MASTER THESIS

THE INFLUENCE OF RETENTION AND COMPENSATORY EMBEDDED QUESTIONS ON THE EFFECTIVENESS OF VIDEO LECTURES

Eszter Szöllősi

FACULTY OF BEHAVIOURAL, MANAGEMENT AND SOCIAL SCIENCES

MASTER EDUCATIONAL SCIENCE AND TECHNOLOGY

EXAMINATION COMMITTEE

Dr. Hans van der Meij Dr. Alieke van Dljk

Date 19.09.2019.

UNIVERSITY OF TWENTE.



Summary

The growing popularity of MOOCs and flipped classrooms increases the need for, and importance of video lectures. A way to make such lectures more effective is to include Embedded Questions (EQs). Although the effectiveness of EQs in video lectures on the knowledge gain is proven, there is a lack of research about what type of EQs are the most effective for increasing learning. This study aims to find an answer to the question *What is the effect of retention and compensatory EQs on knowledge gain and video engagement?* Retention questions draw attention to information that is already mentioned in the lecture. Compensatory questions invite students to find an answer to issues for which the lecture offers incomplete information.

The research was done in the form of a quasi-experiment. Participants answered a total of four EQs, two retention and two compensatory. Based on the research, the effectiveness of retention and compensatory EQs was analysed.

In general, knowledge gain happened by watching the video since the post-test score was higher than the pre-test score. The participants of this research scored high on video engagement as well, which might be explained by the inclusion of EQs. Contrary to the expectations, there was no significant difference found between the retention and compensatory questions when it comes to knowledge gain and video engagement. A number of possible explanations for the findings are presented in this paper as well.

This research should be seen as a first exploration about the effect of different types of EQs in video lectures. It is therefore highly valuable from an innovative point of view. Additionally, it provides guidelines for analysing videos via the Component Display Theory of David Merrill, which could be useful for educational professionals and researchers.

Acknowledgement

From starting with a blanking cursor in an empty Word document to finishing this thesis, there have been interesting conversations, hundreds of pages of drafts printed, long brainstorm sessions, big achievements, sacrifices and tears.

I would like to thank my supervisor, Dr. Hans van der Meij for providing guidance, pushing me to do better and always taking the time to talk things through. I would also like to thank my second reader, Dr. Alieke van Dijk, because her feedback led to valuable improvements in my thesis. Thank you Kyra for providing not just professional but personal support.

Thank you for Remko and Peter from OAB Dekkers for providing not just the video but a lot of valuable and critical feedback and support on the way.

I am grateful to my Mom, Dad and sister, who are my biggest personal support system! You always gave me the means that I was missing to pursue my dreams!

Nóri and Levi, you have been checking in on me and being there for me all the time. Thank you!

Axel, you have seen me at my worst and was still there to pick me up and to help me. You made me believe that I can do it, even when I did not believe in myself. I cannot even begin to describe how thankful I am to you. Christien, merci pour tout les temps que t'as m'aidé et que t'as me donné des commaintaires impressionants.

Finally, I would like to dedicate this paper to my biggest fan, who believed in me every step of the way, put the smile on my face and still gives me an enormous amount of strength, even from up there. This is for you, VNDadó!

Table of Contents

Ac	know	ledgement3
1.	Intr	oduction6
2.	The	oretical Framework8
	2.1.	Embedded questions and the testing effect8
	2.2.	Embedded questions and learner engagement9
	2.3.	Embedded questions and learning in texts10
	2.4.	Development of embedded questions in the digital era13
	2.5.	Retention and compensatory embedded questions14
	2.6.	Curiosity, spreading activation and compensatory embedded questions15
	2.7.	Component Display Theory (CDT) and the video audit17
3.	Res	earch question and hypotheses19
4.	Met	thod20
4	4.1.	Research design20
	4.2.	Video audit20
4	4.3.	Participants22
	4.4.	Procedure
	4.5.	Instrumentation24
	4.6.	Normality26
	4.7.	Data analysis27
	4.8.	Reliability
5.	Res	ults31
	5.1.	Scores of the embedded questions
	5.2.	Overall knowledge gain32
	5.3.	Knowledge gain and retention embedded questions32
	5.4.	Knowledge gain and compensatory embedded questions

5.5.	Interaction between retention and compensatory questions	33
5.6.	Video engagement	34
6. Dis	cussion and conclusion	36
6.1.	Answering the research question	36
6.2.	Limitations	40
6.3.	Future research	41
6.4.	Implications	42
6.5.	Conclusion	43
Referer	nces	44
Append	lices	51

1. Introduction

With the recent and rapid development of technology, Massive Open Online Courses (MOOCs) and Flipped Classrooms are becoming more important teaching methods (Evans & Baker, 2016; Gilboy, Heinerichs & Pazzaglia, 2015). MOOCs provide students access to video lectures, tests and exams in order to make learning more accessible everywhere (Conache, Dima, & Mutu, 2016). In case of a flipped classroom approach, students are expected to study the required material before class. This is usually done with the help of instructional videos so that class activities can be used to deepen and extend the already learned material (Limniou, Schermbrucker, & Lyons, 2018).

