gap would be larger for those who took the text-based PK-SJT (increase of 4.49 score-points for every extra unit in the reading comprehension scale) compared to those who took the animated PK-SJT (increase of 3.60 score-point for every extra unit in the reading comprehension scale). A detailed discussion of these findings is presented in the final chapter of the thesis

4.5.1. Additional analysis using the extended scales

As explained in Chapter 3, experienced teachers from Greece and Ireland were asked to rate the suitability of every practice statement to inform the scoring of pre-service teachers' responses on the PK-SJT. Comparisons between native (i.e., Irish) and non-native (i.e., Greek) English speakers were important in the context of this study, hence, a common 16-item scoring scale, for which experienced teachers' ratings from the two countries did not statistically significantly differ, was developed to assess pre-service teachers' practical knowledge. For the first research question, the main analysis was conducted using the common 16-item PK-SJT scale. However, analyses for which comparisons between native and non-native speakers was not required, can also be conducted for each country separately using all of the available practice statements that

[30] A decrease from 37.2% to 33.7% is a change of 3.5 percentage points and, when expressed into percentages, a decrease of 9.4%.

were categorised as either *Good* or *Bad* by the experienced teachers in each country. These analyses were conducted and the results are presented in Appendix I. Overall, the results using the extended scales for each country were almost identical to the results as presented earlier.

4.6. Research Question 2: What Impact Does the Use of Animated Videos Have on Test-Takers' Reactions to the Test?

The second set of research questions focused on test-takers' reactions to the PK-SJT that they completed. The data from the Likert-type questions they answered were used to create a number of measurement scales, as described in Chapter 3. The final scales for each factor were computed by taking test-takers' average scores on the relevant items[31]. Table 4.10 presents participants' minimum, maximum and mean scores along with the standard deviations on each one of these scales.

Table 4.10

Perceptions and invested effort scales

Scales	*Minimum*	*Maximum*	*M*	*SD*
Face validity	2.78	4.78	3.85	0.37
Enjoyment	3.33	5.00	4.48	0.47
Invested effort	2.60	5.00	4.24	0.56

As Table 4.10 shows, it could be concluded that there was reasonable evidence of face validity and enjoyment, as the means of both scales were closer to the upper end. Participants also reported investing substantial effort in the assessment. These three scales were used to compare the text-based and the animated version of the PK-SJT in terms of face validity, enjoyment and invested effort. The impact of animations on test-takers' perceived difficulty of the test was also examined.

Research question 2.1: *What impact does the use of animated videos have on the face validity of the PK-SJT?*

The results of the Independent-Samples T-test indicated that the animated version had statistically significantly higher levels of face validity (M = 3.91, SD = 0.36) compared to the text-based version of the PK-SJT (M = 3.77, SD = 0.37), $t(127) = -2.133$, $p = .035$.

[31] Scores ranged from 1 to 5.

In other words, test-takers perceived the animated PK-SJT to be more closely related to someone's ability to deal with challenging social situations that may be encountered in the teaching profession, more predictive of a teacher's behaviour in the classroom and a fairer indicator of teachers' practical knowledge. The effect size of this difference was small to moderate, $d = 0.38$.

To determine the extent to which this finding was consistent across the native and non-native English-speaking groups, the statistical significance of the interaction term between *Test format* and *Native language* was tested in a multiple linear regression with *Face validity* as the outcome variable and *Test format*, *Native language* and their interaction term as predictors. The results of the regression analysis showed that the interaction between the two variables was not statistically significant ($p = .303$), indicating that the animation of the original text-based PK-SJT had the same, positive impact on face validity across native and non-native English speakers.

Research question 2.2: *What impact does the use of animated videos have on test-takers' enjoyment of the PK-SJT?*

The test-takers who completed the animated version of the test had statistically significantly higher mean enjoyment ratings ($M = 4.59$, $SD = 0.44$) compared to their peers who took the text-based version of the test ($M = 4.36$, $SD = 0.48$), $t(127) = -2.881$, $p = .005$. In other words, the animated PK-SJT was found to be more interesting and enjoyable than the text-based PK-SJT. The effect size of this difference was moderate, $d = 0.50$.

To examine whether this impact was consistent across different groups (i.e., native and non-native English speakers), the significance of the interaction term between *Test format* and *Native language* was tested in the relevant multiple linear regression model. The interaction term was not statistically significant ($p = .087$), indicating that the positive impact of animations on test-takers' enjoyment applied in the case of both native and non-native English speakers.

Research question 2.3: *What impact does the use of animated videos have on test-takers' perceptions of the difficulty of the PK-SJT?*

The two statements examining the perceived difficulty of (i) the content of and (ii) the language used in the assessment were treated as independent variables and not as part of

a scale. Figure 4.7 presents the extent to which the participants agreed or disagreed with the two difficulty-related statements. Overall, test-takers found neither the content of nor the language used in the assessment to be particularly challenging. However, they reported being more concerned about the difficulty of the content rather than the language used in the assessment.

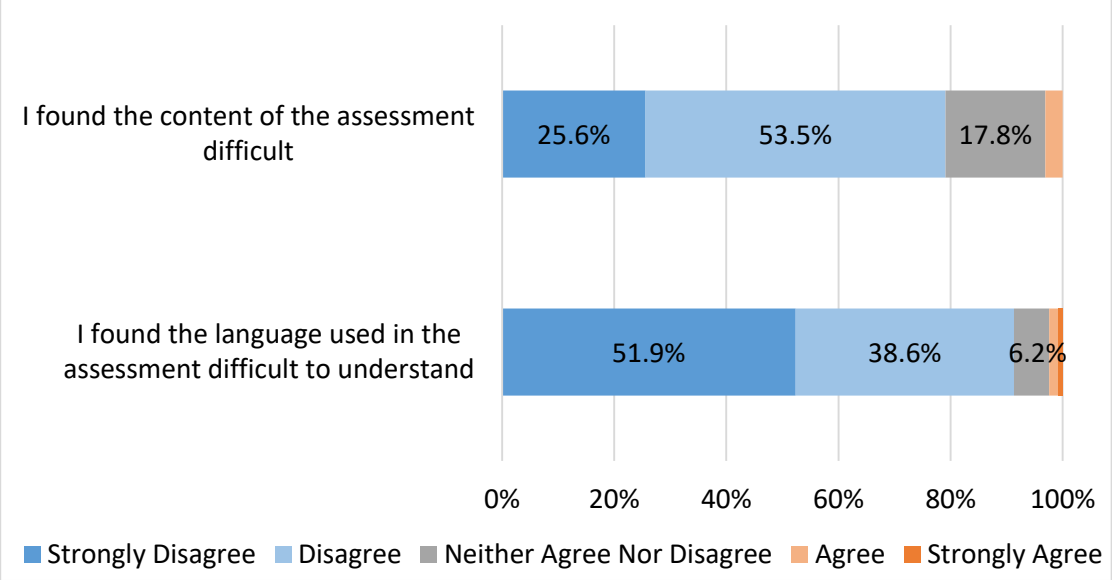


Figure 4.7. Perceived difficulty of the PK-SJT.
Note. Data labels for percentages smaller than 5% are not included in the graph.

Almost 80% of the sample disagreed or strongly disagreed that the content of the assessment was difficult, while more than 90% of the test-takers disagreed or strongly disagreed that the language of the assessment was difficult to understand. Unsurprisingly, non-native English speakers reported that they found the language used in the assessment more difficult, compared to their native peers (U = 1529, p = .013), with the effect size of this difference being small to moderate, η^2 = 0.048. However, this was not the case for the perceived difficulty of the content of the assessment (U = 1804.5, p = .328)[32].

The Mann-Whitney U test indicated that there were no statistically significant differences between the animated and the text-based version of the PK-SJT in terms of the perceived difficulty of the content of the assessment, U = 2033.5, p = .813. However, the results were different for the perceived difficulty of the language used in the

[32] The Mann-Whitney U test was used instead of an Independent-Samples T-test because the dependent variables were ordinal and not continuous.

assessment. The Mann-Whitney U test revealed statistically significant differences between the animated and the text-based version of the test, with those taking the text-based version of the test being more likely to perceive the language used in the test more difficult to understand, compared to those who completed the animated test, $U = 1631$, $p = .018$. The effect size of this difference was small to moderate, $\eta^2 = 0.043$. These differences are illustrated in Figures 4.8 and 4.9.

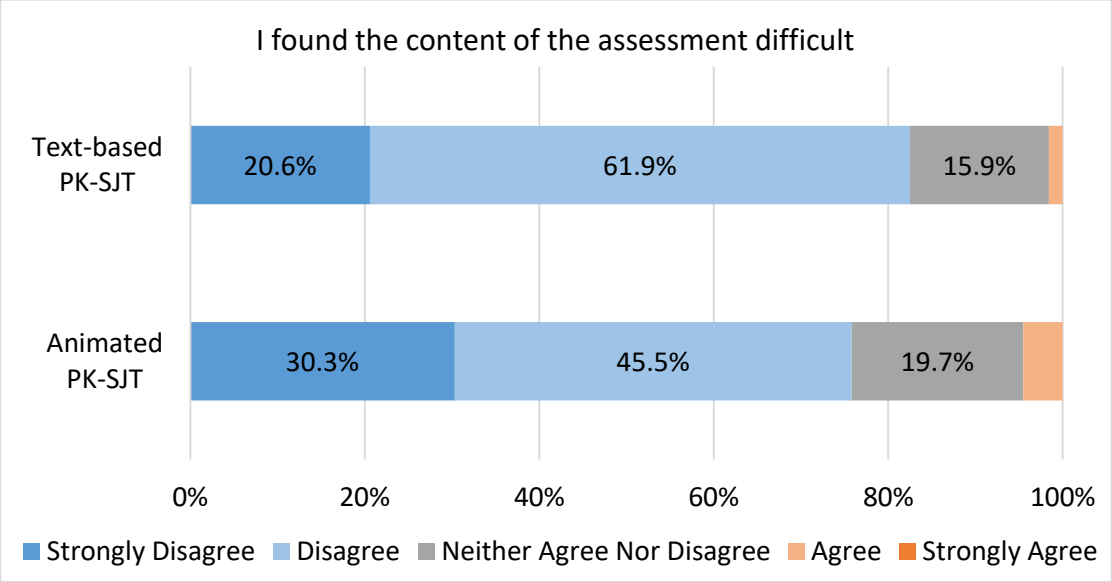


Figure 4.8. Perceived difficulty of the content of the assessment.
Note. Data labels for percentages smaller than 5% are not included in the graph.

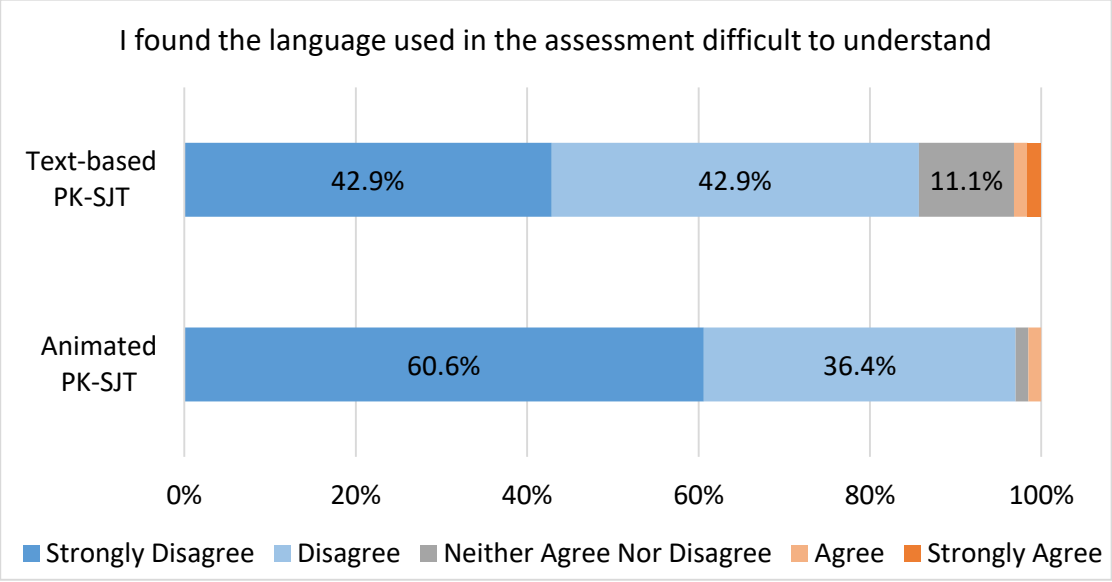


Figure 4.9. Perceived difficulty of the language used in the assessment.
Note. Data labels for percentages smaller than 5% are not included in the graph.

To examine whether the impact of animations on the perceived difficulty of the language used in the PK-SJT was constant across native and non-native English speakers, a multiple ordinal regression model, with the perceived difficulty of the language of the test as the outcome variable was conducted. *Test format, Native language* and their interaction (i.e., *Test format*Native language*) were added as predictors in the model. The interaction between the two variables was not statistically significant (p = .833), indicating that the positive impact of animations on reducing the perceived difficulty of the language used in the assessment was constant across the two groups as defined by their native language.

Research question 2.4: *What impact does the use of animated videos have on test-takers' invested effort in the PK-SJT?*

To answer this research question, an Independent-Samples T-test, comparing the control to the experimental group in terms of the effort test-takers put into the assessment was applied. The results showed that there was not a statistically significant difference between the text-based (M = 4.19, SD = 0.54) and the animated PK-SJT (M = 4.28, SD = 0.57) in terms of the effort participants reported investing in the assessment, $t(127)$ = -.899, p = .370. This finding indicates that, based on participants' self-reports, the animation of the text-based PK-SJT did not lead to significantly greater invested effort in the assessment and, as a consequence, any differences in the performance of the two groups cannot be attributed to this.

4.7. Summary

This chapter presented results on the impact of animations on construct-irrelevant variance and test-takers' reactions to the test. The chapter provided evidence on the equivalence of the control and the experimental groups, in terms of participants' background characteristics, English proficiency and reading comprehension skills. The research questions were addressed in two sets. The first set of research questions examined the effectiveness of animations in reducing construct-irrelevant variance and unintended subgroup differences attributed to language and reading skills. The relevant statistical analyses indicated that animations led to smaller gaps between groups characterised by differing native language and proficiency levels in English, and also, to test scores that were less dependent on test-takers' reading comprehension ability.

However, none of the individual effects was large enough to be statistically significant. Overall, though, the animations contributed to a 9.4% decrease in the variance in PK-SJT scores explained by construct-irrelevant factors.

As for test-takers' reactions to the assessment, the animated version of the PK-SJT was perceived as more valid and enjoyable by the participants. Additionally, test-takers found the animated PK-SJT to be less difficult, in terms of the language used, but not in terms of its content. Finally, participants did not report putting more effort in the animated compared to the text-based PK-SJT. The next chapter discusses these results in detail and outlines the conclusions that can be drawn based on the findings of this study.

Chapter 5: Discussion and Conclusions

5.1. Introduction

This thesis examined the extent to which animated videos can be a useful alternative to written text for assessing complex knowledge and skills. More specifically, it investigated the potential of animations to (i) reduce construct-irrelevant variance attributed to language-related skills in the practical knowledge situational judgment test (PK-SJT) and (ii) affect test-takers' reactions to the test. This final chapter summarises the major findings of the study, discusses their meaning and how they relate to previous studies in the field, and acknowledges the limitations of this research. Additionally, this chapter provides recommendations for policy, practice, and future research in the field of animated assessments.

5.2. The Impact of Animated Videos on Reducing Construct-Irrelevant Variance

One of the main aims of this study was to examine whether the use of animated videos, as opposed to written text, can mitigate the impact of language-related construct-irrelevant factors (i.e., native language, proficiency in English, reading comprehension ability) on test performance. The fact that those who were administered the animated version of the test performed statistically significantly better than those who took the text-based PK-SJT suggested that, indeed, there were differences between the two formats. However, this evidence, on its own, does not provide adequate information regarding the improved quality of the animated versus the text-based version of the test.

To this end, the relationship between PK-SJT performance and a series of construct-irrelevant factors was examined. A sample of 129 pre-service teachers from Ireland (native English speakers) and Greece (non-native English speakers) was used for this purpose. Overall, the animations succeeded in reducing the performance gap between native and non-native English speakers, as well as between advanced and non-advanced English speakers. In addition, pre-service teachers' performance on the animated version of the PK-SJT was less dependent on their reading comprehension, compared to the performance of their peers who completed the text-based version of the same test.

However, none of these individual effects was large enough to be statistically significant.

In order to provide a more comprehensive picture of the role of animations in reducing construct-irrelevant variance in test performance, regression models including all three construct-irrelevant factors as predictors of pre-service teachers' performance on the PK-SJT were constructed. Both in the animated and the text-based version of the PK-SJT, participants with advanced proficiency and higher levels of reading comprehension ability in English performed better than their counterparts with low proficiency and poorer reading comprehension, respectively. After accounting for test-takers' proficiency and reading comprehension in English, participants' native language was no longer significant, indicating that the performance gap between native and non-native English speakers may be attributed to the lower English proficiency and reading comprehension scores of the latter group.

