Id: 927114

Participatory Educational Research (PER)
Vol.9(4), pp. 33-52, July 2022
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.22.78.9.4

Investigating Animation-Based Achievement Tests According to Various Variables

Elif GÜVEN DEMİR[*]
Primary Education Department, Düzce University, Düzce, Turkey
ORCID: 0000-0001-6685-5341

Yücel ÖKSÜZ
Educational Sciences Department, Ondokuz Mayıs University, Samsun, Turkey
ORCID: 0000-0002-9310-7506

Article history

Received:
24.04.2021

Received in revised form:
20.12.2021

Accepted:
10.01.2022

Key words:
animation; achievement; open-ended test; multiple-choice test, science

This research aimed to investigate animation-based achievement tests according to the item format, psychometric features, students' performance, and gender. The study sample consisted of 52 fifth-grade students in Samsun/Turkey in 2017-2018. Measures of the research were open-ended (OE), animation-based open-ended (AOE), multiple-choice (MC), and animation-based multiple-choice (AMC) achievement tests developed for the unit "Motions of the Earth" of the subject area "Earth and Universe" of the science course. Tests were compared to item difficulty, item discrimination, internal consistency levels, and students' performance and gender. Additionally, predicting students' science grades was tested by animation-based open-ended and multiple-choice tests. Paired sample t-tests, Pearson's correlation coefficient, Friedman test, Wilcoxon test, two-way mixed ANOVA tests, and multiple regression analysis were applied to analyze research data. Research results show no significant difference between animation-based and traditional tests' validity and reliability levels. Research result also shows that students' test scores were significantly higher at animation-based tests than traditional tests. Besides, students' test scores differed according to the animation type used in tests. However, gender was not a significant variable on students' test scores. Finally, despite using animation as a significant predictor of Science grades, just animation based multiple-choice test has significantly predicted students' Science grades. Future research can investigate variables that affect students' participation in animation-based tests and their opinions on animation-based tests. The finding regarding the animation type variable can be investigated in-depth in terms of the effect levels of the visual and auditory elements added to the animations by including an equal number of items on the test.

Introduction

Research shows that the digital content included in the educational process develops and increases the students' interest and motivation for learning (Burke, Snyder, & Rager,

[*] Correspondency: elifguvendemir@duzce.edu.tr

2009; Chao, Chen, Star, & Dede, 2016). Mayer (2005) states that individuals learn better with words and pictures instead of just words. It is easy to recognize, memorize, and recall pictures evoking both a verbal code and a visual code than the words, that words enter long-term memory with one verbal code (Dewan, 2015). The dual encoding of pictures advocates easily remembered, as two stored representations potentially cause a higher probability of retrieval success (Ally, 2012). In other words, it would be fair to state that it is easier to place and recall multiple stimuli in memory. The teaching technologies are involved in the inclusion of these multiple stimuli in the training process.

Instruments used to deliver an instructional message and the representational formats used to present the instructional message or the sense modalities the learner uses to receive the instructional message are the ways of arose the term multimedia (Mayer, 2009). Additionally, the animations may offer a multimedia opportunity to be used in educational processes since they address the visual and auditory senses. Indeed, there are many studies on the use of animations in educational processes in the literature (Chan, 2015; Daşdemir & Doymuş, 2016; Ercan, Bilen, & Bulut, 2014). Research results show that animations in education increase students' success and motivation levels (Hwang, Tam, Lam, & Lam, 2012; Kayaoglu, Akbaş, & Öztürk, 2011; Yezierski & Birk, 2006). Dalacosta, Kamariotaki-Paparrigopoulou, Palyvos, and Spyrellis (2009) have found that the use of animation increases the level of knowledge of specific science subjects that often have misconceptions and are challenging to understand. Additionally, movements in the animation could cause excitement to stimulate positive attitudes towards Science classes, which, in turn, may have motivated students to continue focusing; thus, animations may encourage emotional engagement (Stromme & Mork, 2020). Besides, it is stated that animations activate three learning styles, such as visual, auditory, and kinesthetic, at the same time and increase the knowledge comprehension levels (Barak, Ashkar, & Dori, 2011). Stith (2004) states that the function of controlling animation with the mouse enables the visual learners and the students who tend to learn kinesthetically to be included in the learning process. At this point, one can suggest that animations with their mentioned superior sides cannot be limited to only the teaching dimension of the educational process and can be integrated into the evaluation dimension.

Measurement and evaluation are an integral part of the teaching and learning process, enabling the evaluation and development of this process and the quality to be increased in the long term (Ferrari, Cachia, & Punie, 2009). Cachia, Ferrari, Ala-Mutka, & Punie (2010) state that the current changes cannot be effective if the evaluation method remains the same while the program and training objectives change. As a matter of fact, traditional measurement, and evaluation methods such as multiple-choice tests have various limitations (Wu, Chang, Chen, Yeh, & Liu, 2010). Particularly, it is difficult to reflect abstract concepts at the micro and macro levels, such as atomic molecule theory in the science field, Earth, Sun, and Moon, and depict traditional test items (Wu et al., 2010). In traditional tests, it is difficult to reflect real-life circumstances to test items (Wu et al., 2010).

The text, video (in which characters are actors), and animations (in which characters are computer-generated avatars) are testing formats most commonly used in the measurement and evaluation process (Pop, Tuzinski, & Fetzer, 2016). The use of longer, more complex pieces of text with difficult vocabulary and sentence structure are issues that emerge when it comes to the measurement of more sophisticated or higher-order skills via text-based assessment (Karakolidis, O'leary, & Scully, 2021).

When it comes to important features that are difficult to highlight in the video, animations are preferred more than videos (Pop, et al., 2016). Animations, such as 3D virtual reality and real-world simulation, present visual cues for students to enhance their cognitive processing (Chan,2013). Animations can be changed and modified relatively easily making it possible to correct errors and/or to keep the instruments up to date over time as the need for use in different countries and cultures (Karakolidis, Scully, & O'Leary, 2021). Besides, the video-based evaluation has limitations as creating a role-playing-dependent process and finding the same actors when updating is needed; however, animation-based testing is more flexible and less expensive in these respects (Pop, et al., 2016).

There are lots of benefits of using animations in the measurement and evaluation process (Dancy & Beichner, 2006; Wouters, Paas, & van Merrienboer, 2008; Wu et al., 2010). First, animations are descriptive presentation tools and can be effective in the presentation of abstract subjects like ebb and flow (Wu et al., 2010). Chan (2015) also emphasizes that animations are effective in visualizing abstract concepts in areas such as Science and mathematics. Secondly, animations can be used instead of reality and can be effective in presenting complex subjects of real-life (Dancy & Beichner, 2006; Wu et al., 2010). Third, animations can increase the motivation of the students with the help of visual attraction (Wouters, Paas, & van Merrienboer, 2008). In addition, it is stated that animation-based tests increase students' test completion levels and interest in taking the test according to traditional methods (Wu et al., 2010). Dancy and Beichner (2006) state that animations are very valuable, not only visually but also as a part of the whole problem and with the ability to clarify the situation caused by misreading or misinterpreting the question in the standard questionnaire. Dancy (2000) has found that misconceptions due to difficulty in reading or problems of uncertainty caused by the problem decrease with the animation-based test.