The instructional core in both MOOCs and flipped classrooms is a video lecture. Generally, a video lecture consists of a record from a real-life event. The video lecture usually switches between lecturer and slides, or the two are shown in a side-by-side view. The availability of video lectures keeps growing (Storme, Vansieleghem, Devleminck, Masschelein & Simons, 2016). The reason for that is that the fact that video lectures have repeatedly been proven to make the learning process more effective when combined with the traditional, face-to-face teaching (Choi & Yang, 2011; Wu, Yang, Zhang, & Huang, 2014). Additionally, there is an increasing availability of portable devices that can provide access to these video lectures (e.g.: mobile phones, tablets, laptops) (Algoufi, 2016). This combination of the proven effectiveness of video lectures and the availability of the devices to access them, creates a need for understanding how the videos can be made even more effective.

Although videos in general are quite effective in maintaining the motivation and attention of learners (Brar & van der Meij, 2017), supplementary design measures are often needed to ensure sufficient processing of lecture content. According to the findings of Callender & McDaniel (2007) and Strouse, O'Doherty, & Troseth (2013), questions asked during a video lecture are a good mean for raising students' knowledge gain. Such questions are also called 'embedded questions'.

There is a paucity of research on what type of questions is more effective. Since the specific types of questions asked might have a significant effect on the knowledge gain, an answer to this issue is important for the creation of more effective video lectures.

The importance of the problem lies in the improvement of the widely used video lectures. Knowing which type of embedded question increases knowledge gain the most could add to the already available information about other aspects of embedded questions. The aspects of the frequency (Rickards & Di Vesta, 1974), response mode (Bing, 1982), and knowledge gain have already been researched. These aspects will be further elaborated on in the theoretical framework (Chapter 2). There is need for a study on the relationship between the type of embedded questions and learning from video lectures. The present study aims to do so. Therefore, the main aim of this paper is to see what types of embedded questions are the most effective in improving the learning from video lectures. In order to provide a detailed overview of the research, several concepts need to be mentioned and described. Below is an in-depth discussion of the literature focusing on how (embedded) questions affect the different aspects of learning.

The first three sub-chapters focus on the relationship between the embedded questions and learning. The following three sub-chapters focus more on the educational videos. The last sub-chapter explains the Component Display Theory and how it has been used in this research as the basis of the video audit.

2.1. Embedded questions and the testing effect

According to Tweissi (2016), embedded questions (EQs) are questions that are put in a video after a content segment of a video lecture. Embedded questions can be considered to be a special type of instruction that can elicit a testing effect. Testing effect refers to the finding that students who have the opportunity to practice on a test before the final test, do better on the final test, compared to students who restudy or those who only take the final test (Adesope, Trevisan, & Sundararajan, 2017). Embedded questions can be looked at as a form of formative assessment, or as practice questions for the final test. This means that they test the knowledge during the lesson and give feedback without the score of the assessment affecting the final score of the student (Black & William, 2009). It can therefore be stated that every embedded question serves as a practice question for the final test (Vojdanoska, Cranney, & Newell, 2010).

The testing effect demonstrated that assessment cannot only be used for evaluating the students but to promote learning as well. Giving a chance to students to practice for a test in the form of embedded questions therefore can have three main advantages. The first one is that, if different from the final test questions, embedded questions can make learners see a different aspect of the studied material, instead of repeating it (McDaniel, Anderson, Derbish, & Morisette, 2007). The second advantage is that, if learners see the correct answer to the embedded question, in case they give the wrong answer to the embedded question, they get feedback straight away. This prevents the learners from learning the wrong information (Agarwal, Karpicke, Kang, Roediger III, & McDermott, 2008). As embedded questions elicit the

testing effect, it can be stated that they indirectly provide a positive influence on the learning process as well. Lastly, McDaniel et al. (2007) have proven that testing effect occurs when the practice questions are different from the final test questions, therefore embedded questions could serve as different practice questions and thereby promote learning.