In both test formats, a large proportion of the variance in PK-SJT scores was explained by these construct-irrelevant factors. However, this variance was lower in the case of the animated versus the text-based PK-SJT by 9.4%. Based on these results, it could be concluded that the use of animations led to a reduction of the construct-irrelevant variance in pre-service teachers' PK-SJT performance.

The findings of this study indicated that there may be merit in using animated videos as opposed to written text in assessments of complex knowledge and skills. However, it should be acknowledged that, in the animated PK-SJT, there was still a large amount of variance in PK-SJT scores explained by construct-irrelevant factors, that the use of animations did not reduce. It is difficult to determine if this variance can be attributed to actual differences between test-takers or whether it constitutes evidence of unintended subgroup differences that were not eliminated through the use of animations. In other words, while it is safe to argue that animations had a positive impact, enhancing the validity of the inferences drawn from test-takers' scores, assertions about the extent to which they are able to eliminate construct-irrelevant variance should be made with caution. Recommendations about how future research could possibly give more definite answers to this question are provided later in this chapter.

The findings of this study could be also interpreted with respect to the two conflicting theories regarding the impact of animations on perceivers' understanding and

comprehension of the provided situation; namely, the *cognitive load* and *additive* theories. It seems that the animated PK-SJT managed to overcome the challenges of the *cognitive load theory*, according to which, the presentation of multiple information elements can place excessive demands on perceivers' working memories, which, in turn, may affect their ability to comprehend the provided material (Sweller et al., 2011). In other words, there was a risk of animations inadvertently introducing measurement error due to the fact that several different sources of information interact and simultaneously convey complex messages that may not be easily digested by the recipients. Instead, it could be argued that the findings of this study provide some support for the *additive theory*, according to which the more information perceivers are provided with, the better they comprehend the conveyed messages (Archer & Akert, 1980). It should be noted, though, that either theory could be correct, as the positive or negative impact that animations may have on the perceivers heavily depends on the exact nature of the animations, how they are developed, and how well the multiple pieces of information are integrated. These are aspects that were all taken into consideration when designing the animated videos, as explained in Appendix A. The meticulous design of the animated videos in this study contributed towards the development of an instrument that seemed to facilitate rather than incommode test-takers.

As explained earlier, there is a lack of research examining the use of animations as an alternative to written text for reducing the impact of language and reading comprehension skills on test-takers' performance, especially in the context of SJTs. Early attempts to replace written descriptions in SJTs with videos took the form of acted videos involving human actors performing a scenario. The most relevant research study in the field of acted video-based tests dates back to 1997, when Chan and Schmitt compared the acted to the text-based format of the same SJT and measured the reading comprehension ability of a sample of Black and White college students. The findings of this doctoral study corroborate and extend Chan and Schmitt's (1997) findings and suggest that the use of videos as an alternative to written text can mitigate the impact of construct-irrelevant factors on test-takers' performance. It should be acknowledged, though, that in Chan and Schmitt's (1997) study, the impact of acted videos was particularly pronounced as the construct-irrelevant variance was not only reduced but became statistically non-significant in the case of the video-based SJT. This means that, in the case of the video-based test, test-takers' performance was not, to any extent,

affected by their reading comprehension. In this study, the use of animated videos did not appear to achieve this to the same degree.

In contrast to acted videos, research on the use of animated videos in assessment is much scarcer. Previous studies in the field either used animations in a different way to this study or focused solely on test-takers' perceptions of the test. Thus, their validity evidence may not be directly comparable to the results of this study. Dancy and Beichner (2006), for instance, examined whether the animation of static images in science items could reduce the dependency of students' science performance on their verbal skills. Despite the fact that their work was among the few studies that examined the use of animations in assessment from a validity point of view, as explained in Chapter 2, the way Dancy and Beichner (2006) used animation technology, the features of the animated items, and the aims of the assessment were very different from the approaches used in this study. Their findings, though, were similar to the findings of this study, indicating that the use of animations may reduce the impact of language-related construct-irrelevant factors on test-takers' performance.

5.3. Test-Takers' Reactions to the Animated Versus the Text-based Test

The findings of this study regarding the impact of animations on test-takers' perceptions of and invested effort in the assessment were more clear-cut, compared to the findings regarding construct-irrelevant variance. Animations managed to enhance the face validity of the PK-SJT. Although the two assessment formats were identical in terms of their content, pre-service teachers who completed the animated PK-SJT considered it to be significantly more relevant to the teaching profession and a fairer indicator of their practical knowledge compared to those who took the same test in its text-based format. The use of animations also positively influenced the extent to which test-takers enjoyed the assessment process. These findings corroborate a great deal of previous research findings in the field of both acted and animated assessments (Bruk-Lee et al., 2016; Kanning et al., 2006; Richman-Hirsch et al., 2000), and further support one of the main advantages of multimedia and, specifically, animated videos over text-based assessments; assessments using multimedia were perceived to be more authentic and evoke more positive test-taker reactions.

Although face validity and "enjoyableness" have been investigated by previous research in the field, this study took a further step towards the examination of test-takers' reactions to the assessment by examining the impact of animations on test-takers' perceived difficulty of and invested effort in the test. As far as the difficulty of the test is concerned, the use of animations did not significantly affect the perceived difficulty of the content of the PK-SJT. On the other hand, the use of animations, which led to a considerable reduction in the use of written text, was found to reduce test-takers' perceived difficulty of the language used in the test. This reinforces the findings about the reduction of construct-irrelevant variance attributed to reading and language skills, as presented earlier. In other words, the positive influence of animations on reducing the adverse impact of language and reading skills on test scores was perceived by the test-takers, who found the language used in this test format less difficult compared to the language used in the text-based PK-SJT.

Finally, invested effort was measured in order to provide some indication of test-takers' engagement with the assessment process. Despite the fact that animations were linked to higher levels of enjoyment among the test-takers, they did not significantly affect the levels of effort participants put in the PK-SJT. Such a finding also implies that the overall impact of animations on test-takers' performance could not be attributed to the fact that those who took the animated test put more effort in it, as, based on participants' responses, this was not the case.

5.4. Contributions

Modern technology offers many opportunities for improving assessment and this study has made an important contribution to the field by examining the use of animated videos in testing. Significantly, this was the first study that compared an animated and a text-based version of the same SJT, using an experimental design methodology to provide validity evidence regarding the extent to which animated videos could mitigate construct-irrelevant variance attributed to language and reading skills. In addition, the study explored the role of animations in improving test-takers' perceptions of a test, the importance of which is often overlooked within the testing community.

On top of that, this was one of the few studies that investigated the effectiveness of animated-video SJTs in a context other than personnel selection. Last but not least, by

documenting the process of animating a text-based test, this study also provided valuable information regarding the decisions made, the steps followed and the challenges encountered in such a venture that was heretofore missing in the research literature. Indeed, the fact that this study has offered and tested a prototype for creating technology-enabled tests that use animations to deal with measurement-related challenges and fairly capture complex knowledge and skills is expected to be of real value to the research community.

5.5. Limitations

There are a number of limitations underlying this study that should be taken into account when interpreting the findings. First of all, it should be noted that this was a small-scale study and that, in many cases, the sample size was not large enough to obtain statistical power at the recommended .80 level (J. Cohen, 1988). Low statistical power leads to increased risk of Type II error, meaning that statistical tests may fail to reject the null hypothesis when it is false (L. Cohen et al., 2011). Therefore, this is an issue regarding effects that were not found to be statistically significant, such as the impact of animations on pre-service teachers' invested effort in the test, where the statistical power of the analysis applied was particularly low for detecting such a small effect (power of .15).

Subgroup analysis results (e.g., analysis of non-native speakers' data), in particular, should be interpreted with more caution because of the smaller number of cases involved in each group, which led to reduced power. For instance, a post-hoc power analysis regarding the non-significant correlation of native English speakers' (n = 51) reading comprehension score with their PK-SJT performance (r = .181) revealed low power (power of .24). On the basis of a correlation (r) of .181 between reading comprehension and PK-SJT scores, a sample size of 237 native English speakers would be needed to obtain adequate statistical power. As previous research in the field is particularly scarce, information about the expected effect sizes and sample sizes required to assure adequate statistical power was limited.

Thus, the small sample size of this study may have impacted the robustness of the conclusions regarding the role of animations in improving assessment validity. More specifically, the influence of animations on reducing the adverse impact of each one of the three construct-irrelevant factors (e.g., native language, English proficiency, and

reading comprehension) on test-takers' performance was tested through the inclusion of the relevant interaction terms in three multiple regression models and was not found to be statistically significant in any of the three cases. The post-hoc analysis indicated that the statistical power of these regression models was adequate (i.e., above .80) for detecting large (f^2 = .35) or moderate (f^2 = .15) effect sizes, but low (power of .24 or below) for detecting smaller effect sizes (f^2 = .02)[33].

The second limitation of this study relates to the nature of the sample that was used. The main sample consisted of pre-service teachers. Although this constituted a convenient sample that was easily accessible, it might be the case that this was not a group of test-takers who would be expected to have severe difficulties with text-based assessments. It could be argued that the use of animated videos as opposed to text-based descriptions might have a stronger impact in cases of assessments of groups with low levels of education and/or weak text-processing skills. Despite the fact that the original intention was to undertake this study with adults who had low literacy levels and/or low levels of education, this was very challenging due to the difficulty of accessing such a group of people.

The third limitation of the study results from the way in which the sample was selected. Participation in the study was voluntary both for pre-service and experienced teachers; as such, all individuals who were approached had the right to reject taking part in the research. The fact that this was a self-selected sample may suggest that participants were more likely to be committed during the research, with greater willingness to provide meaningful data (Colman, 2008). This is likely to lead to self-selection bias as those volunteering to participate in the study may have different characteristics from those who did not participate. Due to ethical constrains, this is a limitation that could not be prevented.

Fourth, although the construct measured by the PK-SJT (i.e., practical knowledge) and the format used to present the stimuli (i.e., animated videos vs. written text) are conceptually distinct, it is often challenging to assure that empirically one does not impact the other. In other words, it could be argued that the use of animations is likely

[33] Post-hoc power analysis was conducted using an online statistics calculator (https://www.danielsoper.com/statcalc/calculator.aspx?id=9). The recommended effect sizes used for all regression models were the following: small (f^2 = .02), medium (f^2 = .15), and large (f^2 = .35), based on J. Cohen's (1988) guidelines.

to affect the actual construct that a test purports to measure. This could lead to having two seemingly identical tests in terms of content, that may capture different constructs.

In order to secure the invariance of the construct measured across the two test formats, this doctoral study used a common set of test items across both test formats. On top of that, the animated videos were strictly based on the written descriptions of the text-based PK-SJT to make the two formats comparable. However, holding the content of test items constant does not necessarily guarantee that the same construct is being measured across the two formats of the same test. While some previous studies in the field (e.g., Chan and Schmitt, 1997) tried to establish measurement invariance across multiple test formats using the multiple group approach to confirmatory factor analysis, this was not feasible in this study due to the very restrictive size of the sample. Despite the efforts to assure measurement invariance across the two test formats, the lack of relevant statistical evidence should be taken into consideration in the interpretation of the findings of this study.

The fifth limitation relates to the low reliability of some of the measures used in this study. The reading comprehension test, which constituted an important measure in this study, had a reliability of .52 (Cronbach's alpha), with values greater than .70 being considered as acceptable (L. Cohen et al., 2011). This could be attributed to the fact that (i) only a small number of items (11) were used to measure participants' reading comprehension and (ii) participants did not really seem to engage with this test as their performance was much lower than expected. Low levels of reliability were also found for the face validity scale (Cronbach's α = .63). The low reliability of these measures should be borne in mind when evaluating the findings regarding these factors.

The sixth limitation relates to the issue of self-report response bias. The measures of test-takers' reactions to the PK-SJT were based on pre-service teachers' self-reports. It should be acknowledged that this could lead to self-report response bias, as participants, intentionally or unintentionally, may have provided distorted responses (R. J. Cohen & Swerdlik, 2009).

Finally, the findings of this study indicated that the use of animations reduced the overall variance attributed to construct-irrelevant factors. However, it is difficult to determine whether animations managed to eliminate or just reduce construct-irrelevant variance. Differential Item Functioning (DIF) analysis could help determine whether there were

still items in the animated version of the test that discriminated against certain groups of test-takers (e.g., non-native English speakers). However, again, the sample of the study was not large enough for conducting such an analysis.

5.6. Recommendations for Policy, Practice, and Future Research

5.6.1. Policy and practice

Although the animations appeared to reduce the impact of construct-irrelevant factors on test-takers' PK-SJT performance, their effect was not large enough to be statistically significant. Hence, the implications of the findings of this study for policy and practice are restricted. However, the fact that animations managed to provide positive outcomes using multiple measurement criteria and reduced the overall variance attributed to construct-irrelevant factors by almost 10% should not be overlooked.

The use of animated videos provided promising results, especially for groups of test-takers with limited proficiency and reading comprehension in English. From a construct-irrelevant variance point of view, animated videos may not be particularly valuable for assessments that are intended to be administered to native speakers with good levels of reading comprehension. Yet, it seems that their advantages should be seriously taken into account when it comes to assessments that are intended to be administered to groups with varying levels of language proficiency and reading comprehension (e.g., non-native speakers). The findings of this study suggest that providing test-takers with visual stimuli (e.g., animated videos) is a promising approach for enhancing the validity of the inferences drawn from their test scores.

The evidence provided by this study is particularly valuable in the context of modern societies and globalisation. Nowadays, the percentage of people living and working away from their country of birth is higher than at any other point in history, rising from 2.8% of the global population in 2000 to 3.4% in 2017 (Stroud, Jones, & Brien, 2018). As an inevitable consequence, many of these people are being taught and assessed in a language other than their mother tongue. Hence, numbers of test-takers with varying levels of language proficiency and reading comprehension across a wide range of fields are ever-increasing. This constitutes one of the most important challenges that education systems have to deal with. The validity of the inferences drawn from these people's test scores must constitute the most fundamental consideration in developing and evaluating

tests for such groups, and the use of new technologies, such as animations, could facilitate towards this goal.

To better interpret the findings of this study and provide further recommendations for policy and practice, the context and sample of the study should be taken into consideration. The participants of this study were higher education students, who were granted entry into their course on the basis of having a certain level of cognitive ability. Hence, it might be argued that they had the threshold level of reading skills necessary that allowed them to successfully enter such competitive college programmes in both countries (i.e., Ireland and Greece). It may be the case that the use of animations would have been more valuable for certain populations, such as test-takers with low levels of education. Other groups of test-takers who might benefit from the use of animations to achieve their potential could be those with certain learning difficulties, such as dyslexia. Dyslexia is linked to inaccurate, slow, and effortful reading, difficulties that can significantly affect the reading comprehension ability of those people facing such a learning difficulty (Petretto & Masala, 2017; Tunmer & Greaney, 2010). However, relevant research in the field of animated and, generally, multimedia assessment that specifically focuses on such groups is very scarce.

Going back to the original idea on which this study was based (as explained in section 1.4), it should be appreciated that the findings of this research are relevant to fields other than education. These include, but are not limited to, licensure, certification, and personnel selection assessments. These assessments can attract candidates with a broad range of profiles that can significantly differ in their language proficiency, education, and ability to process written text. Furthermore, the use of animated videos could also be valuable in non-cognitive measurement contexts. For instance, the use of animated videos instead or on top of verbal messages could accommodate young children who do not have fully functional verbal language allowing them comprehend items in psychological assessments.

In addition to providing validity evidence regarding the dependency of the animated and the text-based PK-SJT scores on pre-service teachers' language and reading skills, this study examined another very important aspect of every assessment process, namely test-takers' reactions to the test. One of the main findings was that the animated version of the test had improved face validity, compared to the text-based one. Although the

presence of face validity does not constitute validity evidence in the mainstream sense, it is still worthy of consideration, especially as assessments continue to evolve at a rapid pace with the introduction of new technologies. Popp et al. (2016) argued that a test should have face validity for three different groups of people: (i) those who make the final decision to use the test, (ii) those who administer the test, and (iii) those who take the test. As far as the latter group is concerned, face validity is a powerful concept, as examinees who believe that they are being assessed on characteristics relevant to the purpose of the test are more likely to place credence on the measure and be motivated to do their best, which, in turn, can affect their performance (Chan et al., 1997; J. C. Scott & Mead, 2011).

In personnel selection assessment contexts, the face validity of the measures used can affect not only the perceived reputation of a company but also candidates' desire to work there (Hausknecht, Day, & Thomas, 2004). In relation to that, Bruk-Lee et al. (2016) concluded that the use of animated videos rather than text-based descriptions in assessment can considerably impact test-takers' perceptions of the organisation and their intention to accept a potential job offer. Even though there is no evidence to support such an argument, based on Hausknecht et al.'s (2004) and Bruk-Lee et al.'s (2016) findings, one could argue that the face validity of the instruments used in educational assessment and measurement may have an impact on test-takers' attitudes towards the credibility of the institution organising the assessment.

Perceived difficulty of several aspects of a test is another important factor that test-developers should take into account as, according to Ryan and Huth (2008), it can affect test-takers' perceptions of and response to an assessment. This study indicated that although the use of animations did not affect test-takers' perceived difficulty of the content of the test, it did affect the perceived difficulty of the language used in it. However, this does not imply that test-developers should make tests easier in order to improve test-takers' perceptions of them. Decisions regarding assessment difficulty should be informed by the aim of the assessment in an effort to further enhance validity.