According to the above mentioned, it is seen that animation-based tests have superior aspects in various aspects. However, Karakolidis, Scully, and O'Leary (2021) present the key issues that need to be considered when developing animated tests as value-added, the fidelity of representation, facial expressions, voice, and movement, cognitive load, situational contexts, response options, testing platform, and financial cost. The effects of animation-based tests may change depending on these concepts.

In the literature, it is seen that there are limited numbers of studies about animation-based assessment. Dancy (2000) compared the animated version of the Force Concept Inventory and the traditional one. Wu et al. (2010) compared an animation-based test and the graphics-based test about earth sciences. King (2015) developed a picture-based assessment tool that evaluates trauma history and symptomatology in children. Both these studies showed that animation often had a positive effect on the assessment process (Dancy, 2000; Wu et al., 2010). However, Dindar et al. (2015) have concluded that using animation did not significantly affect students' success in the study, which compared the use of static graphics and animated graphics in the multiple-choice tests. Recently, Bardach, Rushby, Kim, & Klassen (2021) compared three conditions in a prospective teacher selection test: video-based (3D animated) with text, video-based (3D animated) without text, and text-based. Results showed that test takers' feeling of engagement was higher on animation-based than text-based tests, but there was no significant difference in student performance. Similarly, Karakolidis, O'Leary & Scully (2021) compared an animated-based and text-based test that measured practical knowledge in the teaching profession of primary school teachers. Results showed that variance attributed to construct-irrelevant factors was lower at the animation-based test than the text-based test, and participants perceived it to be more valid, fair, and enjoyable than those who took the text-based test (Karakolidis, O'Leary & Scully, 2021). Due to the differing research results and lack of evidence from the younger age groups' performance associated with a school course, there is a need for further studies on animation-based assessment and evaluation studies. Additionally, there is still need evidence for test item format, psychometric features, and animation type about using animations in the assessment and evaluation process.

Unlike some previous studies, animations used in this study not only the visualization of science subjects but involve daily life circumstances in a given scenario, such as playing snowball, watching clouds, seasonal changes in the weather and environment, observing the movement of the Sunlights. Besides, animations contain question sentences embodied in the animation and vocalized by a cartoon character as a girl, boy, or an adult, depending on the animation scene. In other words, the animated characters vocalized questions and items in animation-based tests.

Additionally, it was aimed to prevent failure caused by the reading disorder or other verbal skills by using an animated character to vocalize the question items without using on-screen text. Insufficient reading comprehension may prevent the use of other skills to a certain extent. Bayat, Şekercioğlu, and Bakır (2014) indicate that reading comprehension is an enabling function that provides success in finding answers in Science test items. In this way, it is aimed to have more advantages related to the assessment process.

Mayer (2009) discusses whether attaching an image of the speaker to the screen - such as a cartoon-like character or even a video of a talking head - helps students learn more in-depth than a multimedia lesson in the context of the image principle. There is evidence that students were more successful who interacted with an animated pedagogical agent who spoke to them than with those of students who received similar explanations as on-screen text without the intervention of a pedagogical agent (Moreno, Mayer, Spires, & Lester, 2001). Pedagogical agents that interacted through text rather than narration and contained some form of animation allow more learning benefits than agents that did not (Schroeder, Adesope, and Gilbert, 2013). Besides, van der Meij, van der Meij, and Harmsen (2015) provide evidence that external properties of animated pedagogical agents show that students' science motivation can be positively affected by the embodied presence (i.e., image and voice) of an agent. Mayer (2009) suggests that when an on-screen pedagogical agent points to a relevant part of the graphic, this may canalize the student's visual attention, reducing external cognitive processing. At this point, the use of animated characters in animation-based questions is expected to attract students' attention to the question and increase the students' test success and contribute positively to the validity and reliability of the test. At this point, the effect of the animation type on the measurement process was determined as one of the research questions. The animations were designed with an animated character who asks the questions and no animated character to see the effect of using an animated character on the screen. Child and adult animated characters were used and presented on the screen in animations depending on the events in the scenario. Besides, some animations have been designed without any animated characters just as a simulation of science concepts. In these animations, questions were vocalized, but an animated character was not used.

The results of a meta-analysis study towards pedagogical agents conducted by Schroeder, Adesope, and Gilbert (2013) show that participants benefited more from a system with a pedagogical agent than a system without an agent and felt they were engaged in social interaction with the pedagogical agent. Nass and Brave (2005) present evidence that people may be more impressed by online spoken messages when they sense the speaker's voice coming from someone similar to their gender, race, ethnicity, or emotional state. Liu and Elms (2019) concluded that character design, voice acting, and dialogues in animated videos are found vital to improving student engagement in the learning process. At this point, it is thought that the feeling of closeness to the voice or image may affect the students' test-taking willingness and so success. Hence gender was investigated as a variable to affect students' success in this study. Another factor that may have been related to test-taking willingness is using familiar contexts, which may affect the students' success. Martins and Vaiga (2001) suggest that using contexts familiar to students and the use of everyday contexts are promising, helpful, and motivating ways to advance the learning of concepts with which they are connected and Science.

The starting point of the current study was to assess the factors relevant to the success of animation-based tests. The item type, item discrimination, item difficulty, predicting success grade, animation type, and gender were the variables tested in the scope of the research.

Multiple-choice tests are tests in which the respondents do not give the answers, the correct answer is given among the options by the tester and the respondents are asked to find the correct answer (Güler, 2017). The advantages of the multiple-choice tests are, it is easy to increase content validity due to the fact that a large number of items can be included and providing scoring reliability in large groups and objective and fast scoring (Güler, 2017; Klufa, 2015; Polat, 2020; Turgut & Baykul, 2012). However, in a multiple-choice test, students are able to get high scores by estimation, and therefore, the actual skill level may not be measured or identified the underlying causes which students' reasons for choosing the options they marked (Bush, 2001; Karataş, Köse, & Coştu, 2003).

The measurement methods based on open-ended test items to avoid these limitations can be used. Thus, student performance can be evaluated according to the strategy and justification skills used, not only based on the correct answer (Birgili, 2014). Open-ended tests eliminate the possibility of guessing the correct answer, measure students' perception, reasoning, and knowledge, and give information about the ability to use the knowledge (Badger & Thomas, 1992; Polat, 2020). However, taking a long time to answer open-ended items may prevent asking too many questions and negatively affect content validity, and the necessity of giving the answers in writing may cause the characteristics that are not wanted to be measured to reflect in the scoring (Öksüz & Güven Demir, 2019; Temizkan & Sallabaş, 2011).

The type of test item, with its distinct advantages and disadvantages, can affect the psychometric properties of animation-based tests and their impact on student performance in different ways. Therefore, animation-based open-ended and multiple-choice tests were compared with the traditional use of tests in terms of those variables in the current study. Besides, the power to predict students' Science success grades were tested in animation-based tests. The specific questions that will be addressed follow.

Is there any significant difference between animation-based and traditional tests' validity and reliability levels (open-ended and multiple-choice tests without animations)?

(1) Is there any significant difference between animation-based test and traditional test scores of students?
(2) Is there any significant difference between students' test scores according to the used animation type in animation-based tests?