The general conclusion is that embedded questions are effective due to the enhanced testing effect. Additionally, the mentioned findings can be looked at as a guidance for making the embedded questions effective. Making the question different from the material, providing the correct answers and making them differ from the final test could all be used in the current research.

2.2. Embedded questions and learner engagement

Learner engagement can be a very crucial part of the learning process. As the main medium of the current research are video lectures, it is important to have an overview of the findings concerning learner engagement in video lectures and how embedded questions can possibly affect learner engagement.

In many of the researches, learner engagement refers to the active relationship of the students with the study material (Connell, 1990; Fiedler, 1975; Koenigs, Fiedler, & Decharms, 1977). As an extension of this definition, Marks (2000) defined student engagement as "the attention, interest, investment, and effort students expend in the work of learning" (Marks, 2000, p. 155). Learner engagement is an essential part of learning (Johnson & Delawsky, 2013), and is therefore a good means to make the learning process more effective.

It has been proven several times that using video lectures in education increase learner engagement compared to the traditional classroom setting (Gilardi, Holroyd, Newbury and Watten, 2015; Gilboy, Heinerichs and Pazzaglia, 2015). According to Conrad and Donaldson (2011) a higher level of learner engagement results in more critical thinking. Furthermore, according to Gill (2008), learners can reach an in-depth understanding of high-level concepts by having attention (or: engagement) and motivation. It can be stated that learner engagement is one of the measures that can give an indication for the effectiveness of video lectures, hence could be used in the current research.

2019

There are other aspects that could are important for learning and could affect the effectiveness of embedded questions. Videos are proven to have a positive effect on motivation (Núñez, 2017), therefore motivation could be further investigated in case of video lectures with embedded questions. Additionally, self-efficacy is considered an important aspect of learning, since it can have significant effects on both motivation and knowledge gain (Vancouver & Kendall, 2006). However, the aforementioned two aspects, namely motivation and self-efficacy, are disregarded in the current study, since investigating their potential relationship with embedded questions would require a separate research.

Despite the positive effects of videos on engagement and on critical thinking, it is important to mention that according to Guo, Kim and Rubin (2014), there is place for improvement in educational videos, since participants still do not engage fully with these videos. They found that in case of a 3-minute-long video, students engage for one minute on average (33%), while in case of a 6-minute-long video, for four minutes on average (66%).

Research about the relationship between learner engagement and embedded questions has been conducted as well. Several studies suggest that embedded questions could play a role in improving learner engagement (Guo, Kim, & Rubin, 2014; Cummins, Beresford and Rice, 2016; Kolås, Nordseth and Hoem, 2016). Both Guo et al. (2014) and Kolås et al. (2016) suggest that the main reason for that is that embedded questions act as a surprise element and therefore are breaking the watching rhythm of the students.

To conclude, learner engagement is important to be taken into account when measuring the effectiveness of embedded questions. Furthermore, if embedded questions are formulated in a way to increase student engagement, they can contribute significantly to making learning more successful. Additionally, since no research has been done measuring the possible difference in learner engagement caused by different types of embedded questions in video lectures, the current research will be a valuable addition to this part of the literature as well.

2.3. Embedded questions and learning in texts

It has been proven that students receiving texts with embedded questions perform better on subsequent tests than students not receiving embedded questions (Rothkopf, 1966; Rothkopf & Bisbicos 1967; Rickards and DiVesta, 1974; Felker & Dapra, 1975; Bing, 1982). In other

2019

words, embedded questions have a positive effect on learning. Besides the pure effectiveness of embedded questions, it has also been proven that it is more beneficial to ask the questions after the relevant paragraph than asking them beforehand (Rothkopf & Bisbicos, 1967; Frase, 1968).

Besides the research about effectiveness of embedded questions and the positive effect on learning from embedded questions after relevant paragraphs, more specific research has been conducted. These areas of research about embedded questions in text and how they affect learning, can roughly be grouped into three main aspects. The three aspects are the (1) Level of embedded questions, (2) Frequency, and (3) Response mode. Following is a detailed description of the research of how these three aspects affect learning.

Level of embedded questions

Research has been conducted about whether the different difficulty levels of embedded questions have different effectiveness in the learning process. The opinions and findings for 'level' seem to be inconsistent.

Anderson (1972) draws a connection between the commonly used Bloom levels and the verbatim and comprehension embedded questions. Verbatim questions are the ones that require the learner to remember specific parts of the mentioned information. This is in line with Bloom level 1 (Remembering). Comprehension questions require the learner to not just remember but understand the material, thereby in line with Bloom level two (Understand). The higher level questions in other studies can also cover Bloom level three (Apply), where the learners need to use their knowledge in a new situation (Krathwohl, 2002).