The implications of the findings about test-takers' perceptions are important, not only in the context of assessments that have short- or long-term consequences for test-takers (e.g., job selection assessments and college admission exams) but in low-stakes contexts as well. As explained earlier, test-takers are expected to be motivated to perform well

on high-stakes tests, independently of the perceived validity and difficulty of the test and the extent to which they enjoy it. However, in low-stakes contexts, test-takers' perceptions of the test may serve as a mediator, affecting their engagement with the assessment tasks. Harlen (2012) argued that motivation and engagement are key components in assessment for learning purposes (i.e., formative assessment) because engaging assessments enhance students' learning and further motivate them to achieve learning goals. Based on the findings of this study, although animations managed to make the assessment look more enjoyable, more valid, and less difficult (all of which are desirable, especially in formative assessment contexts), they did not lead to higher levels of effort. However, in order to provide definitive answers, more comprehensive measures of engagement should be undertaken.

5.6.2. Cost – benefit evaluation

The cost and development challenges are probably the two main reasons why video-based assessments are not very popular. An important question is whether the benefits of animated tests in terms of reducing construct-irrelevant variance and improving test-takers' perceptions can justify the resources required for their development. As outlined in Appendix A, the cost for the development of the 15 animated videos was approximately €15,000; a particularly competitive price.

With regards to construct-irrelevant variance and assessment validity, based on the findings of this study, it could be argued that the advantages of animated videos over text-based descriptions might justify the money and effort required for their development, but only in certain cases and for certain groups of test-takers. On the other hand, animations made a clear difference in improving test-takers' perceptions of the test. Such perceptions are likely to affect test-takers' attitudes towards and performance on the test, but the degree of this impact might vary across different types of tests (e.g., high-stakes, low-stakes, summative, formative). It is up to the key decision makers to determine whether the positive impact of animations on test-takers' perceptions of the test, as evidenced by the present and previous studies, can justify the extra cost and the work required.

To facilitate future applications of and decision-making about similar projects within the field of educational assessment and beyond, this thesis provided a detailed discussion of the challenges and the cost involved in developing such an animated assessment

(Chapter 3 and Appendix A). It should be acknowledged that the increased cost and complexity of animated assessments may preclude them from being used in small-scale assessment projects. In other words, there is so much work and cost involved in the development of animated tests that it would be very difficult for a teacher to develop such a test for classroom assessment purposes. Animated assessments are expected to better facilitate large-scale assessment projects, such as national and international assessments, university assessment programmes (e.g., teacher education courses), personnel selection procedures, and credentialing exams by large organisations and, generally, assessments that are intended to be administered to groups of people, large enough to justify their cost.

5.6.3. Future research

In exploring the potential of animations to enhance validity and improve test-takers' perceptions of tests, this study has contributed to the research literature pertaining to video-based assessments. Notwithstanding this contribution, the field of animated assessments remains relatively unexplored. More research in the area is necessary to secure clearer conclusions about the value of animated videos over text-based testing. The limitations of this study as well as aspects of the research problem that were not explored in the study now lead naturally to recommendations for what might be addressed in the future.

New research projects should ensure that they recruit samples large enough to detect even small effect sizes. This would allow for further analyses (e.g., measurement invariance and DIF analysis) that were not possible in this study due to the restrictive size of its sample. The relevant literature in the field of measurement invariance and DIF analysis provides some useful rules of thumb regarding the minimum sample sizes required for such analyses (Meade, 2005; N. W. Scott et al., 2009).

Moreover, future research should explore the effectiveness of animated videos for alternative populations, such as test-takers with poor education or learning difficulties. Such evidence would further inform policy and practice about the groups of test-takers for which the use of animated videos, as opposed to written text, might be beneficial.

In an effort to further examine the advantages of animated videos over text-based descriptions, research should consider exploring the potential of fully-animated

assessments. In these assessments both the stimuli and the response options would be presented via the use of animated videos and no text would be provided in written format. As explained in Appendix A, such an approach can be very challenging as it may impact the measured construct, threatening the measurement invariance between different testing formats. However, the possibility should be worth investigating.

It is also advised that research studies in the future use more complete and reliable measures of reading comprehension, as this is an important indicator of construct-irrelevant variance. Finally, as far as test-takers' engagement with the test is concerned, attention should be placed on examining how the use of animations can affect student engagement with the assessment by using more accurate and extensive measures of this construct and/or taking advantage of process data.

5.7. Epilogue

This doctoral study is the first-of-its-kind in setting out to investigate in-depth what animations can contribute over and above conventional text-based SJTs, and its findings are very promising. However, given the complicated nature of language- and reading-related issues in assessment and the multidimensionality of the proposed solutions to these issues (i.e., animated videos), it is clear that further investigations are required before more definitive answers to the questions posed in this study are attained. That said, as this study has shown, the use of animated videos has the potential to enhance the validity of the inferences drawn from the scores of certain groups of test-takers, while at the same time significantly improving the fidelity, face-validity, and "enjoyableness" of a test. These findings should not escape the attention of test developers, especially those working in large-scale testing programs. It is clear that the research road ahead is filled with exciting possibilities.

References

Abedi, J. (2004). The No Child Left Behind act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*(1), 4–14. https://doi.org/10.3102/0013189X033001004

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 377–398). New Jersey, NJ: Lawrence Erlbaum Associates.

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. 129–105). California, Ca: SAGE.

Abedi, J., Lord, C., & Plummer, J. (1997). *Language Background as a Variable in NAEP mathematics Performance*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior*, *28*(5), 1816–1828. https://doi.org/10.1016/j.chb.2012.04.023

Adamson, F., & Darling-Hammond, L. (2015). Policy pathways for twenty-first century skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach* (pp. 293–310). Dordrecht, Netherlands: Springer.

Agard, C., & von Davier, A. (2018). The virtual world and reality of testing: Building virtual assessments. In H. Jiao & R. W. Lissitz (Eds.), *Technology Enhanced Innovative Assessment: Development, Modeling and Scoring from an Interdisciplinary Perspective* (pp. 1–30). Charlotte, NC: Information Age Publishing.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Archer, D., & Akert, R. (1980). The encoding of meaning: A test of three theories of social interaction. *Sociological Inquiry*, *50*(3/4), 393–419.

Arrindell, W., & Van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, *9*, 165–178.

Bauer, T., Truxillo, D., Sanchez, R., Craig, J., Ferrara, P., & Campion, M. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, *54*(2), 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x

Bialik, M., Martin, J., Mayo, M., & Trilling, B. (2016). *Evolving Assessments for a 21st Century Education*. Center for Curriculum Redesign.

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). Dordrecht, Netherlands: Springer.

Boyce, A. S., Corbet, C. E., & Adler, S. (2013). Simulations in the selection context: Considerations, challenges, and opportunities. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for Personnel Selection* (pp. 17–42). New York, NY: Springer.

Bridgeman, B., Lennon, M. Lou, & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*(3), 191–205. https://doi.org/10.1207/S15324818AME1603_2

Bruk-Lee, V., Drew, E. N., & Hawkes, B. (2013). Candidate reactions to simulations and media-rich assessments in personnel selection. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for Personnel Selection* (pp. 43–60). New York, NY: 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.

Buzzetto-More, N., Sweet-Guy, R., & Elobaid, M. (2007). Reading in a digital age: E-books are students ready for this learning object? *Interdisciplinary Journal of E-Learning and Learning Objects*, *3*(2003), 239–250.

Cabrera, M. A. M., & Nguyen, N. T. (2001). Situational judgment tests: A Review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*(1&2), 103–113. https://doi.org/10.1111/1468-2389.00167

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. Gage (Ed.), *Handbook on Research in Teaching* (pp. 1–80). Chicago, IL: Rand-McNally.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on Situational Judgment Tests: A content analysis and directions for future research. *Human Performance*, *27*(4), 283–310. https://doi.org/10.1080/08959285.2014.929693

Center for Advanced Research on Language Acquisition. (2014). Test Construction. Retrieved October 25, 2018, from University of Minnesota website: http://carla.umn.edu/assessment/vac/research/construction.html

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.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance*, *15*(3), 233–254. https://doi.org/10.1207/S15327043HUP1503_01

Chan, D., & Schmitt, N. (2004). An agenda for future research on applicant reactions to selection procedures: A construct-oriented approach. *International Journal of Selection and Assessment*, *12*(1/2), 9–23. https://doi.org/10.1111/j.0965-075X.2004.00260.x

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, *82*(2), 300–310. https://doi.org/10.1037/0021-9010.82.2.300

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York, NY: Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research Methods in Education.* (7th ed.). London, England: Routledge.

Cohen, R. J., & Swerdlik, M. E. (2009). *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (7th ed.). New York, NY: McGraw-Hill.

Colman, A. M. (2008). *A Dictionary of Psychology* (3rd ed.). https://doi.org/10.1093/acref/9780199534067.001.0001

Creswell, J. W. (2012). Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research. In *Educational Research* (4th ed., Vol. 4). https://doi.org/10.1017/CBO9781107415324.004

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). London, England: SAGE.

Cucina, J. M., Su, C., Busciglio, H. H., Harris Thomas, P., & Thompson Peyton, S. (2015). Video-based testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment*, *23*(3), 197–209. https://doi.org/10.1111/ijsa.12108

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

Darling-Hammond, L. (2000). Teacher quality and student achievement : A review of state policy evidence previous research. *Education*, *8*(1), 1–44. https://doi.org/10.1038/sj.clp

Darling-Hammond, L. (2014). *Next Generations Assessment: Moving Beyond the Bubble Test to Support 21st Century Learning* (L. Darling-Hammond, Ed.). San Francisco, CA: Jossey-Bass.

Dicerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology*, *18*(S1), 7–18.

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). https://doi.org/10.1007/978-3-642-33197-8_62

Dindar, M., Yurdakul, I. K., & Dönmez, F. I. (2013). Multimedia in Test Items: Animated Questions vs. Static Graphics Questions. *Procedia - Social and Behavioral Sciences*, *106*, 1876–1882. https://doi.org/10.1016/j.sbspro.2013.12.213

Edwards, B. D., & Arthur, W. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *The Journal of Applied Psychology*, *92*(3), 794–801. https://doi.org/10.1037/0021-9010.92.3.794

Eklöf, H. (2010). Student motivation and effort in the Swedish TIMSS advanced field study. *4th IEA International Research Conference*. Gothenburg, Sweden.

Elliott, J. G., Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., & Hoffman, N. (2011). The socially skilled teacher and the development of tacit knowledge. *British Educational Research Journal*, *37*(1), 83–103. https://doi.org/10.1080/01411920903420016

Eurostat. (2016). *Women Teachers Largely Over-Represented in Primary Education in the EU (news release)*. Eurostat Press Office.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037//1082-989X.4.3.272

Fetzer, M., & Tuzinski, K. (Eds.). (2013). *Simulations for Personnel Selection*. https://doi.org/10.1007/978-1-4614-7681-8

Field, A. (2017). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). London, England: SAGE.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to Design and Evaluate Research in Education* (8th ed.). New York, NY: McGraw-Hill.

Fröhlich, M., Kahmann, J., & Kadmon, M. (2017). Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. *International Journal of Selection and Assessment*, *25*(1), 94–110. https://doi.org/10.1111/ijsa.12163

Gardner, J. (2012). Assessment and learning: Introducation. In *Assessment and Learning* (2nd ed., pp. 1–8). London, England: SAGE.

Garrett, H. (1937). *Statistics in Psychology and Education*. New York, NY: Longmans, Green.

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

Golubovich, J., Seybert, J., Martin-Raugh, M., Naemi, B., Vega, R. P., & Roberts, R. D. (2017). Assessing perceptions of interpersonal behavior with a video-based situational judgment test. *International Journal of Testing*, *17*(3), 191–209. https://doi.org/10.1080/15305058.2016.1194275

Gravetter, F., & Forzano, L. (2012). *Research Methods for the Behavioral Sciences* (4th ed.). Belmont, CA: Wadsworth.

Griffin, P., & Care, E. (2015a). Preface. In *Assessment and Teaching of 21st Century Skills: Methods and Approach* (pp. v–x). Dordrecht, Netherlands: Springer.

Griffin, P., & Care, E. (2015b). The ATC21S method. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach* (pp. 3–33). Dordrecht, Netherlands: 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. 1–16). Dordrecht, Netherlands: Springer Berlin Heidelberg.

Grigorenko, E. L., Sternberg, R. J., & Strauss, S. (2006). Practical intelligence and elementary-school teacher effectiveness in the United States and Israel: Measuring the predictive power of tacit knowledge. *Thinking Skills and Creativity*, *1*(1), 14–33. https://doi.org/10.1016/j.tsc.2005.03.001

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–275. https://doi.org/10.1037/0033-2909.103.2.265

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *5058*(October), 1–19. https://doi.org/10.1080/15305058.2017.1297817

Halabi, A. (2012). *Perceptions of a Text-Based SJT Versus an Animated SJT (Master's Thesis)*. Minnesota State University, Mankato, Minnesota.

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

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

Harlen, W. (2012). The role of assessment in developing motivation for learning. In J. Gardner (Ed.), *Assessment and Learning* (2nd ed., pp. 171–183). Lancaster, England: SAGE.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

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*. San Diego, CA.

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

Hawkes, B. (2013). Simulation technologies. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for Personnel Selection* (pp. 61–82). New York, NY: 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, P. C. Kyllonen, & M. Zhu (Eds.), *Innovative Assessment of Collaboration* (pp. 95–111). Cham, Switzerland: Springer.

Hegarty, M. (2004). Dynamic visualizations and learning: Getting to the difficult questions. *Learning and Instruction*, *14*(3), 343–351. https://doi.org/10.1016/j.learninstruc.2004.06.007

Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, *17*(6), 722–738. https://doi.org/10.1016/j.learninstruc.2007.09.013

Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal*, *27*(3), 406–422. https://doi.org/10.1080/09585176.2016.1156004

Ingersoll, R. (2003). *Is There Really a Teacher Shortage?* University of Washington, DC: Center for the Study of Teaching and Policy.

Jamali, H. R., Nicholas, D., & Rowlands, I. (2009). Scholarly e-books: The views of 16,000 academics. *Aslib Proceedings*, *61*(1), 33–47. https://doi.org/10.1108/00012530910932276

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy and Practice*, *23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420

Kan, A., Bulut, O., & Cormier, D. C. (2018). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment*. https://doi.org/10.1080/10627197.2018.1545569

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

Kingston, N. M. (2008). Comparability of computer- and Paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education*, *22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Kirschner, P. A., Park, B., Malone, S., & Jarodzka, H. (2017). Toward a cognitive theory of multimedia assessment (CTMMA). In M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, Design, and Technology*. Cham, Switzerland: Springer.

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

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. *Medical Education*, *47*(2), 182–189. https://doi.org/10.1111/medu.12089

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *5058*(October), 1–7. https://doi.org/10.1080/15305058.2017.1309857

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, *90*(3), 442–452. https://doi.org/10.1037/0021-9010.90.3.442

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

Lievens, F., & De Soete, B. (2015). Situational judgment test. In J. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., pp. 13–19). Oxford, England: Elsevier Ltd.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, *9*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

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

Liu, Z. (2005). Reading behavior in the digital environment. *Journal of Documentation*, *61*(6), 700–712. https://doi.org/10.1108/00220410510632040

Macan, T. H., Avedon, M., Paese, M., & Smith, D. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology*, *47*, 715–739. https://doi.org/10.1111/j.1744-6570.1994.tb01573.x

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40.

MacCann, C., Lievens, F., Libbrecht, N., & Roberts, R. D. (2016). Differences between multimedia and text-based assessments of emotion management: An exploration with the multimedia emotion management assessment (MEMA). *Cognition and Emotion*, *30*(7), 1317–1331. https://doi.org/10.1080/02699931.2015.1061482

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*. Vancouver, Canada.

MacDorman, K. F., Coram, J. A., Ho, C.-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.

MacDorman, K. F., Green, R. D., Ho, C. 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

Malone, S., & Brünken, R. (2013). Assessment of driving expertise using multiple choice questions including static vs . animated presentation of driving scenarios. *Accident Analysis and Prevention*, *51*, 112–119. https://doi.org/10.1016/j.aap.2012.11.003

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61–68. https://doi.org/10.1016/j.ijer.2012.12.002

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, J. D., Salovey, P., & Caruso, D. R. (2002). *The Mayer, Salovey, and Caruso Emotional Intelligence Test: Technical manual*. Toronto, Ontario: Multi-Health Systems.

Mayer, R. E. (2005). Principles in managing essential processing in multimedia learning: Segmenting, pretraining, and modality principles. In R. E. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning* (pp. 169–182). New York, NY: Cambridge University Press.

Meade, A. W. (2005). Sample size and tests of measurement invariance. *20th Annual Conference of the Society for Industrial and Organizational Psychology*. Los Angeles, CA.

Mertens, D. M. (2014). An introduction to research. In *Research and Evaluation in Education and Psychology* (4th ed., pp. 1–46). Thousand Oaks, CA: SAGE.

Mislevy, R. J. (2011). *Evidence-Centred Design for Simulation-Based Assessment (CRESST Report 800)*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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.