(3) Is there any significant difference between traditional test scores and animation-based test scores of students according to the gender variable?
(4) Do animation-based test scores predict the Science success grade of students significantly?

Method

Participants

The study was carried out with 52 (girl: 22, boy:30) 5th-grade secondary school students aged 10-11 years who attend state schools 2017-2018 academic year in Samsun, Turkey.

Measurement

Measures of this study were open-ended (OE), animation-based open-ended (AOE), multiple-choice (MC), and animation-based multiple-choice (AMC) achievement tests developed for Earth and Universe concept at Science classes by Güven Demir (2018) and Öksüz & Güven Demir (2019). Tests were developed by authors in previous studies but used in the current study with different participants and research purposes. Animations used in tests were developed by Güven Demir (2018) in a doctoral dissertation to measure students' academic achievement levels on Earth rotation and revolution concept. Additionally, OE and MC were developed by Öksüz & Güven Demir (2019) in a previous study, which was about comparing students' performance according to the test items.

AOE and AMC consist of 16 2D vector-based animations with open-ended and multiple-choice question items. OE and MC were the same with AOE and AMC tests, exclude animations. While questions in AOE and AMC were presented via animations, tests were applied as paper-based with static images in OE and MC. A rubric was used to score open ended tests developed by Güven Demir (2018).

Animations used in AOE and AMC are the same animations, which is about three to five minutes in length and include cartoons about Earth rotation and revolution and animated child, adult characters were prepared by using Go Animate For School program. Some of the animations were designed in a story-based and events evolved around a child cartoon character (n:4) and an adult character (n:10) and while others included just simulation of science concepts without any character (n:2). Results of revolution and rotation of Earth, axial tilt were reflected in the childhood activities like playing snowball (see Figure 1) throughout the story in animations.

Figure 1. Sample Scenes from Animations

The difference between AOE and AMC was the type of question items asked by animated characters. The educational software and multimedia guidelines suggested in the literature were considered while preparing the animations (Denning, 1992; Koumi, 2006; Moussiades, Kazanidis & Iliopoulou, 2019; Schwartz & Hartman, 2007).

The item difficulty and discrimination proportions of the tests were calculated respectively .50, .50, .53, and .64; .73, .78, .89, and .81 for OE, AOE, MC, and AMC (see Table 2). Accordingly, it can be said that the current tests have medium difficulty and superior item discrimination levels (Ebel & Frisbie, 1991; Güler,2017; Quaigrain & Arhin,2017).

Cronbach Alpha values were calculated 0.68 for OE and 0.69 for AOE. Besides, Kuder-Richardson 20 values were calculated as 0.65 for MC and 0.69 for AMC. Authors made an extensive range of different descriptions to interpret alpha values (Taber, 2016). van Griethuijsen et al. (2015) adopted “0.7 or 0.6 as acceptable values for Cronbach Alpha in their article. Schmitt (1996) has put forward that there is no general level (such as 0.70) where alpha becomes acceptable, but rather that instruments with quite a low value of alpha can still prove useful in some circumstances. It is stated that the Cronbach Alpha value can decrease up to 0.60 (Hair, Black, Babin, & Anderson; 2010). Accordingly, it can be interpreted that the reliability coefficients regarding the tests are at an acceptable level.

The data collection process was conducted at two stages. Firstly, OE and following AOE were applied to the students during two lessons. In the first lesson OE, and in the second lesson, AOE was applied to the students. MC and following AMC were applied to the students four weeks later than OE and AOE. AOE and AMC consisted of 2D animations, and the questions were asked at the end of the video by animated characters differently from OE and MC.

Data Analysis

The z test was used to compare the item difficulty and item discrimination indices, and the Feldt test, which can be measured with the formula proposed by Alsawalmeh and Feldt (1994), was used in the comparison of the Cronbach Alpha and KR-20 reliability coefficients.

Paired sample t-tests were used to compare OE and AOE, MC, and AMC. Pearson’s correlation coefficient was computed to assess the relationship between AOE and AMC. The Friedman and Wilcoxon tests were conducted to compare the effect of animation type (child, adult, and simulation) on the AOE and AMC scores of students. Two-way mixed ANOVA tests were applied to compare OE and AOE according to the gender variable, as well MC and AMC. Multiple regression with using the enter method was carried out to investigate whether AOE and AMC could significantly predict students’ Science grades.

Results

Descriptive statistics and Shapiro-Wilk test results were presented in Table 1.

Table 1. Descriptive Statistics and Shapiro-Wilk Test Results Towards OE, AOE, MC and AMC

Tests	N	Mean	SD	Shapiro -Wilk
OE	52	12.4231	5.05401	.057*
AOE	52	13.5962	5.90533	.529*
MC	52	19.0769	5.74574	.242*
AMC	52	20.8077	5.66045	.366*

$^{*}p>.05$

Table 1 shows that the research data provide the normality assumptions required by parametric tests. Z tests were conducted to compare item difficulty proportions of OE, AOE, MC, and AMC (see Table 2).

Table 2. Z test Results Towards Comparing of Test Item Difficulty Proportions

Animation Type	Item No	Open-Ended Tests				Multiple Choice Tests			
		p_j^1	p_j^2	z	p	p_j^3	p_j^4	z	p
Child Character	1	0.89	0.89	0.01	0.984	0.53	0.60	-0.72	0.471
	2	0.07	0.17	-1.56	0.116	0.50	0.50	0	1
	3	0.50	0.50	0	1	0.50	0.50	0	1
	4	0.28	0.42	-1.49	0.133	0.50	0.50	0	1
Adult Character	5	0.25	0.25	0	1	0.32	0.39	-.745	0.453
	6	0.50	0.50	0	1	0.50	0.50	0	1
	7	0.50	0.50	0	1	0.50	0.50	0	1
	8	0.50	0.50	0	1	0.50	0.50	0	1
	9	0.50	0.50	0	1	0.67	0.75	-0.89	0.368
	10	0.92	0.67	3.15	0.001^{*}	0.96	0.96	0	1
	11	0.50	0.50	0	1	0.50	0.75	-2.63	0.008^{*}
	12	0.50	0.50	0	1	0.50	0.78	-2.9	0.002^{*}
	13	0.35	0.28	0.76	0.441	0.50	0.50	0	1
	14	0.78	0.78	0	1	0.50	0.50	0	1
Simulation	15	0.50	0.50	0	1	0.50	0.50	0	1
	16	0.50	0.50	0	1	0.50	0.50	0	1
Total	Average	0.50	0.50	0	1	0.53	0.64	-0.63	0.528

$^{*}p<.05$, p_j^1 = OE, p_j^2= AOE, p_j^3=MC, p_j^4=AMC

Research results show no significant difference between test item difficulty means of OE-AOE and MC - AMC. However, a significant difference was found between the item difficulty value of OE and AOE at the item numbered 10. There were significant differences between MC and AMC's item difficulty values at the items numbered 11 and 12. Z tests were conducted to compare item discrimination proportions of OE, AOE, MC, and AMC (see Table 3).