Several studies confirmed that higher level embedded questions facilitate learning more than lower level ones (Anderson, 1972; Felker & Dapra, 1975; Hamaker, 1986; Andre & Thieman, 1988). Some of the mentioned studies also state that the effect is circumstantial. Felker & Dapra (1975) suggest that higher level questions need to be followed by a complex problemsolving task in order to be the most effective. Andre & Thieman (1988) highlighted that higher level questions only increase knowledge gain when not combined with lower level questions. To the contrary, Bing (1982) claims that lower level embedded questions are more effective than higher level ones, regardless of the level of questions asked in the post-test.

To conclude, the findings about how the different levels of embedded questions influence learning are not unanimous, however there seems to be a majority of studies stating that higher level questions are more effective. Including both levels in the current research might provide additional insight.

Frequency of embedded questions

Research has shown that the effectiveness of embedded questions on different dimensions of learning is influenced by how often embedded questions appear in the text.

Rickards and DiVesta (1974) studied the effect of frequency in case of high level and low level embedded questions. One group received one question after every second paragraph, while the other group had two questions after every fourth paragraph. They found that increase in frequency had a beneficial effect for higher level questions and that there was no significant effect of frequency when it came to verbatim (lower level) questions. This study therefore was focusing on the knowledge gain effect of embedded questions.

In their research, Frase, Patrick, and Schumer (1970) either asked five embedded questions after each fifth paragraph or one embedded question after each paragraph. The main finding concerning the frequency was that when the embedded questions were asked more frequently, they had an influence on the motivational effects of learning. In other words, this study pointed out how frequency indirectly influences the learning process.

Despite the fact that the frequencies differ, all researches have proven that as long as the questions are asked with moderate frequency and on the proper level, they have a beneficial effect for learning. In case of the current research it is therefore important to ask the embedded questions accordingly.

Response mode of embedded questions

Embedded questions can be categorised based on the nature of the answer as well. Several studies have been conducted about whether the content or the type of the answer facilitates learning.

Rothkopf & Bisbicos (1967) distinguished between four different types of response modes (a common phrase, a technical phrase, a measurement or a place or name). They concluded that questions that required more specific answers (technical phrase, measurement) have a stronger effect on learning.

Frase (1968) researched the difference between multiple choice and short-answer embedded questions. It was assumed that short-answer questions resulted in higher retention, since students have to come up with the answer for the short answer questions themselves. Contrary to their initial assumption, the question mode did not have a significant effect, which was later reconfirmed by Bing (1982) as well.

It can therefore be concluded that the content of the correct answer appears to be more important for the learning process than the desired form of the answer.

The literature in the aspects of the level, frequency and response mode of embedded questions have provided basic guidelines to follow in this study, and thereby making the different types of embedded questions more effective.

2.4. Development of embedded questions in the digital era

The above discussed studies mostly use embedded questions in written texts. This seems to be the general tendency in the area. The recent study of Bridges, Stefaniak & Baaki (2018) (re)confirmed that even nowadays the main focus of embedded question (EQ) research is still written text on paper. However, as digital tools are a crucial element in everyday activities, they are used more and more in education as well. It is therefore important to know more about the current application of embedded questions in digital educative tools.

Digital educative tools do not only include videos but eBooks as well. Sorva & Sirkiä (2015) concentrated on the different types of embedded questions used in eBooks, by performing a literature review. They categorised three new, different kinds of EQs, although stating that the classification is not mutually exclusive, meaning that there can be overlap between the categories. They were (1) EQs that introduce content, (2) EQs that reinforce learning, and (3) EQs that highlight content. Since this categorization of EQs is new, it might suggest that embedded questions have different influence on learning when changing the medium where

2019

It has been proven in earlier researches that people learn differently from videos than from written texts (Michas & Berry, 2000; Felton, Keesee, Mattox, McCloskey, & Medley, 2001; Butcher, 2014). The multimedia principle states that people – regardless of their learning preference – learn more when they are exposed to several types of visual and verbal content as opposed to only reading a text (Butcher, 2014). This statement is supported by several researches. Michas & Berry (2000) have demonstrated that participants learn more and better from a video compared to a text or drawings. The study of Felton et al. (2001) concluded that video instruction is a valuable addition to the traditional classroom teaching, since it helps students understand the material more.

The perceived effectiveness of embedded questions in video lectures has been researched as well. Several studies (e.g.: Callender & McDaniel, 2007; Szpunar, Jing & Schacter, 2014) have proven that embedded questions in video lectures are beneficial for knowledge gain compared to video lectures without embedded questions.