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 178–197. https://doi.org/10.1080/0969594X.2015.1037241

Norcini, J. J., Lipner, R. S., & Grosso, L. J. (2013). Assessment in the context of licensure and certification. *Teaching and Learning in Medicine*, *25*(sup1), S62–S67. https://doi.org/10.1080/10401334.2013.842909

Organisation for Economic Co-Operation and Development. (2016). PISA 2015 science test questions. Retrieved October 9, 2017, from OECD Publishing website: http://www.oecd.org/pisa/pisa-2015-science-test-questions.htm

Organisation for Economic Co-Operation and Development. (2017). *PISA 2015 Technical Report*. Paris, France: OECD Publishing.

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

Petretto, D. R., & Masala, C. (2017). Dyslexia and specific learning disorders: New international diagnostic criteria. *Journal of Childhood & Developmental Disorders*, *03*(04). https://doi.org/10.4172/2472-1786.100057

Pitsia, V., Karakolidis, A., & Emvalotis, A. (2016). Attempting interdisciplinary approaches in education through the use of Google Earth (in Greek). *Education Sciences*, (3), 177–186.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Popp, E. C., Tuzinski, K., & Fetzer, M. (2016). 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). https://doi.org/10.4324/9781315871493

Quellmalz, E. S., & Silbertglitt, M. D. (2018). SimScientists: Affordances of science simulations for formative and summative assessment. In H. Jiao & R. W. Lissitz (Eds.), *Technology Enhanced Innovative Assessment: Development, Modeling and Scoring from an Interdisciplinary Perspective* (pp. 71–94). Charlotte, NC: Information Age Publishing.

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/Doi 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). New York, NY: Routledge.

Russell, M., & Airasian, P. (2012). *Classroom Assessment* (7th ed.). New York, NY: McGraw-Hill.

Ryan, A. M., & Huth, M. (2008). Not much more than platitudes? A critical look at the utility of applicant reactions research. *Human Resource Management Review*, *18*(3), 119–132. https://doi.org/10.1016/j.hrmr.2008.07.004

Ryan, A. M., & Patrick, H. (2001). The Classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal*, *38*(2), 437–460. https://doi.org/10.3102/00028312038002437

Sanchez, C. A., & Goolsbee, J. Z. (2010). Character size and reading to remember from small displays. *Computers & Education*, *55*(3), 1056–1062. https://doi.org/10.1016/j.compedu.2010.05.001

Sangmeister, J. (2017). Commercial competence: Comparing test results of paper-and-pencil versus computer-based assessments. *Empirical Research in Vocational Education and Training*, *9*(3), 1–19. https://doi.org/10.1186/s40461-017-0047-2

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly*, *40*(4), 715–738. https://doi.org/10.2307/40264305

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

Scott, J. C., & Mead, A. D. (2011). Foundations for measurement. In N. Tippins & S. Adler (Eds.), *Technolgy-Enhanced Assessment of Talent* (pp. 21–65). San Francisco, CA: Jossey-Bass.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*(3), 288–295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation*, *22*(4), 1–13.

Selingo, J. (2016). Education for a jobless future: Are colleges preparing students for the workforce? Retrieved September 13, 2017, from The Washington Post website: https://www.washingtonpost.com/news/grade-point/wp/2016/06/21/education-for-a-jobless-future-are-colleges-preparing-students-for-the-workforce/?utm_term=.170c2880df26

Smither, J., Reilly, R., Millsap, R., Pearlman, K., & Stoffey, R. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, *46*(1), 49–76. https://doi.org/10.1111/j.1744-6570.1993.tb00867.x

Spring, J. (2009). *Globalization of Education: An Introduction*. New York, NY: Routledge.

Stemler, S. E., Aggarwal, V., & Nithyanand, S. (2016). Knowing what NOT to do is a critical job skill: Evidence from 10 different scoring methods. *International Journal of Selection and Assessment*, *24*(3), 229–245. https://doi.org/10.1111/ijsa.12143

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

Stemler, S. E., Elliott, J. G., O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). A cross-cultural study of high school teachers' tacit knowledge of interpersonal skills. *American Education Research Association (AERA) Annual Meeting*. https://doi.org/10.302/1304486

Stemler, S. E., Elliott, J., McNeish, D., Grigorenko, E. L., & Sternberg, R. J. (2012). Examining the construct and cross-cultural validity of the Teaching Excellence Rating Scale (TERS). *The International Journal of Educational and Psychological Assessment*, *9*(2), 121–138.

Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. Ployhart (Eds.), *Situational Judgment Tests* (pp. 107–131). https://doi.org/10.4324/9780203774878

Sternberg, R. J. (1997). *Successful Intelligence: How Practical and Creative Intelligence Determine Success in Life*. New York, NY: Plume.

Sternberg, R. J. (1999). The theory of successful intelligence. *Reveiw of General Psychology*, *3*(4), 292–316.

Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence*, *34*(4), 321–350. https://doi.org/10.1016/j.intell.2006.01.002

Sternberg, R. J., & Grigorenko, E. L. (2001). *Practical Intelligence and the Principal*. Washington, DC: Institute of Education Sciences.

Stevens, J. P. (2009). Applied Multivariate Statistics for the Social Sciences. In *Group* (5th ed.). https://doi.org/10.4324/9780203843130

Stroud, P., Jones, R., & Brien, S. (2018). *Global People Movements*. London, England: Legatum Institute.

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

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research*, *81*(1), 4–28. https://doi.org/10.3102/0034654310393361

Tinwell, A. (2015). *The Uncanny Valley in Games & Animations*. Florida, FL: Taylor & Francis Group, LLC.

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

Trilling, B., & Fadel, C. (2009). *21st Century Skills: Learning for Life in Our Times*. San Francisco, CA: Jossey-Bass.

Tunmer, W., & Greaney, K. (2010). Defining Dyslexia. *Journal of Learning Disabilities*, *43*(3), 229–243. https://doi.org/10.1177/0022219409345009

Tuzinski, K. (2013). Simulations for personnel selection: An introduction. In M. Fetzer & K. Tuzinsk (Eds.), *Simulations for Personnel Selection* (pp. 1–16). New York, NY: Springer.

U.S. Office of Personnel Management. (n.d.). Assessment glossary. Retrieved January 11, 2017, from https://www.opm.gov/policy-data-oversight/assessment-and-selection/assessment-glossary/

United Nations Educational Scientific and Cultural Organization. (2014). *Global Citizenship Education: Preparing Learners for the Challenges of the 21st Century*. Paris, France: UNESCO.

Vandeweyer, M. (2016). Soft skills for the future. Retrieved September 14, 2017, from OECD: Skills and Work website: https://oecdskillsandwork.wordpress.com/2016/06/17/soft-skills-for-the-future/

Vinayagamoorthy, V., Steed, A., Street, G., & Slater, M. (2005). Building characters: Lessons drawn from virtual environments. *Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop*, 119–126. Stresa, Italy.

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the likert scale. *Educational and Psychological Measurement*, *72*(4), 533–546. https://doi.org/10.1177/0013164411431162

Ward, W. (2002). Test models. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-Based Testing* (pp. 37–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology*, *50*(1), 25–49. https://doi.org/10.1111/j.1744-6570.1997.tb00899.x

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*. Geneva, Switzerland: 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

Wu, H. C., Chang, C. Y., Chen, C. L. D., Yeh, T. K., & Liu, C. C. (2010). Comparison of Earth Science achievement between animation-based and graphic-based testing designs. *Research in Science Education*, *40*(5), 639–673. https://doi.org/10.1007/s11165-009-9138-9

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 359–377). New Jersey, NJ: Lawrence Erlbaum Associates.

Appendix A: The Process of Animating a Text-Based Situational Judgment Test

A1. Introduction

This study aimed to investigate the effectiveness of animations in reducing construct-irrelevant variance linked to language skills and improving test-takers' perceptions of the test. To examine this research problem, an animated version of a text-based practical knowledge situational judgment test (PK-SJT) was developed. As outlined in the main body of this thesis (Chapter 3), the development of an animated version of the original text-based PK-SJT was particularly challenging, as no previous literature had discussed the intricacies involved in each step of this process in sufficient detail. Keeping this in mind, the aim of this section is to provide information about the process of animating a text-based PK-SJT to develop the main instrument of this study (i.e., the animated PK-SJT). This information is expected to inform future research and practice in the field of video-based assessment. Specifically, the following topics are addressed:

- the steps involved in animating text-based SJTs,
- the critical decisions that had to be made, as well as the rationale behind them,
- the challenges faced during the animation process, and
- the timescale for completing a project of this scale.

A2. Decisions Regarding the Main Features of the Animated Videos

A2.1. Animated characters' appearance

As explained in Chapter 3, not all tests can or should be animated. The first and most critical decision that has to be made is whether or not the assessment of interest has characteristics that make it suitable for animation, keeping in mind that the aim should always be to improve the quality of a test. After confirming that the PK-SJT selected for this study had the required characteristics (as explained in the main body of this thesis), the next key decisions concerned the main features of the animations, such as their level of fidelity and realism. In other words, what should the animated characters look like? Should the animated videos use 2D caricatured, 3D caricatured or 3D realistic characters

and environments? As mentioned in Chapter 3, previous studies in the field of video-based assessment did not provide details about the approaches they followed during the animation process. However, other research literature either from the field of education and assessment, more generally, or from the field of animation technology, has provided some valuable information that helped inform the decision making regarding the main features of the animations in this study.

Levels of fidelity can vary significantly among different types of tests that use multimedia (e.g., audio, acted videos, and animations) (Motowidlo et al., 1990). 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). 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 hypothesised 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 is exceeded, viewers experienced an eerie sensation. This phenomenon is called the "Uncanny Valley". The term "valley" is used to describe the sharp fall of the levels of affinity perceivers have with humanlike objects that are very realistic. The term "uncanny" refers to the fact that the objects that transcend a specific point of human likeness and dip into the "valley" do not just fail to elicit empathy, but they even cause a sensation of eeriness (Misselhorn, 2009). Through his theory, Mori (1970) suggested that it is feasible to create a safe level of familiarity by deliberately developing a not-so-realistic design.

Although the theory of the Uncanny Valley originated in the field of robotics, it seems that 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, Green, Ho, & Koch, 2009). Some of the most famous animation productions, such as *The Polar Express*, have been criticised 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.

To empirically explore the Uncanny Valley theory, Schneider, Wang, and Yang (2007) collected individuals' (N = 60) opinions on 75 animated characters. The authors concluded 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. This conclusion is consistent with other research studies that investigated test-takers' reactions to four different formats of the same items (i.e., acted videos, 2D caricatured animations, 3D caricatured animations and 3D realistic animations) (Hawkes, 2012b, 2012a). Even though based on participants' perceptions, the 3D realistic animations had uncanny features, this was not the case for the 2D and 3D caricatured animated characters. However, these findings were in contrast with a study conducted by Bruk-Lee et al. (2016) comparing 2D, 3D (realistic), and acted-video SJTs. The results of that study revealed that the 3D assessment was more engaging and face valid than the 2D one. The acted-video SJT, though, received the most favourable reactions.

Apart from human likeness, there are other factors that may influence how robots and 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 organising the assessment (Popp et

al., 2016). 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, Coram, Ho, & Patel, 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.

The review of the available literature facilitated the decision making process regarding the design of the animated videos, and in particular, the level of authenticity and the features of the animated characters used in this study. Taking all of the evidence and information mentioned above into consideration, the aim was to obtain a balance between quality and affordability in creating an animated assessment that would enhance the fidelity of the original text-based SJT.

Firstly, it was decided to opt for a 2D rather than a 3D animated environment. Not only was this a much more affordable option, but based on the above literature, it also seemed like an appropriate choice in order to avoid the Uncanny Valley effect. The animated characters were caricatured to provide satisfactory authenticity without falling into the "valley". The aim was to offer test-takers a pleasant virtual experience which would not evoke any eerie sensations that could distract them from focusing on the assessment. Therefore, aesthetically pleasing, cartoon-style animated characters were used, based on the relevant guidelines provided by the literature. Figure A1 presents some examples of the main animated characters used in this study.

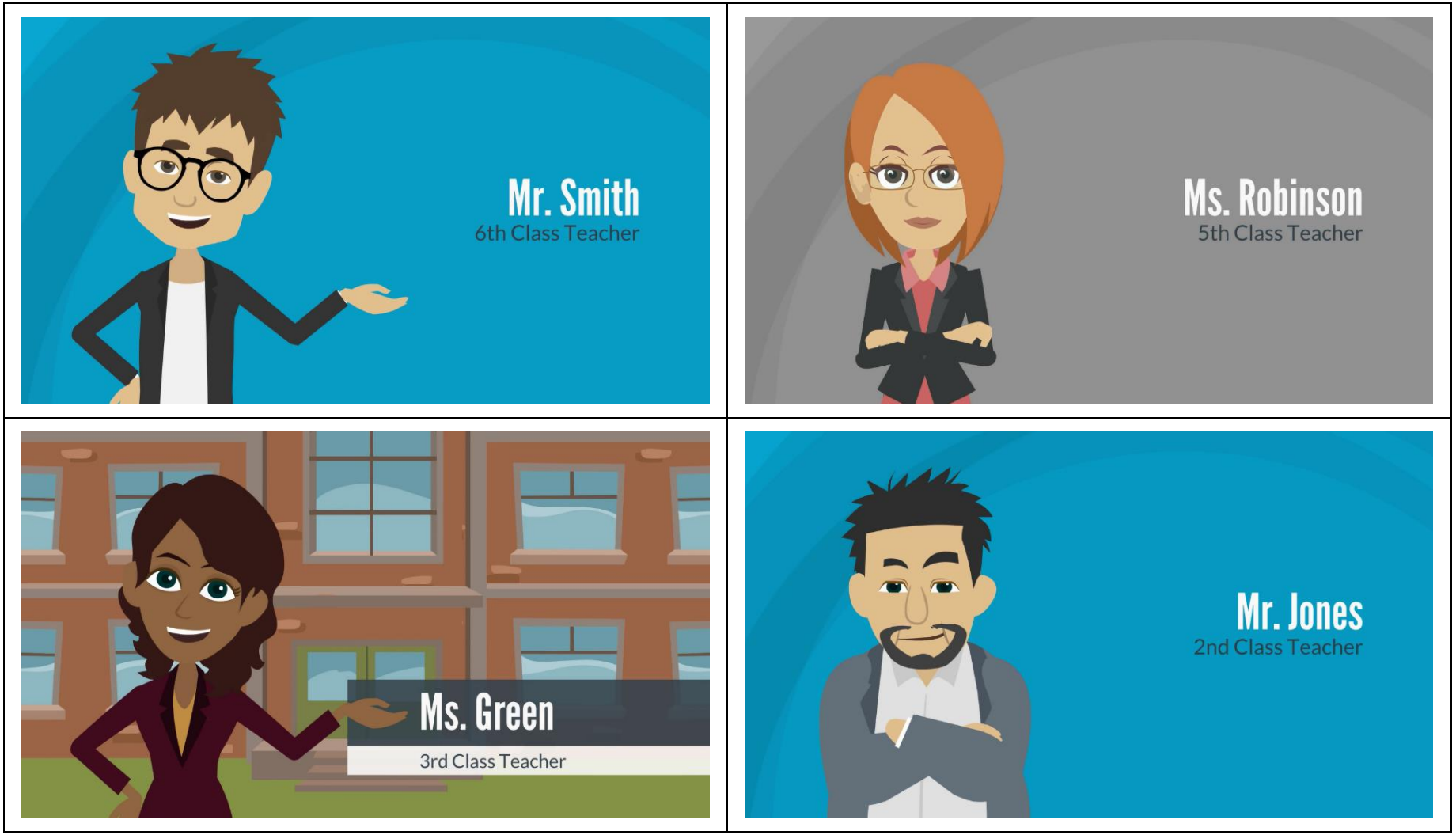


Figure A1. The animated teacher characters.

A2.2. 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, Grimshaw, Nabi, and Williams (2011) and Tinwell, Grimshaw, and Williams (2010). 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 more likely to be 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, on the other hand, was found to increase the uncanniness (Tinwell et al., 2010).

The same authors examined how speech qualities are linked to characters' familiarity. The findings indicated that monotony and slowness of speech were factors that increased the uncanniness of animated characters (Tinwell et al., 2010). 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 favourable outcomes (Mitchell et al., 2011; Tinwell et al., 2010; Vinayagamoorthy, Steed, Street, & Slater, 2005). 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.

Regarding body movement, White, McKay, and Pollick (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, Vinayagamoorthy et al. (2005) pointed out that, in general, animated characters' appearances should match their behaviour 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, 2015).

Regarding the animated characters used in this study, although they were able to express their feelings in a non-verbal way, it could not be argued that their facial expressions, especially in the upper part of the face, could convey complex emotions. This was a result of the limitations linked to the use of 2D caricatured characters, the nature of which restricted characters' motion as well. However, through the use of other means,

such as thought bubbles, characters were able to convey emotions and thoughts in a non-verbal way. Two examples are presented in Figure A2.