Table 3. Z test Results Towards Comparing of Test Item Discrimination Proportions

Animation Type	Item No	Open-Ended Tests				Multiple Choice Tests			
		r_j^1	r_j^2	z	p	r_j^3	r_j^4	z	p
Child Character	1	0.21	0.21	0	1	0.92	0.78	1.99	0.045^{*}
	2	1	1	0	1	1	1	0	1
	3	0.14	0.35	-2.48	0.012^{*}	1	1	0	1
	4	0.57	0.85	-3.14	0.001^{*}	1	1	0	1
Adult Character	5	0.50	0.50	0	1	0.64	0.78	-1.57	0.116
	6	1	1	0	1	1	1	0	1
	7	1	1	0	1	1	1	0	1
	8	1	1	0	1	1	1	0	1
	9	1	1	0	1	0.64	0.50	1.44	0.149
	10	0.14	0.64	-5.22	0^{*}	0.07	0.07	0	1
	11	1	1	0	1	1	0.50	5.88	0^{*}
	12	1	1	0	1	1	0.42	6.51	0^{*}
	13	0.71	0.57	1.48	0.136	1	1	0	1
	14	0.42	0.42	0	1	1	1	0	1
Simulation	15	1	1	0	1	1	1	0	1
	16	1	1	0	1	1	1	0	1
Total	Average	0.73	0.78	-0.32	0.741	0.89	0.81	0.63	0.528

$^{*}p<.05$, r_j^1=OE, r_j^2=AOE, r_j^3= MC, r_j^4=AMC

Research results show no significant difference between test item discrimination means of OE, AOE, MC, and AMC. However, discrimination values of OE and AOE differed significantly at the items numbered 3, 4, and 10. There were significant differences between MC and AMC's item discrimination values at items 1, 11, and 12.

Table 4. Feldt's Test Results Towards Comparing Test Reliability Levels

Tests	N	k	Cronbach Alpha /KR 20	W
OE	52	16	.68	.968
AOE	52	16	.69	
MC	52	16	.65	.885
AMC	52	16	.69	

Feldt's test results are seen in Table 4, show that reliability levels of OE and AOE, besides MC and AMC, did not differ significantly (W < F(51,51) = 1.59). This result may have been interpreted as using animation is not a significant variable to make a difference in the tests' reliability level.

Table 4. Paired Sample T-test Results

Tests	Mean	Std. Deviation	Std. Error Mean	t	df	Sig. (2-tailed)
OE- AOE	-1.17308	3.88926	.53934	-2.175	51	.034*

*p<.05

Research results show that OE (M:12.4, SD:5) and AOE (M: 13.5, SD: 5.9) scores differ significantly, t(51) = -2.175, p = 0.034, d= 0.209). This result shows that AOE scores are higher than OE scores mean and suggest that students were more successful at the AOE test than OE. The One-Way Repeated Measures ANOVA test was used to compare AOE scores according to the Animation Type (AT) (see Table 5).

Table 5. Friedman's Test Results Comparing AOE scores according to AT

	Animation Type	Mean	Sd	Mean Rank	df	X^2	p
AOE	Child	37.50	16.78877	2	2	16.41	.000*
	Adult	47.4038	24.56465	2.38			
	Simulation	27.8846	27.85372	1.62			

*p<.05

Results show that AOE scores of students differ significantly according to the used animation type (X2F=16.41, p=.000). Post-hoc tests using a Wilcoxon signed-rank test with a Bonferroni adjusted alpha level of .016 (.05/3) showed that students' AOE scores were higher at animations include adult (Mdn=50) than child (Mdn=37.5) character (T=300.5, z=-2.952, p=.003); and simulation (Mdn=25) character (T=245, z=-3.522, p=.000). However, AOE scores of students based on test items include child (Mdn=37.5) and simulation (Mdn=25) character did not differ significantly (T=389, z=-2.057, p=.040). The paired-sample t-test was used to compare the means of students' MC and AMC test scores (see Table 6).

Table 6. Paired Sample T-test Results

Tests	Mean	Std. Deviation	Std. Error Mean	t	df	Sig. (2-tailed)
MC-AMC	-1.73077	3.52089	.48826	-3.545	51	.001*

*p<.05

Results show that scores for MC level (M:19, SD:5.7) and AMC level (M: 20.8, SD: 5.6) differ significantly, t(51) = -3.545, p = 0.001, d= 0.303). This result shows that AMC scores are greater than MC scores mean and suggest that students were more successful at AMC test than MC. The One-Way Repeated Measures ANOVA test was used to compare AMC scores according to AT (see Table 8).

Table 7. Friedman Test Results Comparing AMC scores according to AT

	Animation Type	Mean	Sd	Mean Rank	df	X^2	p
AMC	Child	68.2692	25.7	2.25	2	21.799	.000*
	Adult	67.8846	18.5	2.24			
	Simulation	44.2308	35.2	1.51			

*p<.05

Results show that AMC scores of students differ significantly according to the used animation type (X2F=21.79, p=.000). Post-hoc tests using a Wilcoxon signed-rank test with a Bonferroni adjusted alpha level of .016 (.05/3) showed that students' AMC scores were lower at animations include simulation (Mdn=50) than child (Mdn=75) character (T=119, z=-3.850, p=.000); and adult (Mdn=70, T=138.5, z=-4.408, p=.000). However, AMC scores of students based on test items include child (Mdn=75) and adult (Mdn=70) character did not differ significantly (T=486, z=-.356, p=.721). The repeated-measures ANOVA was conducted to compare test scores OE and AOE across the gender variable (see Table 8).

Table 8. Two Way Mixed Design ANOVA on OE and AOE According to the Gender

	Mean square	df	F	Sig.	Partial Eta Squared
Intercept	38.708	1	5.071	0.029	0.092
Animation*gender	4.092	1	0.563	0.467	0.011
Error	7.633	50			

Table 8 shows the results for comparing scores of tests across using animations in girl and boy groups. The analysis showed that the OE and AOE scores of the students did not differ significantly in girl (MOE= 13.1, SDOE= 5.6; MAOE= 13.9, SDAOE= 5.6) and boy (MOE= 11.4, SDOE= 4; MAOE= 13, SDAOE= 5.7) groups, (F(1,50) = 0.536, p=0.467) across animations. Figure 2 shows a line graph with the growth trends for boy and girl groups' achievement scores, displaying each OE and AOE's achievement scores.

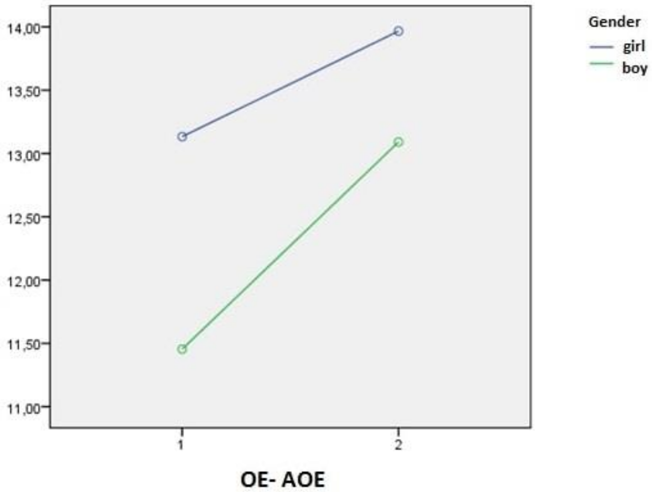

Figure 2. A-Line Graph On OE – AOE Test Scores According to The Gender

Table 9. Two Way Mixed Design ANOVA on MC and AMC According to the Gender

	Mean square	df	F	Sig.	Partial Eta Squared
Intercept	85.886	1	14.151	.000	.221
Animation*gender	12.655	1	2.085	.155	.040
Error	6.069	50			

Research results show that there was no significant difference between the MC and AMC scores of the students in girl (MMC= 19.6, SDMC= 5.6; MAMC= 20.8, SDAMC= 6.2) and boy (MMC= 18.2, SDMC= 5.9; MAMC= 20.8, SDAMC= 4.8) groups (F(1,50) = 2.085, p=0.155) across using animation. Figure 3 shows a line graph with the growth trends for boy and girl groups' achievement scores, displaying the achievement scores means for each of MC and AMC.