It can be stated that research concentrating on the effect of EQs in written text can provide a direction, however it has to be kept in mind that EQs might act completely differently in video lectures. In addition, the availability of research in the topic of specific type of embedded questions in video lectures is lower than desired, therefore more research is needed in the area.

2.5. Retention and compensatory embedded questions

As mentioned, the specific types of embedded questions are not researched in video lectures. More concretely, no research has been found that concentrates on the embedded questions sorted based on their connection to the video content. As stated above by Sorva & Sirkiä (2015), new types of embedded questions appeared in eBooks. They stated that there was overlap between the categories. For this reason, and to avoid making the current research too complex, the three categories are reduced to two. More specifically, EQs that reinforce learning and EQs that highlight content are merged into the same category, as both require retention of the information.

Concerning the types of embedded questions, no categorization has been done in the medium of video lectures. For this reason, the categorization of Sorva & Sirkiä (2015) based on eBooks has been used as a foundation for the current research. The above mentioned two categories were strongly related to the content and appeared in a digital medium, which are important aspects of the current study as well.

The first category is EQs that highlight or reiterate content, in other words *retention embedded questions*. The second category is EQs that introduce new content, in other words *compensatory embedded questions*. These types of embedded questions could be introduced in video lectures as well.

As far as the available literature goes, retention and compensatory embedded questions have not been used in video lectures yet. As outlined later, these two types of embedded questions have the potential to make learning more effective individually and by interacting with each other. Since they are strongly related to the content and presentation of the lecture, an audit of the content of a chosen video lecture needs to be done to find the good place and type of embedded questions.

2.6. Curiosity, spreading activation and compensatory embedded questions

As mentioned, retention embedded questions make learners recall information that was already mentioned in the video. Since the answer is accessible, retention questions are expected to be answered significantly better compared to compensatory questions. But what makes the compensatory questions effective?

There are two main reasons for the potential effectiveness of the presence of compensatory embedded questions in educational videos.

The first reason is that not knowing the answer to the embedded question can make the viewers curious. Building on the statements of Guo et al. (2014) and Kolås et al. (2016), that embedded questions are effective because of the surprise element, it can be assumed that compensatory embedded questions have a higher surprise effect than retention embedded questions. The reason for that might be that the participants cannot find the literal answer for the question in the video and therefore their curiosity gets triggered. This might even make them more engaged to the video. Several researches have already proven that curiosity

can strengthen the learning effectiveness (e.g.: Nojavanasghari, Baltrusaitis, Hughes, & Morency, 2016; Kang et al., 2006)

Furthermore, Kang et al. state that "curiosity enhances learning from new information" (Kang et al., 2006, p 5.). This can be meaningful from two aspects. Firstly, new information is referred to when participants give the wrong answer to a question and then see the correct feedback. Secondly, new information can be presented by the compensatory questions and their answers.

The second reason for the potential effectiveness of the presence of compensatory embedded questions in educational videos is that the spreading activation effect of compensatory questions might make participants learn more from retention embedded questions. The spreading activation theory is a theory in cognitive psychology about how the human brain is capable of remembering concepts, and relationships between concepts by receiving information that is connected to the given concept (Quillian, 1962). The idea is that every concept in the brain is stored as a so-called 'node' and the connection between concepts is prioritised based on strength and importance. For instance, "vegetable" is a concept (node) and it is connected to the concepts "cucumber", "broccoli" and "cauliflower". Broccoli is connected to the concept "green" and cauliflower is connected to the concept "white". All our knowledge therefore forms a complex network. The association that helps the retrieval can be static or dynamic. Static means that the concepts and the relationship between them is already known, it only has to be recalled. For example, already knowing that broccoli is green. Dynamic means that by receiving the information, the brain creates the relationship 'on-the-go' (Crestani, 1997). An example for that is seeing the picture of a cucumber and therefore connecting the concept of "cucumber" to the concept of "green" as well.

Based on the spreading activation theory, it is assumed that compensatory embedded questions have the potential to provide information that enhances the knowledge that is also required by the retention questions. In other words, because of the extra information, the two types of questions might have an interaction with each other.

2.7.Component Display Theory (CDT) and the video audit

It was already mentioned that because of the earlier revealed strong relation between the embedded questions and the content and presentation of the lecture, an audit of a chosen video lecture needs to be done to find a good place and type for the embedded questions. This audit will be done using the Component Display Theory (CDT) theory of David Merrill (1983).