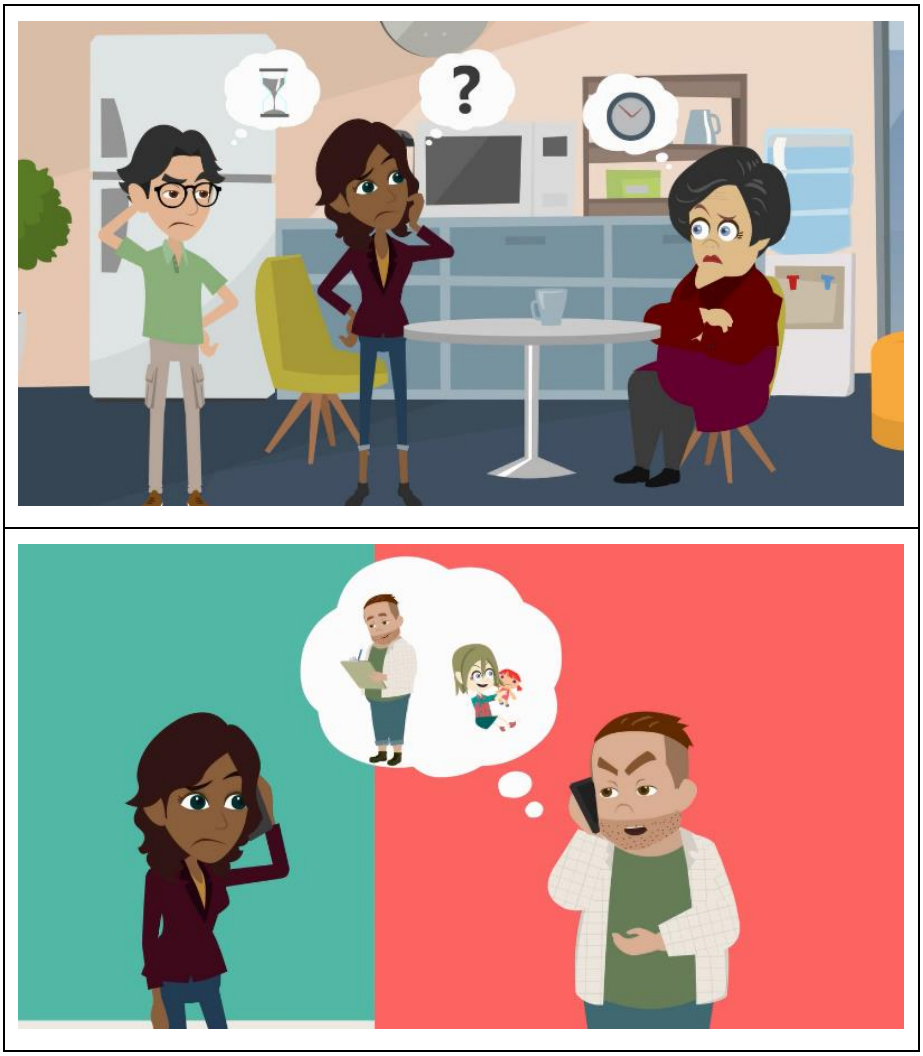


Figure A2. Examples of the use of thought bubbles in the animations.

Moreover, a human voice, rather than a computer-generated one, was selected to cover the audio used in the animations in order to fit with the characters' appearance and further contribute to their relatability. The audio was recorded in English using an American accent, which is the most recognisable among the different English accents and is regarded as being relatively easy to understand, especially for non-native speakers (Scales, Wennerstrom, Richard, & Wu, 2006). Overall, the animated characters' voice and movement were carefully handled to fit with their appearance naturally.

A3. Principles for Creating Viewer-Friendly Animated Videos

Animations and generally video-based tests have higher levels of fidelity than written tests, however, this does not suggest 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/on their own pace. In animations, many different sources of information such as objects and human representations interact and simultaneously convey sophisticated messages. This can create substantial extraneous cognitive load, which can place excessive demands on learners' working memories, and may affect learners' 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 (2005) argued, in animated representations, designers should familiarise 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 (e.g., Koumi, 2006). Koumi's (2006) guidelines were taken into account to make the animated videos of this study more viewer-friendly and easier to comprehend. His key points were:

- *Avoid 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. Keeping this in mind, the animated PK-SJT scenarios developed for this study included no written text. By keeping the use of written text in the

animations to a minimum, the danger of test-takers missing critical information communicated via the animations was avoided. This is the main argument behind the decision not to use subtitles; although they might have helped non-native speakers to fully comprehend the content of the scenarios, they may have also distracted them from the messages conveyed by the animations. The elimination of text was also consistent with the overarching goal of reducing construct-irrelevant variance related to language and reading proficiency. The process of communicating complex concepts and emotions in the absence of text proved challenging at times, however, this was eventually achieved through the creative use of thought bubbles containing vivid, universally recognisable images and symbols, as outlined previously.

- *Indicate where to look at the screen*. In a multimedia environment that communicates a lot of information in multiple ways, there is a risk that test-takers may be distracted by peripheral information and miss the main message that animations seek to convey. Therefore, when new characters were introduced in the 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.

- *Give users enough time to perceive the scene.* Although the aim was to keep the assessment and the animated videos short, pauses between the different scenes of a scenario, or even between different messages conveyed in the same scene, were introduced; the intention was that these pauses would allow test-takers to reflect on and better comprehend the content of the scenarios.

A4. Selection of the Animation Company

Following the selection of a suitable test for animation and having a relatively clear idea about how the final animated test should look, the next step was to source a company that would be able to work with us for developing the animated videos, whilst taking into account the available budget (approximately €15,000). In total, more than ten animation companies in Ireland and the UK were approached. Some of these companies indicated that animating scenarios on demand was not a service that they could provide,

while the rest submitted their offers. The price offers for animating the 15 SJT scenarios, using 2D technology ranged from €7,000 to €45,000. This large range of prices was due to the fact that some companies offered to design new characters for the scenarios, whereas others would use characters that had been already developed in the past. 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. It should be mentioned that the price offers for developing 3D animations were much higher.

After considering the offers, an agreement was reached with a UK-based company; *Animated Scenarios (Animated Scenarios* is a product of *LMS Global UK Ltd.)*. Their previous experience in working with clients in the field of education, transforming written scenarios into animations was a significant factor in this decision; indeed, they demonstrated a thorough understanding of the needs of this project within a short period of time. They utilised technologically-advanced software that included a range of ready-to-use characters, movements, facial expressions, environments, and objects, allowing them to make the animation process much quicker and more affordable.

A5. The Animation of the PK-SJT Practice Statements

One of the first decisions pertaining to the animation process was whether or not the text-based PK-SJT should be fully animated (i.e., with both the scenarios and the practice statements being animated). After reviewing the relevant literature, it was found that the majority of the studies that attempted to develop a video-based version of a text-based SJT focused on the scenarios, while the practice statements 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 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 "actual" validity, was not explored.

In the initial animation of the first PK-SJT scenarios, the practice statements were animated in an attempt to explore whether such an approach would enhance the quality of the test. However, it quickly transpired that the animation of the practice statements,

on top of the scenarios, created a number of additional complexities. After administering the fully animated scenario for review to five university staff members, they all agreed that their responses to the given situation could be affected not only by the nature of the strategy described in each statement but also by the way the practices were animated. The animation of the practice statement "Mr Smith should send William to the principal" provides an illustrative example. This animation presented a teacher who was angry and a student who was very upset, almost crying. Reviewers agreed that test-takers may be less inclined to select this option after watching this animation, 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.

The second issue linked to the animation of the practice statements was practical in nature. After administering both the animated and the text-based versions of the test, it was found that the animation of the scenarios increased the length of the original text-based test by approximately 15 minutes. The animation of the practice statements, on top of the scenarios, would significantly increase the time required to complete the assessment. This would create an unnecessary burden on test-takers, which could cause ethical concerns (Lingler, Schmidt, Gentry, Hu, & Terhorst, 2014) and also lead to fatigue and thus, less accurate responses. Last but not least, the animation of the practice statements would almost double the cost of the project.

For the aforementioned reasons, it was decided that, as an initial step to explore the use of animations in assessment, the animation of the practice statements could be omitted and emphasis should be put on the animation of the scenarios. It should be mentioned, though, that, in the animated version of the test, the practice statements, although not animated, were presented to the test-takers both in text and audio format to provide a multimedia experience. Thereby, reading loads, and hence, reliance on reading comprehension, were expected to be reduced.

A6. Animation of Elements That Were Not Described in the Text-Based PK-SJT

Another challenge faced during the animation process related to the way in which certain aspects of the scenarios were depicted in the animated videos. Even though the animated test was based entirely on the script of the text-based version of the test, animations necessitated the visualisation of aspects that, in the text-based SJT, were not necessary to describe. For instance, the written PK-SJT scenarios did not provide any information about how the characters and their physical environment looked (i.e., skin colour, facial characteristics, body, clothes, classroom design, wall colours, decoration, the way that students sit and work). Thus, the researcher had to make all these important decisions that gave life to the written descriptions – a particularly painstaking process.

Decisions regarding the animation of written scripts can cause some extra complexities, especially around issues that are particularly sensitive, such as how students of a particular gender or ethnic background are depicted. Emphasis was put on creating animations that depicted a contemporary, diverse school environment, avoiding any stereotypical references. Students coming from various ethnic backgrounds were presented to work together in small groups, with the role of the teachers being supportive, reflecting contemporary, collaborative pedagogical practices. The teacher, principal and parent characters were also designed in such a way to ensure gender balance, and age and ethnic diversity. Emphasis was put on the following matters:

- *Gender balance*: Although in the written scenarios it was important to keep a balance in the references to named male and female characters, on top of that, in the animated videos, it was necessary to ensure that there were roughly equal numbers of males and females presented, even in the background of the animated videos.
- *Ethnic diversity*: In the animated version of the scenarios, the ethnic background of the characters, an aspect that was not of major importance in the text-based test as references to ethnic origin did not exist, was considered. The aim was to ensure the ethnic diversity of the animated characters, in order to depict a multicultural environment. This not only served to increase fidelity, given that nowadays, such a school environment is common, but it also ensured that the assessment would be equally relevant for test-takers from various ethnic

backgrounds. It was important to make sure that, not only the main characters but also the characters in the background varied in terms of ethnic characteristics. Figure A3 presents two illustrative examples of gender balance and ethnic diversity.

Figure A3. Examples of gender balance and ethnic diversity.

- *Age diversity of adult characters*: In the text-based scenarios, there was no reference to the age of the adult characters, as this was not necessary. However, in the animated version of the test, it was impossible to avoid depictions of characters' ages. Therefore, it was ensured that the teacher, principal and parent characters represented a wide range of age groups. Figure A4 presents an illustrative example.

Figure A4. An example of teachers' age diversity.

- *Contemporary classroom environment*: In the text-based version of the test, it was not necessary to include any detailed reference to the classroom and school environment. However, in the animated assessment, the classrooms and the other environments where the scenarios took place had to be depicted. The aim was to design pleasant, colourful classrooms that follow contemporary principles with students collaborating using technology, and teachers having a supportive role. Figure A5 presents an example of such a classroom.

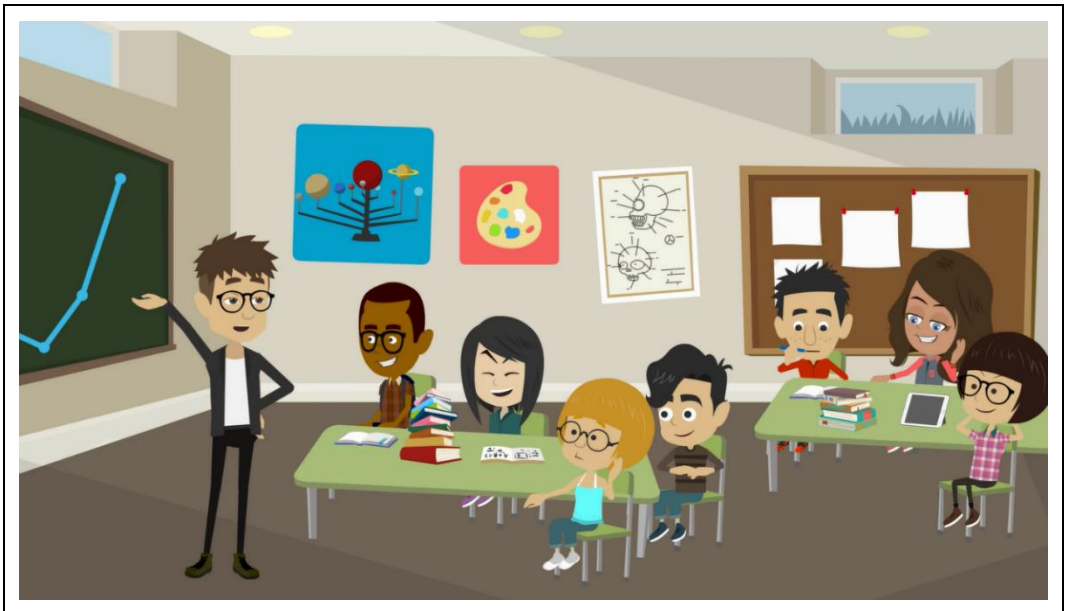


Figure A5. Sample classroom designed for the animated test.

- *Characters' reactions and facial expressions*: In the text-based version of the test, adjectives such as disappointed, frustrated, worried and destructive were used to describe certain behaviours. 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. Such issues were resolved mainly through trial and error. The animation company was informed about the desirable levels of intensity of characters' reactions in the different scenarios in advance, but a final decision was made only after the draft animations were reviewed.

Although it can be challenging to control for all these factors, it could be argued that the animation of written text is likely to further standardise 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.

A7. An Indicative Timeline

The development of an animated testing instrument is a particularly challenging process and the time required to complete it should not be underestimated. Insights into the time required to complete such a demanding project can be helpful for future research in the area. For this research study, approximately a month was required to select the 15 scenarios for animation and make all the necessary adjustments to render the scenarios more suitable for animation and appropriate for the European school context, as the scenarios had initially been developed for the US education system, as explained in Chapter 3. Once the final scripts were ready to be animated, they were gradually submitted to the animation company. It took approximately four months for the animation process to be completed. This was a dynamic, ongoing process of feedback exchange between the research team and the animation company.

When the animations were completed, they were administered to a number of educators and researchers in an attempt to receive constructive feedback. This process lasted almost a month. Finally, two months were required for the design and piloting of the

new testing platform, as explained in Chapter 3. Table A1 summarises the time required to complete these tasks working fulltime on this project. Of course, this is just an indicative timeline and the time required to complete a similar project may vary significantly.

Table A1

Indicative timeline

Task	Time required
Selection and adjustment of the 15 scenarios	One month
Animation of the written scenarios	Four months
Pre-piloting of the animations	One month
Design and piloting of the new testing platform	Two months
Total	Eight months

Appendix B: The Text-Based Situational Judgment Test Used in This Study

Scenario 1

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 recognises 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.

Scenario 2

Tom is a student in Mr. Smith's class. He comes to school, plays with other children, and does his homework. Tom is fine, but his parents can be difficult at times. Tom's mother thinks that Mr. Smith treats Tom unfairly; she believes that Mr. Smith pays more attention to some other students than he does to Tom. Anna, a new student, has now joined the class. Anna's reading is very weak, so Mr. Smith offers her extra help after class. Tom's mother learns about this and she phones looking for an explanation for why Tom is not getting the same kind of help. Tom's reading skills are not very strong, but he is doing reasonably well and, in Mr. Smith's opinion, does not need extra help at this time.

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. continue to give Anna help and wait to see if Tom needs help later in the year.
2. ask Tom's mother to contact the principal directly if she wants to discuss the issue.
3. make it clear to Tom's mother that, as his teacher, he is in the best position to make a judgment about Tom's needs and that if she has any doubts, she should contact the principal about moving Tom to another class.
4. tell Tom's mother that he is willing to offer Tom extra help for one week only.
5. tell Tom's mother that he would like to set up a meeting between the two of them and the principal to talk about her complaints.
6. make it clear to Tom's mother that only those students who perform poorly will be provided with extra help.
7. encourage Tom's mother to help him with his reading at home.

Scenario 3

Mr. Smith's class took an important maths test this week. During the last two weeks, Mr. Smith worked very hard to ensure that all the students were well prepared to take the test. He really wanted them to do well. In particular, he spent a great deal of time with Peter. Peter has difficulty understanding many maths concepts, but he worked really hard to prepare for this test. Mr. Smith hoped that, given their work together, this test would result in a good outcome for Peter. However, when Mr. Smith marked Peter's test, he was really disappointed. Both had worked very hard for such disheartening results.

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. overlook some of Peter's basic computational errors so that he gets a higher score.
2. continue to give additional help to Peter on a one to one basis so that he may do better on the next test.
3. decide that, from now on, students can receive extra marks for correcting their wrong answers on tests, so that students like Peter can improve their performance.
4. talk privately to Peter about his performance before returning the tests to the class.
5. ask another teacher for advice.
6. talk to the principal about arranging learning support in maths for Peter.
7. use Peter's test as a guideline when going over the answers with the class, to make sure he covers all of the areas with which Peter had trouble.

Scenario 4

Mr. Smith is usually very patient. It is important to him that every student understands the material he teaches, and that everyone feels encouraged to ask questions during and after lessons. Lately, Lilly, one of the low-achieving students in his class, has been asking a lot of questions.
Yesterday, when the class was supposed to be doing independent work, Lilly got out of her seat every couple of minutes to ask Mr. Smith questions. Although Lilly's questions were relevant to the assigned task, the time it took Mr. Smith to answer them made him unavailable to others in the class. Mr. Smith is planning a similar independent activity for the students tomorrow.