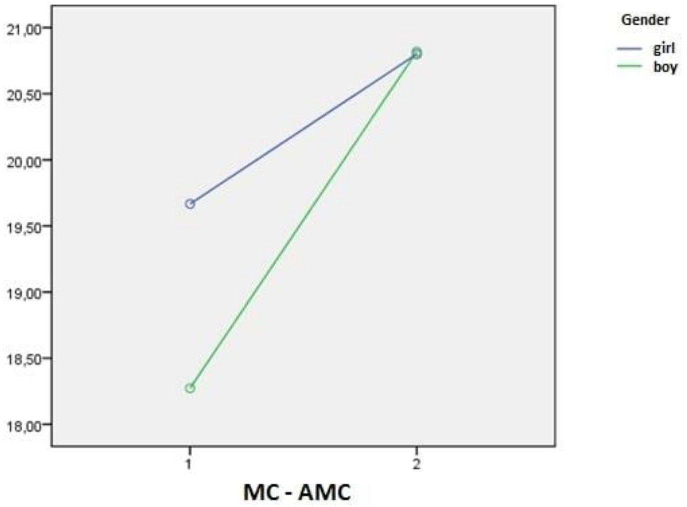

Figure 3. A line graph on MC – AMC test scores according to the gender

Table 10. Results for Variables Predicting Science Grade

Variable	B	SE B	β	t	p
Constant	58.360	5.828			
AOE	.103	.352	.045	.292	.771
AMC	1.361	.367	.574*	3.709	.001

R^2=.366, *p<.05

The regression results indicated that the model explained 36.6 % of the variance and that the model was a significant predictor of Science grade F(2,49) = 14.149, p=.000. The analysis shows that AOE did not significantly predict the students' science grades (B = .103, p=.771). However, AMC significantly predicts Science grade (B = 1.361, p<.05). The final predictive model was Science Grade = 58.360 + (.103*AOE) + (1.361*AMC).

Discussion

In this study, it was found that there was no significant difference between average item difficulty and discrimination values, and reliability levels of the traditional and animation-based tests. In other words, using animation has no significant effect on test validity and reliability levels. However, some of the items' item difficulty and discrimination level were significantly different depending on using animation on assessment than traditional tests. The validity and reliability of the animation-based and traditional tests were acceptable, yet the reliability level needed improvement. Contrary to our expectations, using animation on assessment would increase validity due to eliminating factors like verbal skills, language barriers, deficiencies in educational opportunities, and expressing abstract concepts rather than traditional tests. Both traditional and animation-based tests have similar validity reliability. Other research results supported this, showing that animation-based tests have similar validity and reliability levels with traditional tests (Dancy, 2000; Wu et al., 2010). However, there are research results show that students perceived more valid (higher face validity) the acted video-based tests than text-based ones (Chan & Schmitt, 1997; Karakolidis, O'leary, & Scully, 2021). These results were not based on item analysis but obtained from test takers' perceptions about the test as an indicator of face validity.

At the point reached, the effect of animations on validity may have depended on the students' features like verbal skills and prior knowledge who have participated in the study. Students who have high knowledge about the subject may not have been affected advantage of the animations on assessment. Mayer (2009) states that design effects are mostly stronger for low-knowledge participants than for high-knowledge participants. High-knowledge learners can form appropriate mental images from words and thus use their prior knowledge to meet for lack of guidance in the presentation (Mayer, 2009). As Wu et al. (2010) conducted in their study, high-knowledge learners could form mental representation from words or images alone; in contrast, low-knowledge learners may find it challenging to construct mental representation in working memory at the same time. Additionally, it is possible to deduce; using animation does not change the test's current validity. Malone and Brünken (2013) concluded that the problem of low criterion validity of the test could not be solved only by a change of presentation mode, which was animated versus static in their study. However, Dancy and Beichner (2006) state that they have found no indicators that computer animation can decrease an assessment instrument's validity, but animation can improve assessment under some conditions by increasing its validity. Using animations could increase the validity of questions since they tend to diminish a confounding variable's effect (Dancy, 2000).

Comparing means of test scores shows that animation-based test scores were higher than traditional test scores. In other words, students were more successful at animation-based open-ended and multiple-choice tests. Other research results in the literature support this finding. Karakolidis, O'leary, and Scully (2021) concluded that participants who completed the animated test performed significantly better than those who took the text-based version of the test. Wu, Yeh, and Chang (2010) also found that students' animation-based test scores were significantly higher than the traditional graphic-based test. Dancy and Beichner (2006) have found that students had a better comprehension of the intent of the question when viewing animation and gave an answer that was more representative of their actual comprehension than a static version of the test. Chan and Schmitt (1997) concluded that students performed better in the video-based test than the paper pencil-based test. Similarly, Wu et al. (2010) also found that animation-based test scores were higher than graphic-based test scores. Dancy and Beichner (2006) state that learners probably misread or misconstrue static questions with words and pictures than a question with information delivered in an animation. Therefore, eliminating verbal skills may have been helped students to answer correctly in animation-based tests. Similarly, English (2020) also has emphasized that the animations effectively assess student outcomes in Science concepts. Unlike the research topic, in another study, it was found that the response rate of the audio and animated cartoon questionnaire was significantly higher than the text-based questionnaire (Qiu, Xia, Tian, et al. 2020). Contrary to our finding, Bardach et al. (2021) found participants' test performance did not differ according to the using animation on the test. Malone and Brünken (2013) found that the participants' animated and static driver license test scores did not differ significantly.

Research results show that animation types used in animations significantly affect students' scores. Students' test scores were higher when evaluated as a whole, respectively in adult, child characters, and simulation-based test items. These findings corresponded to one of Mayer's multimedia design guidelines called personalization, voice, and image principles (Mayer, 2014). Specifically, Mayer recommended that multimedia instructional messages should be presented in conservation style rather than formal style; spoken words should come from a standard-accented human voice and use computer-based character while designing animated pedagogical agents. At this point, higher scores in both animation-based open-ended and multiple-choice tests may be derived from the advantages of conversational style of question words, spoken in standard accented in the human voice, and using animated character presented in the animations-based tests. In other words, personalization principle in multimedia learning expressed by Mayer (2014) may have been effective on students' higher success on animations used child and adult cartoon character than simulation. Atkinson, Mayer, and Merrill (2005) indicate that using verbal and visual social signs in computer-based learning environments can create a sense of partnership by encouraging students to consider their interaction with the computer to become more similar to what they would expect from a human-to-human conversation. Students may have experienced the same personalization principle towards child and adult animated characters. Atikonson (2002) found that students with higher test scores learn to solve mathematics problems via a cartoon-like character on screen rather than a no-character presented group. From another angle, having an animated agent, child, and adult cartoon character in this sample, present in a multimedia learning environment, can positively influence the learner's perceptions of their educational experience (Lester et al., 1997). Voice acting, character design, and dialogues in the animations are the essential elements that promote learners' interest, engagement, and enjoyment during the learning process (Liu & Elms, 2019).