The CDT of David Merrill (1983) categorizes the elements of a lecture based on their content, their complexity and their form of presentation. Lectures analysed and changed based on the CDT are proven to be more effective in case of the traditional lecture setting. It can be stated that CDT is not only suitable for the topic but for the medium as well. Since Tweissi (2016) did not use the CDT for a complete video audit and in strong connection with the content, but he did use CDT for designing embedded questions.

Both the content (objectives) and the presentation format of an educational video are important. The first feature in CDT is the performance-content matrix (See Figure 1). This matrix characterizes ten main ways in which learning objectives and items can be sorted based on two dimensions, namely type of content and student performance.

The type of content ranges from the simple facts all the way to the principles which determine and explain the cause and effect relationship between several concepts. Facts are, for instance, dates and events, hence the simplest type of information. Concepts are abstract ideas of grouped items. Concepts often need further explanation (or embedded questions) in order to be clear for the learner. An example of a concept could be the possible consequences of drinking and driving. Procedures are processes that can be described by a list of steps, such as starting a car. Finally, principles elaborate on, or determine, cause-and-effect relationships between concepts. An example for a principle is how to respond when a traffic light turns from green to red.



Figure 1. Performance-content matrix.

Primary Presentation Forms (PPF) are not only an important element of the theory of Merrill but also are crucial in case of any learning activity. Merrill (1987) distinguished between four main types of PPFs, based on how specific the presented information is, and the direction of interaction between the instructor and the learner (See Figure 2). As the chosen research video only contains expository presentation forms, PPF analysis will be a crucial part of analysing the video and choosing the place of the embedded questions.



Figure 2. Primary Presentation Forms Matrix.

The CDT specifies that instruction is more effective when it contains all necessary presentation forms. Thus, a complete lesson would consist of an objective, followed by a combination of rules, examples, recall, practice and feedback appropriate to the subject matter and learning task.

With CDT it is possible to detect rounded off sections in a video. When such a presentation is found, an embedded question (with feedback) is called for. As mentioned earlier, in the present study we investigate two kinds of embedded questions for these places, namely retention and compensatory questions. The description of the video audit process can be found in the Method section (4.2).

3. Research question and hypotheses

Based on the problem and theoretical framework, the main aim of the study is to find whether the type of embedded questions (henceforth simply questions) affects knowledge gain and video engagement. Therefore, the research question is:

What is the effect of retention and compensatory embedded questions on knowledge gain and video engagement?

Sub-questions:

- 1. What is the difference between the scores of the retention and compensatory embedded questions?
- 2. What is the effect of retention questions on knowledge gain?
- 3. What is the effect of compensatory questions on knowledge gain?
- 4. Is there a difference between learning resulting from the presence of retention and compensatory questions?
- 5. Is there an interaction between the two types of questions?
- 6. Does video engagement interact with the type of questions?

The literature in combination with the research question lead to the following hypotheses:

Hypothesis 1: Retention questions are answered more correctly than compensatory questions.

Hypothesis 2: Compensatory questions will make participants learn more from retention questions as well because of the spreading activation effect and curiosity.

Hypothesis 3: Retention questions yield to less video engagement than compensatory questions.

4. Method

This chapter includes the description of the experiment, starting with the research design, followed by the process of the video audit, including examples. Next, the sample is described, the steps of the experiment are outlined and the creation of the instruments used in the experiment is detailed. Lastly, the normality measures are presented, then the explanation of how the incoming data was analysed in order to find an answer to the research question, followed by reporting the reliability measures.

4.1. Research design

The research made use of a quantitative research approach, including an experiment using a video lecture in which two types of embedded questions were studied. The design was a *one group pre-test post-tests design*, in which, by definition every participant gets the same treatment (Bell, 2010). In case of the current research it meant that every participant watched the same video with two retention and two compensatory embedded questions and completed a pre-test and a post-test. In order to see possible differences, the order of the type of questions was varied in four different ways, without changing the content of the video. An overview of the different versions is presented later in Figure 6. This quasi-experimental design does not include a control group, since the effectiveness of embedded questions has already been proven.

4.2.Video audit

This research makes use of a video lecture about the creation of closed-ended questions. CDT, which was elaborated in the theoretical framework, was applied to analyse and prepare the video. Here follows a brief demonstration of how CDT was applied for the purposes of this research. The video-analysis was performed in an Excel-sheet which is available upon request.