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. ask a colleague what they would do when presented with a similar situation.
2. draw up classroom guidelines for when and how students should seek assistance from him.
3. tell Lilly to do the best she can and to come see him after class if she still has questions.
4. highlight Lilly's questions by repeating them to the class.
5. set up a time with Lilly outside of class when they can go over her questions together.
6. pair Lilly up with a high-achieving student who is willing to help her.
7. spend time answering each of Lilly's questions.

Scenario 5

Ms. Green is the 3rd class teacher, and Katie is one of her students. Katie struggles with learning her multiplication tables and Ms. Green helps Katie by working with her after class. Neither in class, nor during the after school sessions has Katie made any progress. Yesterday, Ms. Green received a note from Katie's mother stating that her daughter could not learn the tables, that the experience frustrated Katie, and that Katie cried when she came home from school. The mother added that she did not see any reason for Katie to learn the multiplication tables. That evening, Ms. Green called Katie's mother and invited her in for a meeting. Katie's mother came and they talked, but unfortunately the conversation did not go well at all. Katie's mother insisted that the learning of multiplication tables was damaging to Katie, and therefore Ms. Green should stop forcing Katie to learn them.

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**

Ms. Green should

1. talk to Katie's mother again, explain the importance of mastering multiplication tables, and try to convince her that**,** with help, Katie will be able to master them and succeed.
2. organise a meeting with Katie's parents and the principal of the school.
3. increase her efforts to ensure that Katie learns her multiplication tables.
4. keep doing what she has been doing.
5. bring the situation to the principal's attention and ask him to contact Katie's mother to resolve this conflict.
6. stop helping Katie after class.
7. suggest to the principal that a new rule be introduced that requires students to get extra help during break or after school if their skills in any subject matter are below average.

Scenario 6

Ms. Green sometimes groups her students when teaching them. Using different colours, she divides the students in her class into three groups: the strongest students comprise the yellow group, the average-achieving students make up the grey group, and finally, the weakest students form the blue group. Groups of different colours receive different tasks and so the method helps to individualise teaching. Ms. Green also forms different groups for different subjects. She makes sure that her students understand why she groups them.

George was initially in the blue group for reading because his reading was below average. However, he was often one of the best in the blue group. A week ago George asked Ms. Green if he could work with the grey group. Ms. Green thought about it and decided that it was worth trying. On the very next day she noticed that George was struggling in the grey group. Reading at that level was really difficult for him. Moreover, there was a difference in how George felt about being in the grey group versus the blue group. In the blue group he was one of the best, while in the grey group he was the weakest student. George is very sensitive to his classmates' and teacher's opinions, and Ms. Green is worried about him. Yesterday on George's 4th day in the grey group, Ms. Green noticed that he was very upset following the reading lesson.

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**

Ms. Green should…

1. make it clear that once a student moves up to a new level, he or she has to stay in that group for a week at least.
2. talk to George on his own and encourage him to stay in the grey group, saying that eventually things will get better.
3. offer George after class help so that he knows that his teacher is trying to help him stay in the grey group.
4. phone George's parents and talk to them about the situation.
5. continue to observe the situation before making her decision.
6. tell George that he needs to return to the blue group because he is not performing well enough in the grey group.
7. get a student from the grey group to help George during class.

Scenario 7

Susan is one of the girls in Ms. Green's class, and she tends to perform below average on most tests and other assessments. However, she did much better on a recent geography project. The difference was so pronounced that Ms. Green began to suspect that it may not have been Susan's work at all and that Susan's parents may have helped her with the project. Ms. Green asked Susan if she had done the work herself and Susan said she had. Susan's father phoned Ms. Green the next day to tell her that he was upset that she could even suspect his daughter would hand in work that was not her own.

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**

Ms. Green should…

1. ask Susan's previous teacher about Susan's performance on tests compared to projects and classwork.
2. inform Susan's father that she has a class rule of following up on any incident of suspected cheating.
3. tell Susan's father that she appreciates his concerns and is reassured now that they have had an opportunity to talk.
4. explain that she became suspicious about Susan's work because her performance on that project was much better than usual.
5. tell Susan's father that she can put him in touch with the principal who can address his concerns.
6. explain to the father that she understands that he wants to help his daughter succeed, but doing Susan's work for her will not help his daughter in the long term.
7. tell Susan's father that she will think about the problem and phone him back at some point.

Scenario 8

Ms. Green's principal e-mailed all the teachers saying that the school had been given a small grant to participate in a research project on cooperative learning. He explained that he needed at least half of the teaches to take part in the project. The project would last for four weeks and would involve 20 minutes of class time per day. Teachers would also be required to attend one after-school workshop. Ms. Green is interested in the project, but she is concerned about the amount of class time it may take up. Many of the teachers have similar concerns. The principal has arranged for an information session this week to provide further details about the project.

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**

Ms. Green should…

1. decide not to participate because it will take up too much valuable time.
2. highlight to the other teachers that the project will take a lot of time away from focusing on students and call for a staff vote at the information session.
3. participate in the project to facilitate the principal on this occasion.
4. make it a personal rule to always participate in school research projects.
5. go to the principal to discuss whether or not she should participate.
6. go with the decision of the majority of teachers at her school.
7. ask for more time to think about what she should do.

Scenario 9

Mr. Jones has a new 2nd class. He discovers early on that one of his pupils, Mark, is a challenge. Mark seems to exhibit little self-control. He gets out of his seat a lot, has difficulty waiting in line and is sometimes very destructive with classroom materials. Mr. Jones wonders if there is a problem at home, or maybe Mark is being bullied, or he is having trouble adjusting to his new environment. After only three weeks of the new school year, Mr. Jones has not yet figured out what underlies Mark's behaviour.

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. Jones should…

1. continue to observe and document Mark's behaviour before doing anything.
2. praise the other children openly for staying in their seats and explain to Mark that his behaviour is unacceptable.
3. make a rule that when a pupil is disruptive in class then he or she must stay inside during break time.
4. talk to Mark and try to figure out why he behaves the way he does.
5. accept that this is how Mark will behave.
6. invite Mark's parents to the school to discuss the situation.
7. write a note to the principal stating that Mark's behavior is very disruptive to both him and the other children and request that something be done about it.

Scenario 10

Mr. Jones is involved in a disagreement with the parents of one of his pupils. The parents do not agree with Mr. Jones' judgment about their child's performance on a recent reading test. They think that their child should get a much better score. Mr. Jones has since reviewed the test again and stands by his initial judgment. For this reason, the parents complained to the **vice** principal, who is a good friend of the family. The vice principal agrees with the parents and tells Mr. Jones that he thinks the test score should be changed.

What should Mr. Jones do?

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. Jones should…

1. not make any decision until everyone has had time to consider the problem.
2. report the incident to the school board.
3. get the opinion of another teacher about the situation.
4. make it clear that changing test scores is against his policy.
5. take the **vice** principal's advice and change the test score.
6. invite the pupil, the parents, and the vice principal to meet so that he may explain his marking scheme.
7. ask another teacher to mark the test and suggest that everyone accept that test score.

Scenario 11

Today Mr. Jones is supervising school breaks when he observes that some pupils are laughing at Mary, one of the girls in his class. He can see that Mary is very upset. A few minutes later, she comes over to Mr. Jones and asks him for a hug. Other colleagues see this happening. Mr. Jones is aware that there is no official school policy on hugging pupils.

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. Jones should…

1. explain to Mary in front of everyone that hugging a teacher is not appropriate.
2. try to distract Mary by giving her a job to do.
3. give Mary a hug.
4. look to his colleagues and make a decision based on their reactions.
5. give Mary a hug, and make it a personal rule to comfort children when they are very upset.
6. ask Mary to report what happened to the principal.
7. talk to Mary about what happened, instead of giving her a hug.

Scenario 12

Ms. Robinson is a 5th class teacher and Jim is one of her pupils. Jim has little interest in maths, and he is absolutely convinced that he is no good at this subject. Ms. Robinson is not sure where this belief came from, but it is very strong. Of course, this belief impacts on Jim's learning, and therefore, he does not perform well in maths. He does not even try to engage with the subject and he is clearly not putting enough effort into his maths work. Ms. Robinson, however, feels that Jim is capable of succeeding.

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**

Ms. Robinson should…

1. encourage Jim's parents to get him a private tutor to help with maths.
2. invite Jim's parents to the school, and make the case that through the efforts of all three parties (Ms. Robinson, Jim and the parents) great progress can be made.
3. continue teaching and challenging Jim as usual.
4. make it clear to all of her pupils that anyone not making an effort in class has to do extra work at home.
5. insist that Jim participates more often during maths class to ensure that he engages with the subject matter.
6. talk to Jim and reassure him that he will have a more positive attitude towards maths by the end of the year.
7. allow Jim to work on other subjects he enjoys more during maths class.

Scenario 13

Ms. Robinson has a great relationship with her pupils and they love being in her class. She also does a lot of extra work outside the classroom, like preparing projects and attending extra-curricular activities. However, her family and friends have been telling her that she does not make enough time for them. She realises that she spends most of her time with activities related to school, so she makes plans to go out with her family for dinner on Thursday evening. Both she and her family are really looking forward to this as they rarely eat out together. On Thursday afternoon, however, one of her pupils, Dennis, tells her that he is looking forward to seeing her at the school musical that evening. Before she can get a word in, he continues, saying that he cannot wait for her to see it and that the musical would not be the same without her.

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**

Ms. Robinson should…

1. tell Dennis that she cannot attend the musical, but that she is sure it will be great.
2. discuss the situation with her family, and make a decision with them.
3. promise herself that, from now on, she will not break any more commitments to her family for a school-related activity.
4. take her family to the musical and go out together afterwards.
5. smile warmly without committing herself and then go out to dinner with her family.
6. tell Dennis that she will let her family decide what she should do, because she has already made plans with them.
7. explain to Dennis that she cannot attend the musical because has arranged something important.

Scenario 14

Patricia is one of Ms. Robinson's pupils and she is often disruptive in class. Patricia talks to her friends while Ms. Robinson is trying to teach, and when she asks her to be quiet, she often responds negatively. Ms. Robinson decides to phone Patricia's parents to talk to them about this issue. When she explains the situation, Patricia's father becomes very irritated, saying that he thinks Ms. Robinson is mistaken. He tells Ms. Robinson that at home Patricia is polite and friendly and he does not believe that she would be disruptive in class. Furthermore, he questions Ms. Robinson's competence to make a judgment about his daughter's behaviour.

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**

Ms. Robinson should…

1. tell Patricia's father that if he does not accept her professional judgment, she is going to end the call.
2. accept that she will not receive much support from Patricia's father and that there is not a lot she can do.
3. highlight the class rules and make clear what the consequences of disruptive behavior will be.
4. let the father explain his view and when he has finished, try to find some points they agree on.
5. suggest a meeting with the father, Patricia, and the principal to find a solution.
6. ask the father to meet in person so they can talk about the topic in a calm and polite way.
7. end the conversation and ask the principal to deal with the situation.

Scenario 15

Ms. Robinson usually gets along well with the other teachers in her school. However, during one staff meeting a colleague verbally attacks her when Ms. Robinson expresses an opinion about the school's extra-curricular activities that is in contrast to what most of the staff believe.

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**

Ms. Robinson should…

1. ask one of the other teachers for advice on how to deal with her colleague's comments.
2. draw her colleague's attention to the policies and procedures for staff meetings.
3. repeat her opinion, but state that she is willing to go along with the group.
4. talk privately with her colleague and say that she felt the personal attack was inappropriate.
5. have the principal speak to the colleague about the incident.
6. continue the conversation, focusing on the value of her proposal.
7. state that she is not willing to respond to personal attacks.

Appendix C: Experienced Teachers' Ratings

Table C1

Mean scores and standard deviations for the PK-SJT practice statements

	Overall sample		Greece		Ireland	
	M	SD	M	SD	M	SD
Scenario 1						
Statement 1	2.53	1.137	2.39	1.202	2.66	1.072
Statement 2	1.85	0.902	1.94	0.893	1.76	0.913
Statement 3	3.70	0.932	3.75	0.967	3.66	0.909
Statement 4	4.09	0.953	4.17	0.971	4.03	0.944
Statement 5	3.96	1.066	3.83	1.254	4.08	0.850
Statement 6	3.96	0.898	3.81	0.951	4.11	0.831
Statement 7	1.65	0.629	1.56	0.607	1.74	0.644
Scenario 2						
Statement 1	3.82	0.970	4.22	0.637	3.45	1.083
Statement 2	2.32	0.938	2.08	0.692	2.55	1.083
Statement 3	2.47	1.101	2.75	1.131	2.21	1.018
Statement 4	1.89	0.653	2.03	0.654	1.76	0.634
Statement 5	3.51	1.076	3.14	1.073	3.87	0.963
Statement 6	3.00	1.205	3.36	1.150	2.66	1.169
Statement 7	3.69	1.134	3.03	1.134	4.32	0.702
Scenario 3						
Statement 1	1.93	0.926	2.25	0.967	1.63	0.786
Statement 2	4.16	0.794	4.42	0.500	3.92	0.941
Statement 3	3.12	1.146	3.56	0.969	2.71	1.160
Statement 4	3.91	0.779	3.58	0.874	4.21	0.528
Statement 5	3.57	0.829	3.25	0.770	3.87	0.777
Statement 6	4.07	0.833	3.67	0.894	4.45	0.555
Statement 7	3.41	1.134	2.97	1.158	3.82	0.955
Scenario 4						
Statement 1	3.68	0.760	3.44	0.909	3.89	0.509
Statement 2	4.05	0.809	3.86	0.762	4.24	0.820
Statement 3	3.34	0.896	3.61	0.871	3.08	0.850
Statement 4	3.04	1.013	3.36	0.990	2.74	0.950
Statement 5	3.42	1.020	3.75	0.874	3.11	1.060
Statement 6	3.68	0.846	3.28	0.974	4.05	0.462
Statement 7	2.88	0.921	2.83	0.878	2.92	0.969

Table C1 (continued)

	Overall sample		Greece		Ireland	
	M	SD	M	SD	M	SD
Scenario 5						
Statement 1	3.82	0.998	3.81	1.009	3.84	1.001
Statement 2	2.95	1.133	2.44	0.998	3.42	1.056
Statement 3	3.51	0.983	3.78	0.959	3.26	0.950
Statement 4	3.07	1.127	3.53	1.082	2.63	0.998
Statement 5	2.68	1.074	2.47	0.971	2.87	1.143
Statement 6	2.14	0.911	1.86	0.723	2.39	1.001
Statement 7	2.16	0.937	2.47	0.941	1.87	0.844
Scenario 6						
Statement 1	2.49	1.162	2.86	1.125	2.13	1.095
Statement 2	3.38	0.989	3.78	0.681	3.00	1.090
Statement 3	3.23	0.959	3.61	0.728	2.87	1.018
Statement 4	3.09	1.062	2.69	0.920	3.47	1.059
Statement 5	3.86	0.799	3.83	0.697	3.89	0.894
Statement 6	1.97	0.596	2.03	0.446	1.92	0.712
Statement 7	3.70	0.918	3.86	0.762	3.55	1.032
Scenario 7						
Statement 1	3.77	0.973	3.39	0.994	4.13	0.811
Statement 2	2.95	1.071	3.42	1.025	2.50	0.923
Statement 3	3.78	0.864	3.50	0.971	4.05	0.655
Statement 4	3.38	0.961	3.36	1.046	3.39	0.887
Statement 5	2.01	0.785	1.83	0.609	2.18	0.896
Statement 6	3.38	1.202	3.81	1.142	2.97	1.127
Statement 7	2.54	1.009	2.19	0.951	2.87	0.963
Scenario 8						
Statement 1	1.80	0.662	1.81	0.467	1.79	0.811
Statement 2	2.23	0.869	2.39	0.803	2.08	0.912
Statement 3	2.96	1.039	2.36	0.867	3.53	0.862
Statement 4	2.68	1.035	2.69	1.037	2.66	1.047
Statement 5	3.58	0.828	3.31	0.889	3.84	0.679
Statement 6	2.85	1.029	2.86	0.990	2.84	1.079
Statement 7	3.32	0.938	3.44	0.843	3.21	1.018
Scenario 9						
Statement 1	4.03	0.793	4.17	0.561	3.89	0.953
Statement 2	3.22	1.114	2.64	0.961	3.76	0.971
Statement 3	2.23	0.837	2.36	0.867	2.11	0.798
Statement 4	4.38	0.716	4.47	0.506	4.29	0.867

Table C1 (continued)