Contrary to our expectations, students' score in the open-ended adult based items was significantly higher than the child and simulation-based items. However, there was no significant difference between the child and simulation-based items. This result may be due to the students' prior knowledge of the subject measured in the test item. As a limitation, the question distribution regarding the animation type is not equal, and the students' prior learning was not measured. Mayer (2009) claims that design's impact is more significant on students with low prior knowledge than students with high prior knowledge.

However, students' scores on child and adult-based items in the multiple-choice test were significantly higher than simulation-based items. However, child and adult items did not differ significantly. This result may be derived from the item type, a multiple-choice test, causing students to remember the correct answers from options. Thus, the difference between test scores caused by multiple-choice test items may be decreased. Besides, finding and remembering the answer from the options may have limited the effect of the design.

At the beginning of the study, we thought that using animations would eliminate the verbal ability required in traditional Science tests and test scores differ by gender and in favor of students with low-verbal skills requiring reading comprehension and prior knowledge about the subject. Indeed, there was also evidence that reading comprehension differs by gender in the literature (Arellano, 2013; Bayat, Şekercioğlu, & Bakır 2014; Çiftçi & Temizyürek, 2008; Kutlu, Yıldırım, Bilican, & Kumandaş, 2011). However, contrary to our expectations, research results show that students' test scores did not differ according to gender, both open-ended and multiple-choice tests. In other words, gender was not an effective variable on test scores, and boys' and girls' science scores change did not differ in terms of using animations on assessment. Dancy (2000) has found that although women performed significantly below men on the animated and traditional tests, the gap was consistently smaller on the animated test. In other words, the difference between the performance of men and women on the traditional test is greater than the difference in performance on the animated test (Dancy, 2000). Additionally, the correlation between gender and performance on questions closely tied to the animation dropped to zero in the animated condition (Dancy, 2000), supporting the present study result. Contrary to our finding, Bardach, et al. (2021) found that gender was not a significant variable on video-based tests (with and without text), but gender significantly predicted test performance in the text-based condition.

Research results show that using animation on assessment explains 36% of students' school Science grades. However, open-ended test items were not a significant predictor of students' success in animation-based assessment, while multiple-choice test items were significant predictors. This result may be derived from a general measurement tendency on measurement base on multiple-choice tests in schools. Some researchers stress that students mostly prefer multiple-choice tests (Sarıgül, 2009). While the measurement tools used in education differ according to the cognitive, affective, and psychomotor field, multiple-choice tests are often preferred in determining students' cognitive field behaviors in every course, including general exams (Karamustafaoğlu & Tutar, 2016). Birgili (2014) also emphasizes that there is a high propensity to assess academic success through the use of multiple-choice tests. Primary school teachers mostly use multiple-choice tests (Anıl & Acar, 2008). Providing scoring reliability in crowded groups is easier with multiple-choice tests (Klufa, 2015). However, there is a possibility of finding the correct answer in multiple-choice tests by guessing and eliminating method.

In multiple-choice tests, students are likely to find the correct answer by chance (Gültekin, 2014). It is stated that students think that multiple-choice tests are easy and do not make deep efforts (Struyven, Dochy, & Janssens, 2005). At this point, the research results in favor of multiple-choice tests may be due to the possibility of reminders of options in multiple-choice tests, the chance factor, and the sample group's habit and readiness for multiple-choice tests. It is stated that the hint effect in multiple-choice tests may cause high scores (Schuwirth, Vleuten, & Donkers, 1996). Braun, Bennett, Frye, and Soloway (1990) state that students can reach the correct answer with the elimination method in multiple-choice items, but this is not the case with open-ended items.

Doğan (2012), as another finding supporting the research findings, concluded that the best predictor is the multiple-choice test in the study where the structured grid and concept maps compare the science grade's predictive level. Unlike the research findings, in the study where Kan and Kayapınar (2006) examined the level of predicting the final success of students by comparing multiple-choice questions with short answer questions in foreign language education, it was concluded that short answer tests predicted success better than a multiple-choice test, and the multiple-choice test was not a significant predictor of success.

Limitations

Research results were limited with test scores of 52 students; besides, the level of prior knowledge of the students about the subject measured in the tests is unknown. The question distribution regarding the animation type is not equal.

Conclusion and Recommendations

In this study, it was found that there was no significant effect of using animation in assessment in terms of test validity and reliability. Comprehensive studies are recommended to explain the effect of using animation in the reliability and validity of assessment in terms of different variables such as reading comprehension. Students’ performance on animation-based tests were higher than traditional tests for both test item type. Future research can investigate variables that affect students' participation in animation-based tests and their opinions on animation-based tests. Using animated cartoon characters affected students’ test scores. In further studies, test-taking motivation may be investigated related to this result. Besides, the finding regarding the animation type variable can be investigated in-depth in terms of the effect levels of the visual and auditory elements added to the animations by including an equal number of items on the test. Gender was no significant variable on animation-based test scores. Using animation on assessment via multiple-choice test items was a significant predictor of students’ school Science grades. It should be noted that findings are limited to the type of test items used in the current study. The effect of animation on different types of test items may be investigated through future research.

References

Ally, B. A. (2012). Using pictures and words to understand recognition memory deterioration in amnestic mild cognitive impairment and Alzheimer’s disease: a review. *Current Neurology and Neuroscience Reports*, *12*(6), 687–694. https://doi.org/10.1007/s11910-012-0310-7

Alsawalmeh, Y. M., & Feldt, L. S. (1994). A modification of Feldt’s test of the equality of two independent alpha coefficients. *Psychometrika*, *59*(1), 49–57. https://doi.org/10.1007/BF02294264

Anıl, D., & Acar, M. (2008). Elementary school theachers’ views on issues they experience through measurement and evaluation processes. *Van Yuzuncu Yil University Journal of Education*, *5*(2), 44-61. https://dergipark.org.tr/tr/pub/yyuefd/issue/13714/166032

Arellano, M. D. C. (2013). Gender differences in reading comprehension achievement in English as a foreign language in compulsory secondary education. *Tejuelo: Didáctica de la Lengua y la Literatura. Educación*, *17*, 67-84.