Process

The chosen video lecture is a video about how to create closed-ended questions. It contains several tips and examples for teachers about what has to be paid attention to when creating closed-ended questions. The 15-minute long video was analysed according to the CDT in order to determine the place and type of the embedded questions needed. As shown in Figure 3,

the video was first divided into four segments, based on the content. The length of the segments was 2 min 11 s, 7 min, 4 min 43 s, and 1 min 35 s, respectively. Afterwards, the segments were divided and categorized with the help of the content performance matrix and the primary presentation forms matrix. In other words, it was determined whether the so called 'video events' are complete or incomplete.



Figure 3. Overview of the video audit.

Example of an incomplete cycle

In Figure 4, an example of an incomplete cycle is shown, according to the analysis. It is incomplete because the definition for "Negative formulation" is not defined. Therefore, the *compensating* embedded question here is: *What is the definition of a negative formulation*?

Weite und evolution landen ligt niet in Afrika? Meite und evolutione landen ligt niet in Afrika? A. Buckhar Faco B. Frank-Grass B. Frank-Grass Mauritanië	Een tip om te voorkomen dat de taalvaardigheid van studenten gaat toetsen in plaats van de stof waar het om gaat, negatieve formuleringen kan je beter vertalen in eenvoudige positieve vraagstellingen. Een voorbeeld is: Welke van de volgende landen ligt niet in Afrika? Wanneer je vragen negatief formuleert, vraag je van studenten dat deze eerst een taaltrip toepast. Hiermee maak je de vraag moeilijker maar op verkeerde gronden.	Voorbeeld van negatieve formulering	Bernember Fact Cancept Procedure Principle Types of Cortent
Welle van de volgende landen liet niet in Afrika? Melle van de volgende landen liet niet in Afrika? A Burkina-Faste B Frans Gyana C Norkust D Mauritanië	Dit komt de validiteit niet ten goede, maar mocht je toch negatieve termen gebruiken, zorg er dan in ieder geval voor dat deze opvallen. Biyoorbeeld door te onderstrepen of dik te drukken.	Relatie tussen concepts	Remember Fact Concept Procedure Principle Types of Content

Figure 4. Video analysis example: Incomplete event.

Example of a complete cycle

Figure 5 shows an example of a segment after which an embedded question emphasizing the content could be inserted. This segment is not missing anything crucial, because it makes the learner remember a concept by giving the definition of what "One aspect per question" is. Then the learner sees how to use the concept through an example. As the last part of the

segment, a principle is presented, connecting two concepts from earlier, namely the example of the "one aspect per question" is connected to the concept of validity and reliability. An example of a compensatory embedded question is: *If a question contains negative formulation, how does that affect the reliability?*



Figure 5. Video analysis example: Complete event.

As a result of the audit, it was decided to be asking embedded question after every segment, resulting in 4 embedded questions during the whole video. One event from each segment was chosen to be the topic of the embedded question at the end of the corresponding segment. Both a retention and compensatory question for the chosen event were created. To conclude, the audit provided the basis for the creation and placement of the total of eight embedded questions in order to create the four different versions of the research (See Figure 6 and Procedure section (4.4) for explanation).

4.3.Participants

Despite the fact that participation in the research was voluntary, the sample is not completely random, because it only contains people from the target group who had the incentive to participate. The experiment was run for two weeks. Participants had to sign up by email to participate and then randomly got assigned – based on the moment of signup – to the Graasp research environment of one of the four video versions. Thereby all participants had an equal chance to be assigned to one of the four versions. Participants were expected to complete

both the pre-test and the post-test. Only if they did so, their results were registered, since otherwise the knowledge gain could not have been accurately checked.

The total number of participants was expected to be at least 30, since the research was a one group design (Bell, 2010). The actual number of respondents was 32. The sample had a mean age of 24.38 (*N*=32, *SD*=2.89), varying between 20 years and 30 years old. 3.1 % of the sample had a vocational degree, 37.5 % had a university of applied sciences degree and 59.4 % had a university degree.

Login codes were sent to participants in an even distribution, however after receiving the login code, it was still the voluntary decision of the participant whether he or she completes the experiment. Therefore, the distribution is not perfectly even. Given the fact that the focus of the study is on different types of embedded questions, Table 1 shows the distribution of embedded questions per type.

Table 1. Answered embedded questions per type

	Embedded	Embedded	Embedded	Embedded
	question 1	question 2	question 3	question 4
Retention	N = 17	N = 19	N = 15	N = 13
Compensatory	N = 15	N = 13	N = 17	N = 19

4.4.Procedure

After granted permission for conducting the research from the Ethical Committee of the University of Twente, a short pilot test was done to determine the approximate duration of the experiment and to ensure content validity. The participants were acquired through the researcher's personal network and social media platforms, such as Facebook and LinkedIn. The potential participants had to send an email to the researcher in order to get the link to the research environment along with their personal login name. The experiment was online and was completed from the personal computers of the participants. The main advantage of such an online experiment was that participants had the chance of completing the experiment at home, which imitates the circumstances in which participants would watch the video if it would be part of a MOOC or flipped classroom method.