	Overall sample		Greece		Ireland	
	M	SD	M	SD	M	SD
Statement 5	1.86	0.626	1.94	0.532	1.79	0.704
Statement 6	4.27	0.626	4.33	0.632	4.21	0.622
Statement 7	2.20	0.951	2.42	1.025	2.00	0.838
Scenario 10						
Statement 1	3.34	1.231	3.11	1.214	3.55	1.224
Statement 2	2.78	1.037	3.06	1.040	2.53	0.979
Statement 3	3.43	1.061	2.78	0.989	4.05	0.695
Statement 4	3.73	1.038	3.56	1.157	3.89	0.894
Statement 5	1.55	0.644	1.53	0.774	1.58	0.500
Statement 6	3.72	1.129	3.61	1.202	3.82	1.062
Statement 7	2.68	1.183	2.08	0.996	3.24	1.076
Scenario 11						
Statement 1	2.78	1.219	2.39	0.994	3.16	1.305
Statement 2	3.23	1.001	2.69	0.856	3.74	0.860
Statement 3	2.72	1.176	3.39	0.994	2.08	0.969
Statement 4	1.81	0.839	1.67	0.676	1.95	0.957
Statement 5	2.34	1.063	2.75	1.079	1.95	0.899
Statement 6	2.50	1.010	2.31	1.009	2.68	0.989
Statement 7	4.18	0.927	3.75	0.967	4.58	0.683
Scenario 12						
Statement 1	2.38	0.932	2.19	.889	2.55	0.950
Statement 2	4.35	0.730	4.31	0.668	4.39	0.790
Statement 3	3.30	1.030	2.92	0.996	3.66	0.938
Statement 4	2.23	0.944	2.61	0.934	1.87	0.811
Statement 5	3.86	0.669	4.08	0.439	3.66	0.781
Statement 6	4.20	0.662	4.17	0.609	4.24	0.714
Statement 7	1.51	0.646	1.61	0.599	1.42	0.683
Scenario 13						
Statement 1	3.86	0.865	3.64	0.867	4.08	0.818
Statement 2	2.85	1.178	3.22	1.072	2.50	1.180
Statement 3	3.30	1.082	2.94	0.893	3.63	1.149
Statement 4	3.32	1.136	3.78	0.832	2.89	1.226
Statement 5	2.41	0.920	2.58	0.906	2.24	0.913
Statement 6	2.18	0.956	2.22	0.866	2.13	1.044
Statement 7	3.88	1.033	3.44	1.081	4.29	0.802
Scenario 14						
Statement 1	2.19	1.056	2.17	0.941	2.21	1.166

Table C1 (continued)

	Overall sample		Greece		Ireland	
	M	SD	M	SD	M	SD
Statement 2	2.42	0.907	2.25	0.906	2.58	0.889
Statement 3	3.99	0.884	3.94	0.826	4.03	0.944
Statement 4	4.08	0.543	3.94	0.532	4.21	0.528
Statement 5	3.45	0.981	3.14	0.931	3.74	0.950
Statement 6	4.15	0.634	3.92	0.649	4.37	0.541
Statement 7	1.99	0.836	1.92	0.732	2.05	0.928
Scenario 15						
Statement 1	3.01	1.000	2.44	0.773	3.55	0.891
Statement 2	3.73	0.764	3.75	0.692	3.71	0.835
Statement 3	3.55	0.862	3.89	0.785	3.24	0.820
Statement 4	3.92	0.772	3.75	0.874	4.08	0.632
Statement 5	2.43	0.877	2.25	0.841	2.61	0.887
Statement 6	3.62	0.823	3.67	0.676	3.58	0.948
Statement 7	3.22	0.983	3.25	0.906	3.18	1.062

Note. *Bad* practices, *Good* practices, Statements for which there were statistically significant differences between Greek and Irish teachers' ratings (based on the results of Mann-Whitney U tests).

Appendix D: Alternative Scoring Approaches

Given the complex nature of the assessment used in this study (i.e., PK-SJT) and the fact that there were no objectively correct responses to the scenarios, it should be acknowledged that there are many alternative methods that could have been used to score participants' responses. In Table D1 some of those approaches are described, along with their levels of reliability (Cronbach's alpha). Similar alternative scoring approaches were used by Stemler and his colleagues in the past to measure complex skills for the purposes of various research projects (e.g., Stemler, Aggarwal, & Nithyanand, 2016; Sternberg, 2006).

As Table D1 shows, apart from the scoring approach used for the purposes of this research, the only scale that had adequate levels of internal consistency was the one computed based on the distance scoring approach. This may be attributed to the large number of items included in that scale (48 items). Table D2 presents the correlations among test-takers' scores computed based on the different approaches, including the one used in this research (i.e., original scoring approach). As expected, there were statistically significant strong correlations among all scoring approaches that focus merely on the *Bad* and *Good* practices. Scores generated from the pure distance approach were correlated to the other scoring approaches to a lesser extent. This was the case due to the fact that the pure distance approach took into account more items, the majority of which, though, were categorised as *Neutral*, based on experienced teachers' ratings. These items were not expected to provide much information about test-takers' ability to detect the G*ood* and avoid the *Bad* practices, as the construct was conceptualised by Stemler and Sternberg (2006).

Table D1

Examples of alternative scoring approaches of the PK-SJT

Scoring approach	Description	Items in the scale	Cronbach's alpha
The *correct response* scoring approach	Test-takers get one score-point if they: • *Agree*/*Strongly Agree* with a practice categorised as *Good* • *Disagree*/*Strongly* Disagree with a practice categorised as *Bad* If they give any other response, they get zero points.	16	.650
The *correct-penalised* scoring approach	Follows the same principles as the *correct response* scoring approach. However, test-takers get penalised and lose one score-point if they: • *Disagree/ Strongly Disagree* with a practice categorised as *Good*. • *Agree/Strongly Agree* or with a *Bad* practice.	16	.648
Distance scoring	For every item, the distance between the experts' (i.e., experienced teachers) mean and the test-takers' provided response is computed. The smaller the overall distance from the experts, the higher the score. This is a scoring approach that significantly differs from all the others as a distance score is computed for all items, regardless of whether they are categorised as *Bad*, *Good* or *Neutral*.	48	.784
Distance scoring scale for the Good and Bad practices	This approach follows the same principles as the pure *distance scoring* approach but it takes into account only test-takers' responses to items that have been categorised as either *Good* or *Bad*.	16	.570

Table D2

Correlations among the different scoring approaches

		Original scoring	Correct response scoring	Correct-penalised scoring	Distance scoring
Correct response scoring	Pearson Correlation	**.885***	-	-	-
	Sig. (2-tailed)	.000			
	N	129			
Correct-penalised scoring	Pearson Correlation	**.886***	**.959***	-	-
	Sig. (2-tailed)	.000	.000		
	N	129	129		
Distance scoring	Pearson Correlation	**.137**	**.356***	**.470***	-
	Sig. (2-tailed)	.121	.000	.000	
	N	129	129	129	
Distance scoring for *Good* and *Bad* practices	Pearson Correlation	**.510***	**.742***	**.832***	**.748***
	Sig. (2-tailed)	.000	.000	.000	.000
	N	129	129	129	129

**p* < .01.

Appendix E: Participants' Feedback on the Situational Judgment Tests of Practical Knowledge

At the end of the post-test survey, participants were presented with two open-ended questions that gave them the opportunity to provide some feedback on the test they completed. Participants' responses to the open-ended questions were not used to address any of the research questions of this study. Instead, they were used to inform future administrations of the PK-SJT or other animated tests.

The two questions used are the following:

1. In general, what do you feel worked well in this assessment, if anything?
2. What could be improved in this assessment, if anything?

The vast majority of the pre-service teachers who completed the PK-SJT provided some feedback by answering at least one of the two open-ended questions (122 out of 129). Most participants gave positive feedback, mentioning one or more aspects of the test that. according to them. worked well (121 out of 122). Fewer participants mentioned something that they did not like or provided improvement suggestions (77 out of 122).

Most of those who provided positive feedback mentioned that they found the assessment to be realistic and related to the job of a primary school teacher (67 out of 121). Some participants acknowledged the fact that the PK-SJT provided a good range of scenarios and response practices (40 out of 121) and/or reported that the close-ended nature of the assessment worked well (11 out 121). Some others reported that the information provided in the test was very clearly presented (30 out of 121). A few participants mentioned that they found the assessment interesting and enjoyable (19 out of 121). Finally, almost half of those who completed the animated version of the PK-SJT and answered the open-ended questions reported that the use of the videos was interesting and/or helpful (28 out of 61).

Regarding participants' responses on what could be improved in the assessment, a number of test-takers mentioned that on top of rating the available response practices, they would like to be able to provide their own responses in an open-ended format (19 out of 77). A few participants reported that they would like to have been provided with

more scenarios (14 out of 77) and/or more response practices (11 out of 77). Some test-takers mentioned that certain scenarios had very similar response practices (6 out of 77). Finally, A few of those who completed the animated version of the test and provided feedback on aspects of the test that could be improved reported that they would prefer to be able to move on and give their responses without having to wait for the voiceover of the response practices to finish (7 out of 35).

Appendix F: Exploratory Factors Analysis: Factor Loadings

Table F1

EFA for the perception- and effort-related statements: Pattern Matrix

	Factor				
	1	2	3	4	5
17. **Effort:** I could have worked harder on this assessment (r)	**.790**			.323	
16. **Effort:** I did NOT give this assessment my full attention (r)	**.670**				
14. **Effort:** I gave my best effort on this assessment	**.583**				
15. **Effort:** I worked on each item in the assessment	**.481**				
6. **Fairness:** This assessment was a fair indicator of someone's knowledge of how to deal with challenging social situations that may be encountered in the teaching profession		**.696**			
7. **Fairness:** This assessment gave me the opportunity to demonstrate my knowledge of how to deal with challenging social situations in teaching		**.572**			
5. **Face Validity:** A person's overall performance on this assessment would predict how well they deal with challenging social situations at school		**.546**			
8. **Fairness:** This assessment would NOT afford everyone the same opportunity to demonstrate their knowledge of how to deal with challenging social situations in teaching (r)		**.477**			
3. **Face Validity:** A person who can successfully tackle challenging social situations in teaching would do well on this assessment		**.345**			
12. **Enjoyment:** This assessment was interesting			**.863**		
11. **Enjoyment:** Participation in this assessment was a positive experience			**.690**		
13. **Enjoyment:** I did NOT enjoy taking this assessment (r)			**.326**		
18. **Effort:** I did NOT try as hard on this assessment as I normally would when taking an assessment at university (r)	**.456**			.599	
9. **Fairness:** This assessment was biased against test-takers who do not have strong language skills in English (r)				**.524**	
10. **Fairness:** Overall, the assessment was fair				**.371**	.358
2. **Face Validity:** It was clear to me that this assessment is related to someone's ability to deal with challenging social situations that may be encountered in the teaching profession					**.718**
1. **Face Validity:** The content of this assessment was clearly related to the job of a primary school teacher					**.401**
4. **Face Validity:** There was NO real connection between this assessment and the job of a primary school teacher (r)					**.333**

Note. Different colours in factor loadings indicate different underlying factors, as they were initially conceptualised. Only loadings greater than .3 are presented; r = reversed.

Table F2

EFA for the validity- and fairness-related statements: Factor Matrix

	Factor		
	1	2	3
2. **Face Validity:** It was clear to me that this assessment is related to someone's ability to deal with challenging social situations that may be encountered in the teaching profession	**.605**	.478	-.402
6. **Fairness:** This assessment was a fair indicator of someone's knowledge of how to deal with challenging social situations that may be encountered in the teaching profession	**.555**	-.411	
7. **Fairness:** This assessment gave me the opportunity to demonstrate my knowledge of how to deal with challenging social situations in teaching	**.488**		
8. **Fairness:** This assessment would NOT afford everyone the same opportunity to demonstrate their knowledge of how to deal with challenging social situations in teaching (r)	**.478**		
10. **Fairness:** Overall, the assessment was fair	**.434**		
3. **Face Validity:** A person who can successfully tackle challenging social situations in teaching would do well on this assessment	**.373**		
4. **Face Validity:** There was NO real connection between this assessment and the job of a primary school teacher (r)	**.320**	.304	
5. **Face Validity:** A person's overall performance on this assessment would predict how well they deal with challenging social situations at school	**.384**	-.426	
1. **Face Validity:** The content of this assessment was clearly related to the job of a primary school teacher		.375	
9. **Fairness:** This assessment was biased against test-takers who do not have strong language skills in English (r)	**.310**	.315	.670

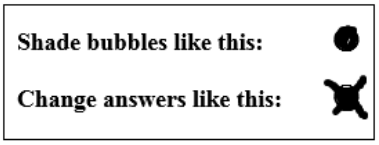
Note. r = reversed

Appendix G: The Reading Comprehension Test

Reading comprehension test

Anastasios Karakolidis, PhD Candidate
anastasios.karakolidis@dcu.ie

Directions: Each passage in this group is followed by questions based on its content. After reading a passage, choose the best answer to each question. Answer all questions following a passage on the basis of what is stated or implied in



Passage 1

Zooplankton, tiny animals adapted to an existence in the ocean, have evolved clever mechanisms for obtaining their food, miniscule phytoplankton (plant plankton). A very specialized feeding adaptation in zooplankton is that of the tadpolelike appendicularian who lives in a walnut-sized (or smaller) balloon of mucus equipped with filters that capture and concentrate phytoplankton. The balloon, a transparent structure that varies in design according to the type of appendicularian inhabiting it, also protects the animal and helps to keep it afloat. Water containing phytoplankton is pumped by the appendicularian's muscular tail into the balloon's incurrent filters, passes through the feeding filter where the appendicularian sucks the food into its mouth, and then goes through an exit passage. Found in all the oceans of the world, including the Arctic Ocean, appendicularians tend to remain near the water's surface where the density of phytoplankton is greatest.

1. It can be inferred from the passage that which of the following is true of appendicularians?

(A) They are exclusively carnivorous

(B) They have more than one method of obtaining food

(C) They can tolerate frigid water

(D) They can disguise themselves by secreting mucus

(E) They are more sensitive to light than are other zooplankton

2. The author is primarily concerned with

(A) explaining how appendicularians obtain food

(B) examining the flotation methods of appendicularians

(C) mapping the distribution of appendicularians around the world

(D) describing how appendicularians differ from other zooplankton

(E) comparing the various types of balloons formed by appendicularians

3. According to the passage, all of the following are descriptive of appendicularians EXCEPT

(A) tailed

(B) vegetarian

(C) small-sized

(D) single-celled

(E) ocean-dwelling

4. The passage suggests that appendicularians tend to remain in surface waters because they

(A) prefer the warmer water near the surface

(B) are unable to secrete mucus at the lower levels of the ocean

(C) use the contrast of light and shadow at the surface to hide from predators

(D) live in balloons that cannot withstand the water pressure deeper in the ocean

(E) eat food that grows more profusely near the surface

GO ON TO THE NEXT PAGE

Passage 2

"Popular art" has a number of meanings, impossible to define with any precision, which range from folklore to junk. The poles are clear enough, but the middle tends to blur. The Hollywood Western of the 1930's, for example, has elements of folklore, but is closer to junk than to high art or folk art. There can be great trash, just as there is bad high art. The musicals of George Gershwin are great popular art, never aspiring to high art. Schubert and Brahms, however, used elements of popular music - folk themes - in works clearly intended as high art. The case of Verdi is a different one: he took a popular genre - bourgeois melodrama set to music (an accurate definition of nineteenth-century opera) - and, without altering its fundamental nature, transmuted it into high art. This remains one of the greatest achievements in music, and one that cannot be fully appreciated without recognizing the essential trashiness of the genre.

As an example of such a transmutation, consider what Verdi made of the typical political elements of nineteenth-century opera. Generally in the plots of these operas, a hero or heroine - usually portrayed only as an individual, unfettered by class - is caught between the immoral corruption of the aristocracy and the doctrinaire rigidity or secret greed of the leaders of the proletariat. Verdi transforms this naive and unlikely formulation with music of extraordinary energy and rhythmic vitality, music more subtle than it seems at first hearing. There are scenes and arias that still sound like calls to arms and were clearly understood as such when they were first performed. Such pieces lend an immediacy to the otherwise veiled political message of these operas and call up feelings beyond those of the opera itself.

Or consider Verdi's treatment of character. Before Verdi, there were rarely any characters at all in musical drama, only a series of situations which allowed the singers to express a series of emotional states. Any attempt to find coherent psychological portrayal in these operas is misplaced ingenuity. The only coherence was the singer's vocal technique: when the cast changed, new arias were almost always substituted, generally adapted from other operas. Verdi's characters, on the other hand, have genuine consistency and integrity, even if, in many cases, the consistency is that of pasteboard melodrama. The integrity of the character is achieved through the music: once he had become established, Verdi did not rewrite his music for different singers or countenance alterations or substitutions of somebody else's arias in one of his operas, as every eighteenth-century composer had done. When he revised an opera, it was only for dramatic economy and effectiveness.