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, *94*(2), 416–427. https://doi.org/10.1037/0022-0663.94.2.416

Atkinson, R., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent’s voice. *Contemporary Educational Psychology*, *30*(1), 117-139. https://doi.org/10.1016/j.cedpsych.2004.07.001

Barak, M., Ashkar, T., & Dori, Y. J. (2011). Learning science via animated movies: Its effect on students’ thinking and motivation. *Computers & Education*, *56*(3), 839–846. https://doi.org/10.1016/j.compedu.2010.10.025

Badger, E., & Thomas, B. (1992). Open-ended questions in reading. *Practical Assessment, Research & Evaluation*, 3(4), 03. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1039&context=pare

Bardach, L., Rushby, J.V., Kim, L.E., & Klassen, R.M. (2021). Using video- and text-based situational judgement tests for teacher selection: a quasiexperiment exploring the relations between test format, subgroup differences, and applicant reactions. *European Journal of Work and Organizational Psychology*, 30:2, 251-264, DOI: 10.1080/1359432X.2020.1736619

Bayat, N., Şekercioğlu, G., & Bakır, S. (2014). The relationship between reading comprehension and success in science. *Education and Science*, *39*(176). http://dx.doi.org/10.15390/EB.2014.3693

Birgili, B. (2014). *Open ended questions as an alternative to multiple choice: dilemma in Turkish examination system* [Master dissertation, Middle East Technical University]. OpenMETU. https://hdl.handle.net/11511/23866

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement*, *27*(2), 93-108. https://doi.org/10.1111/j.1745-3984.1990.tb00736.x

Burke, S. C., Snyder, S., & Rager, R. C. (2009). An assessment of faculty usage of YouTube as a teaching resource. *Internet Journal of Allied Health Sciences and Practice*, *7*(1), 8. https://nsuworks.nova.edu/ijahsp/vol7/iss1/8/

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, 25(2), 157–163. https://doi.org/10.1080/03098770120050828

Cachia, R., Ferrari, A., Ala-Mutka, K., & Punie, Y. (2010). *Creative learning and innovative teaching: Final report on the study on creativity and innovation in education in the EU member states*. JRC European Commission. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC62370/jrc62370.pdf

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

Chan, C. K. Y. (2013). Use of animation in engaging teachers and students in assessment in Hong Kong higher education, *Innovations in Education and Teaching International*, DOI: 10.1080/14703297.2013.847795

Chan, C. K. Y. (2015). Use of animation in engaging teachers and students in assessment in Hong Kong higher education. *Innovations in Education and Teaching International*, *52*(5), 474-484. https://doi.org/10.1080/14703297.2013.847795

Chao, T., Chen, J., Star, J. R., & Dede, C. (2016). Using digital resources for motivation and engagement in learning mathematics: reflections from teachers and students. *Digital Experiences in Mathematics Education*, *2*(3), 253-277. https://doi.org/10.1007/s40751-016-0024-6

Çiftçi, Ö., & Temizyürek, F. (2008). Measurement understanding of reading skills in 5th classes of primary schools. *Mustafa Kemal University Journal of Social Sciences Institue*, *5*(9), 109-129.

Dalacosta, K., Kamariotaki-Paparrigopoulou, M., Palyvos, J. A., & Spyrellis, N. (2009). Multimedia application with animated cartoons for teaching Science in elementary education. *Computers & Education*, *52*(4), 741–748. https://doi.org/10.1016/j.compedu.2008.11.018

Dancy, M.H. (2000). *Investigating animations for assessment with an animated version of the Force Concept Inventory* [Doctoral dissertation, North Carolina State University, Raleigh]. https://www.learntechlib.org/p/128748/

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

Daşdemir, İ., & Doymuş, K. (2016). The effect of use animation in unit electricity on academic achivements of 6th students, retention of the knowledge learned. *Pegem Education and Instruction Journal*, *2*(3), 33-42.

Denning, D. (1992). Video in theory and practice: Issues for classroom use and teacher video evaluation. Innature Productions.

https://www.academia.edu/4666293/Video_in_Theory_and_Practice_Issues_for_Classroom_Use_and_Teacher_Video_Evaluation

Dewan, P. (2015). Words versus pictures: leveraging the research on visual communication. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *10*(1). https://doi.org/10.21083/partnership.v10i1.3137

Dindar, M., Kabakçı Yurdakul, I., & İnan Dönmez, F. (2015). Measuring cognitive load in test items: static graphics versus animated graphics. *Journal of Computer Assisted Learning*, *31*(2), 148-161. https://doi.org/10.1111/jcal.12086

Doğan, S. (2012). *A comparision of concept map and structural grid techniques with multiple choice tests* (Publication No. 319636) [Master dissertation, Mersin University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall of India. https://ebookppsunp.files.wordpress.com/2016/06/robert_l-ebel_david_a _frisbie_essentials_of_edbookfi-org.pdf

English, J. (2020). Animation in online school science assessment: the validation of assessment for learning and individual development program. In: Unsworth L. (eds) *Learning from animations in science education. innovations in science education and technology*, vol 25. Springer, Cham. https://doi.org/10.1007/978-3-030-56047-8_11

Ercan, O., Bilen, K., & Bulut, A. (2014). The effect of web-based instruction with educational animation content at sensory organs subject on students' academic achievement and attitudes. *Procedia - Social and Behavioral Sciences*, 116, 2430-2436. https://doi.org/10.1016/j.sbspro.2014.01.587

Ferrari, A., Cachia, R., & Punie, Y. (2009). *Innovation and creativity in education and training in the EU member states: Fostering creative learning and supporting innovative teaching*. JRC European Commission. http://ftp.jrc.es/EURdoc/JRC52374_TN.pdf

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics: and sex and drugs and rock 'n' roll* (4th ed.). Sage.

Güler, N. (2017), *Eğitimde ölçme ve değerlendirme* (10^{th} ed.), Pegem Akademi Publishing.

Gültekin, C. (2014). *Comparison of abilities on drawing, reading and interpreting of graphs of the secondary education students and university students in change of state, solutions and solubility subjects* [Doctoral dissertation, University of Balıkesir]. Dspace. https://hdl.handle.net/20.500.12462/2836

Güven Demir, E. (2018). *Effect of applications based on flipped classroom on academic achievement and planning skills of 4th grade primary school students* (Publication No.519317) [Doctoral dissertation, University of Ondokuz Mayıs]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Hair, J., Black, W. C., Babin, B. J. & Anderson, R. E. (2010), *Multivariate data analysis* (7th ed.). Pearson Education International.

Hwang, I., Tam, M., Lam, S. L., & Lam, P. (2012). Review of use of animation as a supplementary learning material of physiology content in four academic years. *Electronic Journal of e-Learning*, *10* (4), 368–377. http://www.ejel.org/volume10/issue4

Kan, A. & Kayapınar, U. (2006). A comparison of the items and specifications of two english tests in multiple choice and short answer format measuring the same behavior in foreign language teaching. *Education and Science*, 32(142). http://egitimvebilim.ted.org.tr/index.php/EB/article/view/844

Karakolidis, A., O'Leary, M., & Scully, D. (2021). Animated videos in assessment: comparing validity evidence from and test-takers' reactions to an animated and a text-based situational judgment test, *International Journal of Testing*, 21:2, 57-79, DOI: 10.1080/15305058.2021.1916505

Karakolidis, A., Scully, D., & O'Leary, M. (2021). Eight issues to consider when developing animated videos for the assessment of complex constructs, *Practical Assessment, Research, and Evaluation*, 26(15). https://doi.org/10.7275/f2s7-yz14

Karamustafaoğlu, S. & Tutar, M. (2016). 6. sınıf dünya'mız, ay ve yaşam kaynağımız güneş ünitesi'ne yönelik bir başarı testi geliştirme [Developing an achievement test for 6th-grade earth, moon, and solar unit as our source of life.]. In Ö. Demirel, S. Dinçer (Eds.), *Eğitim bilimlerinde yenilikler ve nitelik arayışı [Education science innovations and the pursuit of qualifications]* (2nd ed., pp.303–320), Pegem Akademi Publishing, 303–320. http://dx.doi.org/10.14527/9786053183563b2.020

Karataş, F. Ö., Köse, S., & Coştu, B. (2003). Öğrenci yanılgılarını ve anlama düzeylerini belirlemede kullanılan iki aşamalı testler. *Pamukkale University Journal of Education*, 13(13), 54–69.