5. The author refers to Schubert and Brahms in order to suggest

(A) that their achievements are no less substantial than those of Verdi

(B) that their works are examples of great trash

(C) the extent to which Schubert and Brahms influenced the later compositions of Verdi

(D) a contrast between the conventions of nineteenth-century opera and those of other musical forms

(E) that popular music could be employed in compositions intended as high art

6. According to the passage, the immediacy of the political message in Verdi's operas stems from the

(A) vitality and subtlety of the music

(B) audience's familiarity with earlier operas

(C) portrayal of heightened emotional states

(D) individual talents of the singers

(E) verisimilitude of the characters

7. According to the passage, all of the following characterize musical drama before Verdi EXCEPT

(A) arias tailored to a particular singer's ability

(B) adaptation of music from other operas

(C) psychological inconsistency in the portrayal of characters

(D) expression of emotional states in a series of dramatic situations

(E) music used for the purpose of defining a character

8. It can be inferred that the author regards Verdi's revisions to his operas with

(A) regret that the original music and texts were altered

(B) concern that many of the revisions altered the plots of the original work

(C) approval for the intentions that motivated the revisions

(D) puzzlement, since the revisions seem largely insignificant

(E) enthusiasm, since the revisions were aimed at reducing the conventionality of the operas' plots

GO ON TO THE NEXT PAGE

9. According to the passage, one of Verdi's achievements within the framework of nineteenth-century opera and its conventions was to

(A) limit the extent to which singers influenced the musical composition and performance of his operas

(B) use his operas primarily as forums to protest both the moral corruption and dogmatic rigidity of the political leaders of his time

(C) portray psychologically complex characters shaped by the political environment surrounding them

(D) incorporate elements of folklore into both the music and plots of his operas

(E) introduce political elements into an art form that had traditionally avoided political content

10. Which of the following best describes the relationship of the first paragraph of the passage to the passage as a whole?

(A) It provides a group of specific examples from which generalizations are drawn later in the passage

(B) It leads to an assertion that is supported by examples later in the passage

(C) It defines terms and relationships that are challenged in an argument later in the passage

(D) It briefly compares and contrasts several achievements that are examined in detail later in the passage

(E) It explains a method of judging a work of art, a method that is used later in the passage

11. It can be inferred that the author regards the independence from social class of the heroes and heroines of nineteenth-century opera as

(A) an idealized but fundamentally accurate portrayal of bourgeois life

(B) a plot convention with no real connection to political reality

(C) a plot refinement unique to Verdi

(D) a symbolic representation of the position of the bourgeoisie relative to the aristocracy and the proletariat

(E) a convention largely seen as irrelevant by audiences

Thank you!

Appendix H: Ethics Approval Letter

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

DCU

Mr Anastasios Karakolidis

School of Policy and Practice

DCU Institute of Education

3 January 2018

REC Reference: **DCUREC/2017/206**

Proposal Title: **The use of animations to assess practical knowledge**

Applicant(s): Mr Anastasios Karakolidis, Professor Michael O'Leary

Dear Anastasios,

This research proposal qualifies under our Notification Procedure, as a low risk social research project. Therefore, the DCU Research Ethics Committee approves this project.

Materials used to recruit participants should state that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee.

Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,

Donal O' Gorman

Dr Dónal O'Gorman

Chairperson

DCU Research Ethics Committee

DCU
Research & Innovation

Taighde & Nuálaíocht Tacaíocht
Ollscoil Chathair Bhaile Átha Cliath,
Baile Átha Cliath, Éire

Research & Innovation Support
Dublin City University,
Dublin 9, Ireland

T +353 1 700 8000
F +353 1 700 8002
E research@dcu.ie
www.dcu.ie

Appendix I: Results Using the Extended Scales

As explained in Chapter 3, experienced teachers from Greece and Ireland were asked to rate the suitability of every practice statement to inform the scoring of pre-service teachers' responses on the PK-SJT. Comparisons between native (i.e., Irish) and non-native (i.e., Greek) English speakers were important in the context of this study, thus a common 16-item scoring scale, for which experienced teachers' ratings from the two countries did not statistically significantly differ, was developed to assess pre-service teachers' practical knowledge. However, it should be appreciated that analyses for which comparisons between native and non-native speakers was required could also have been conducted for each country separately using all the available practice statements that were categorised as either *Good* or *Bad* by the experienced teachers in each country. The *Good* and *Bad* items for each country are presented in Appendix C (Table C1).

This section provides supplementary results for the research questions 1.2 and 1.3, based on analyses conducted using these extended PK-SJT scales. The scale for the Greek participants included their responses to ten items categorised as *Bad* practices and nine categorised as *Good* practices, based on the Greek teachers' ratings; the practical knowledge scale for the Greek pre-service teachers ranged from 19 to 95. The performance scale for the Irish participants included 14 *Bad* practice and 22 *Good* practice items, based on the Irish teachers' ratings; the practical knowledge scale for the Irish pre-service teachers ranged from 36 to 180. Table I1 provides the minimum, maximum and mean scores along with the standard deviations and reliability levels for each scale.

Table I1

Information about the extended PK-SJT scales for each country

	Minimum	*Maximum*	*M*	*SD*	Cronbach's alpha
Extended PK-SJT scale for Greek participants	*59*	*91*	75.31	6.42	.699
Extended PK-SJT scale for Irish participants	121	167	148.26	7.75	.800

Research question A1.1: *Is the performance of non-native English speakers on the PK-SJT related to their level of proficiency in English?*

As mentioned in Chapter 4, based on non-native English speakers' responses regarding their level of proficiency in English, two groups of English proficiency were formed: (i) the advanced English speakers (i.e., participants who had English level of C1 or above) and (ii) the non-advanced English speakers (i.e., participants who had English level of B2 or below). To answer this research question, an Independent-Samples T-test comparing these two groups in terms of their PK-SJT performance was applied, using the extended scale for Greek pre-service teachers. The analysis showed that advanced English speakers (M = 78.39, SD = 5.64) outperformed their non-advanced counterparts (M = 73.58, SD = 6.22) in the PK-SJT assessment, $t(76) = -3.385$, $p = .001$. The effect size of this gap was large, $d = 0.80$.

This finding is consistent with the main findings of the study, where the common 16-item scale was used, as presented in Chapter 4.

Research question A1.2: *Is the relationship between non-native English speakers' performance and their level of proficiency in English weaker in the case of the animated PK-SJT?*

Running separate analysis for those who took the animated and the text-based version of the test, it was found that the average performance difference between advanced and non-advanced non-native English speakers, although statistically significant in both formats, was smaller in the case on the animated PK-SJT (mean difference of 4.70 score-points) compared to those taking the text-based PK-SJT (mean difference of 5.37 core points); animated PK-SJT: $t(38) = -2.354$, $p = .024$, text-based PK-SJT: $t(36) = -2.744$, $p = .009$. Figure I1 illustrates these differences.

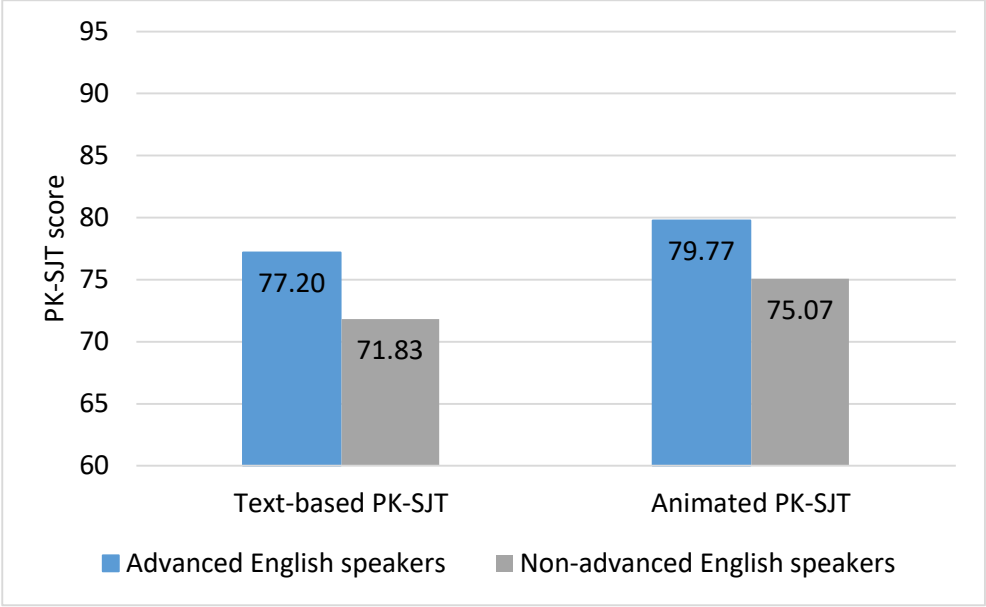


Figure I1. Advanced and non-advanced non-native English speakers' performance across test formats (extended PK-SJT scale).

To examine whether the performance gap between advanced and non-advanced non-native English speakers was significantly smaller in the case of the animated assessment, a hierarchical multiple linear regression with participants' PK-SJT as the outcome variable was conducted (Table I2). *Test format* and *English proficiency* (advanced/non-advanced English speaker) were entered as predictor variables in step one, whilst the *Test format*English proficiency* interaction term was added in the model in step two, while controlling for the original variables.

Table I2

Test format - Proficiency in English regression model (non-native English speakers, extended PK-SJT scale)

Predictors	*B*	*SE B*	*β*	R^2
Step 1				.186
Test format	3.004*	1.333	0.235	
Proficiency in English (Advanced)	5.040**	1.389	0.379	
Step 2				.187
Test format*Proficiency in English	-0.679	2.795	-0.040	

Note. The analysis was based on a sample of 78 test-takers.
*$p < .05$. **$p < .01$.

The regression analysis results showed that the interaction term was not statistically significant. Even though the animation of the text-based test resulted in reducing the gap

between advanced and non-advanced English speakers, the interaction effect was not large enough to be statistically significant.

This finding is also consistent with the main findings of the study where the common 16-item scale was used, as presented in Chapter 4.

Research question A2.1: *Is test-takers' PK-SJT performance related to their performance on the reading comprehension test?*

This relationship was investigated within the two subgroups of the study (i.e., native and non-native English speakers) using the extended scales for each country, namely the 19-item PK-SJT scale for Greece and the 36-item PK-SJT scale for Ireland. Reading comprehension ability was a statistically significant predictor of practical knowledge, only in the case of non-native English speakers; non-native speakers: $r(78) = .422$, $p < .001$, native speakers: $r(51) = .149$, $p = .296$[34].

This finding is consistent with the main findings of the study where the common 16-item scale was used, as presented in Chapter 4.

Research question A2.2: *Is the relationship between test-takers' PK-SJT performance and their performance on the reading comprehension test weaker in the case of the animated PK-SJT?*

The results pertaining to research question A2.2 were slightly different when the extended PK-SJT scales for each country were used, compared to the results provided in Chapter 4. More specifically, although the impact of reading skills on non-native speakers' performance was smaller in the case of the animated assessment, animated PK-SJT, $r(40) = .389$, $p = .013$, text-based PK-SJT: $r(38) = .471$, $p = .003$, the same was not the case for native English speakers, animated PK-SJT: $r(26) = .320$, $p = .111$, text-based PK-SJT: $r(25) = .063$, $p = .763$[35]. In other words, the use of animations helped to reduce the adverse impact of reading comprehension skills on practical knowledge performance only in the case of non-native English speakers, while the opposite was the case for native English speakers. This is illustrated in Figures I2 and I3. However, none

[34] Subgroup analysis results should be interpreted with more caution because of the smaller number of cases in each group, which may lead to reduced power.
[35] Subgroup analysis results should be interpreted with more caution because of the smaller number of cases in each group, which may lead to reduced power.

of the correlations between PK-SJT performance and reading comprehension ability was statistically significant for the native English-speaking sample.

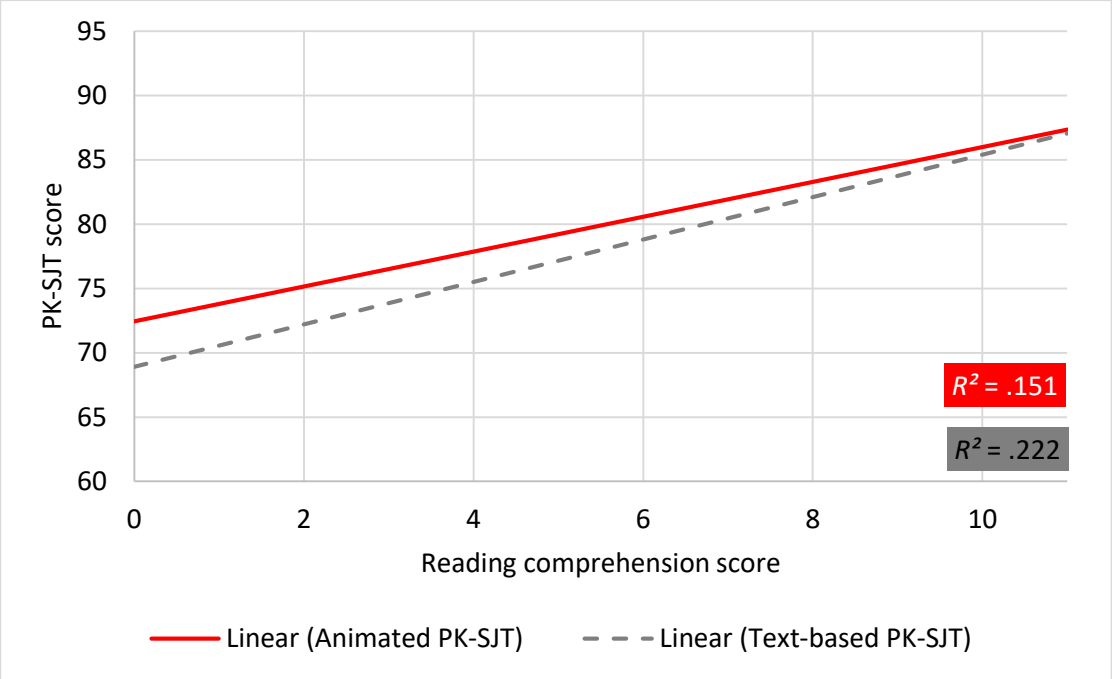


Figure I2. Correlation between reading comprehension and PK-SJT performance (non-native English speakers, extended PK-SJT scale).

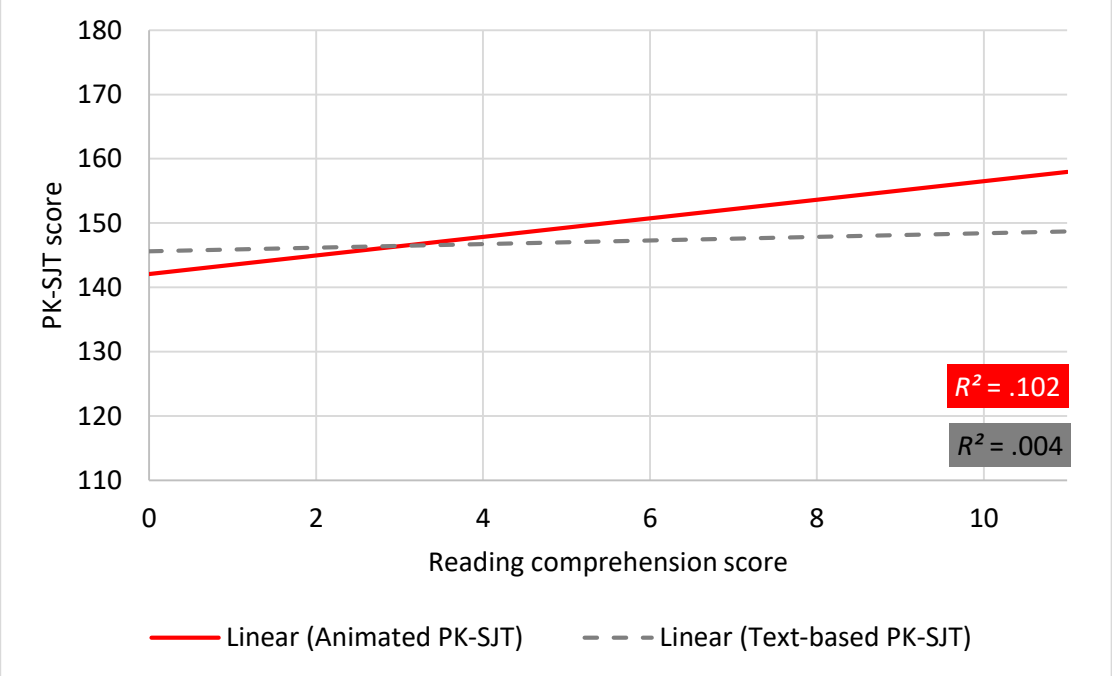


Figure I3. Correlation between reading comprehension and PK-SJT performance (native English speakers, extended PK-SJT scale).

Hierarchical linear regression that was applied separately only for the group of non-native English speakers, where the impact of reading skills on PK-SJT performance was

statistically significant, using the 19-item PK-SJT scale, revealed similar results regarding the interaction between test format and participants' reading comprehension skills. Again, the interactions between the two variables were negative, indicating a reduced adverse impact when animations were used, but this was not statistically significant. Table I3 summarises the results of the regression model.

Table I3

Test format - Reading comprehension regression model (non-native English speakers, extended PK-SJT scale)

Predictors	*B*	*SE B*	*β*	R^2
Step 1				.220
Test format	2.619*	1.301	0.205	
Reading comprehension	1.501**	0.364	0.420	
Step 2				.222
Test format*Reading comprehension	-0.291	0.733	-0.091	

Note. The analysis was based on a sample of 78 test-takers.
**p < .05. **p < .01.*

In most of the above cases, the results using the extended scales for each country were identical to the results as presented in the main body of this thesis. Hence, the overall conclusions regarding the effectiveness of animation in reducing construct-irrelevant variance would be similar in both cases.