Kayaoglu, M. N., Akbas, R. D., & Öztürk, Z. (2011). A small scale experimental study: Using animations to learn vocabulary. *TOJET: The Turkish Online Journal of Educational Technology*, *10* (2). https://files.eric.ed.gov/fulltext/EJ932222.pdf

King, J. A. (2015). *getting the picture: a cartoon-based assessment tool for complex trauma in children* [Doctoral dissertation, University of Pennsylvania].https://repository.upenn.edu/edissertations_sp2/68/

Klufa, J. (2015). Multiple choice question tests–advantages and disadvantages. In K. Psarris (Ed.), *Proceedings of the 2015 International Conference On Education and Modern Educational Technologies* (pp. 39-42). http://www.inase.org/library/2015/

Koumi, J. (2006). *Designing video and multimedia for open and flexible learning*. Routledge. https://doi.org/10.4324/9780203966280

Kutlu, Ö., Yıldırım, Ö., Bilican, S., & Kumandaş, H. (2011). İlköğretim 5. sınıf öğrencilerinin okuduğunu anlamada başarılı olup-olmama durumlarının kestirilmesinde etkili olan değişkenlerin incelenmesi [Investigation of the variables that are effective in predicting the success or failure of primary school 5th-grade students in reading comprehension]. *Journal of Measurement and Evaluation in Education and Psychology*, *2*(1), 132-139. https://dergipark.org.tr/tr/pub/epod/issue/5806/77235

Lester, J., Callaway, C.B., Stone, B., & Towns, S.G. (1997). Mixed initiative problem solving with animated pedagogical agents. *AAAI Technical Report*. https://www.aaai.org/Library/Symposia/Spring/1997/ss97-04-018.php

Liu, C., & Elms, P. (2019). Animating student engagement: The impacts of cartoon instructional videos on learning experience. *Research in Learning Technology*, *27*. https://doi.org/10.25304/rlt.v27.2124

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

Martins, I. P. & Veiga, L. (2001). Early science education: exploring familiar contexts to improve the understanding of some basic scientific concepts, *European Early Childhood Education Research Journal*, 9:2, 69-82. https://doi.org/10.1080/13502930185208771

Mayer, R. E. (2005). C*ognitive theory of multimedia learning: The cambridge handbook of multimedia learning (R. E. Mayer, Ed.).* Cambridge University Press. 31–48. https://doi.org/10.1017/CBO9780511816819.004

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction*, *29*, 171-173. https://doi.org/10.1016/j.learninstruc.2013.04.003

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177–213.https://doi.org/10.1207/S1532690XCI1902_02

Moussiades, L., Kazanidis, I., & Iliopoulou, A. (2019). A framework for the development of educational video: An empirical approach. *Innovations in Education and Teaching International*, *56*(2), 217-228. https://doi.org/10.1080/14703297.2017.1399809

Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press.

Öksüz, Y., & Demir, E. G. (2019). Comparison of open ended questions and multiple choice tests in terms of psychometric features and student performance. *Hacettepe University Journal of Education*, *34*(1), 259-282. https://doi.org/10.16986/HUJE.2018040550

Polat, M. (2020). Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. *Novitas-ROYAL (Research on Youth and Language),* 14(2), 76-96.

Pop, E. C., Tuzinski, K., & Fetzer, M. (2016). Actor or avatar? Considerations in selecting appropriate formats for assessment content, technology and testing. In F. Drasgow (Eds.), *Improving educational and psychological measurement*, Routledge Taylor& Francis Group New york and London.

Qiu, S., Xia, Y., Tian, F. et al. (2020). Using a cartoon questionnaire to improve consent process in children: a randomized controlled survey. *Pediatric Research*. https://doi.org/10.1038/s41390-020-01227-2

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Sarıgül, Z. (2009). *An investigation into the effectiveness of multiple choice tests, structural communication grid and concept maps technique on the students' success in the aspect of measurement process and students' views about these techniques* (Publication No. 241764) [Master dissertation, University of Abant İzzet Baysal]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, *49*(1), 1-39. https://doi.org/10.2190/EC.49.1.a

Schuwirth, L. W. T., van der Vleuten, C. P. M., & Donkers, H. H. L. M. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education*, *30*(1), 44-49. https://doi.org/10.1111/j.1365-2923.1996.tb00716.x

Schwartz, D. L., & Hartman, K. (2007). *It is not television anymore: Designing digital video for learning and assessment* (R. Goldman, R. Pea, B. Barron, S. J. Derry, Eds.). Video research in the learning sciences, 335-348. https://doi.org/10.4324/9780203877258

Stith B. J. (2004). Use of animation in teaching cell biology. *Cell Biology Education*, *3*(3), 181–188. https://doi.org/10.1187/cbe.03-10-0018

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: a review 1. *Assessment & Evaluation in Higher Education*, *30*(4), 325-341. https://doi.org/10.1080/02602930500099102

Strømme, T.A., & Mork, S.M. (2020). Students' conceptual sense-making of animations and static visualizations of protein synthesis: a sociocultural hypothesis explaining why animations may be beneficial for student learning. *Research in Science Education*. https://doi.org/10.1007/s11165-020-09920-2

Temizkan, M., & Sallabaş, M. E. (2011). Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması. *Dumlupınar University Journal of Social Sciences*, 30, 207–220.

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi

van der Meij, H., van der Meij, J., & Harmsen, R. (2015). Animated pedagogical agents effects on enhancing student motivation and learning in a science inquiry learning environment. *Educational Technology Research and Development*, *63*(3), 381-403. https://doi.org/10.1007/s11423-015-9378-5

van Griethuijsen, R.A.L.F., van Eijck, M.W., Haste, H., den Brok, P. J., Skinner, N. C., Mansour, N., Savran Gencer, A., Boujaoude, S. (2015). Global patterns in students' views of Science and ınterest in Science, *Research in Science Education*, *45*, 581–603. https://doi.org/10.1007/s11165-014-9438-6

Wouters, P., Paas, F., & van Merriënboer, J. J. (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

Wu, H. C., Yeh, T. K., & Chang, C. Y. (2010). The design of an animation-based test system in the area of Earth sciences. *British Journal of Educational Technology*, *41*(3), 53-57. https://doi.org/10.1111/j.1467-8535.2009.00977.x

Yezierski, E. J., & Birk, J. P. (2006). Misconceptions about the particulate nature of matter. Using animations to close the gender gap. *Journal of Chemical Education*, 83, 954-960. https://doi.org/10.1021/ed083p954