After clicking on the link and signing in with their personal login names, participants had to give informed consent. (All the respondents gave consent.) The experiment started with a pre-test. Then the participants had to watch four video segments, each of which were followed by an embedded question, which they had to answer.

As mentioned earlier, in order to see the possible difference in the effects of retention and compensatory questions, the order of the kinds of questions was different. The content of the video was not changed. Figure 6 shows all the different versions and the overview of the procedure.

After answering each embedded question, the participants had the chance to see the correct answer to the question. After answering the fourth embedded question and checking the correct answer, participants filled in the post-test. Lastly, they had to answer four demographic questions. The expected total completion time was approximately 45 minutes.

	Pre-test	Embedded question 1	Embedded question 2	Embedded question 3	Embedded question 4	Post-test	Demographic questions
Sample group 1	0	Compensatory	Compensatory	Retention	Retention	0	0
Sample group 2	0	Compensatory	Retention	Retention	Compensatory	0	0
Sample group 3	0	Retention	Retention	Compensatory	Compensatory	0	0
Sample group 4	0	Retention	Compensatory	Compensatory	Retention	0	0

Figure 6. Overview of the procedure and the different video versions.

4.5.Instrumentation

Pre- and post-test

A pre-, and post-test were created in order to measure the knowledge gain of the participants. Both tests consisted of a mix of open ended and closed-ended questions related to the content of the video. The video was divided into main topics and both the pre-test and the post-test were covering the same topics. All the pre-test and post-test questions were also categorized based on whether they are connected to a retention or compensatory embedded question, hence named "Retention" or "Compensatory". In case the test question was connected to the content of the video, but not directly connected to any of the embedded questions, it was categorized as "Other" (See categories and topics in Figure 7).

Pre-test		Post-test	
Othor	Question 1	Othor	
Other	Parts of a question	Other	
Othor	Question 2	Othor	
Other	Stem	Other	
Potontion	Question 3	Potention	
Retention	Alternatives	Retention	
	Question 4		
Other	Quality requirements of a closed-ended	Other	
	question		
Compensatory	Question 5	Compensatory	
compensatory	Validity	compensatory	
Potontion	Question 6	Potention	
Retention	Reliability	Retention	
Componentary	Question 7	Componentary	
compensatory	Usability	compensatory	
Componentary	Question 8	Componentary	
compensatory	Transparency	compensatory	

Figure 7. Categories and topics of the pre-test and post-test questions.

See pre-test including the correct answers and the score division in Appendix D and post-test including the correct answers and the score division in Appendix E.

The main reason for adding the "Other" category questions to both the pre-test and the posttest was to make them more life-like. In other words, they do not intend to measure the effect of embedded questions but the general knowledge gain caused by the video. Concerning the embedded question-related elements of the tests, it is important to highlight that the difference between the amount of retention and compensatory elements was not intentional. It has only been noticed after the completion of the experiment and therefore could not be corrected.

Video

The chosen video is a Dutch speaking, approximately 15 minute long lecture about tips and examples of creating closed-ended questions (See video transcript in Appendix A). The main reason for choosing this video was that it seemed to represent the quality of an average, educational video that is available to be used by educators.

Embedded questions in the video

The two types of embedded questions were inserted based on the video audit since the embedded questions are in strong connection with the content of the video. The video audit was conducted using the Component Display Theory of David Merrill (1983). The video was divided into four segments and at the end of each segment a retention or compensatory question was asked.

An example of an retention question is: '*If a question contains negative formulation, how does that affect the reliability?*', given the fact that both reliability and the definition of negative formulation are specified in the video (See embedded questions including correct answers and the score division in Appendix C).

An example of a compensatory embedded question is '*What is the definition of transparency* of a closed ended question?', given the fact that the definition of transparency is not included in the video.

Usage logs

A built-in video analysis tool was used in Graasp to measure the differences in video engagement. There were three relevant variables in case of video engagement, namely unique play time, replay time and play time. Unique play time represents the amount of time (in percentages) during which the participants played the video once. The maximum unique play time can be the full length of the video, meaning if the video is 132 seconds, the unique play time will not exceed that time (100 %). Replay time is the amount of time spent by rewatching the video. This represents how much of the video has been played more than once. Lastly, play time stands for the total time of the played video. These measures are in line with the video engagement analysis in several other researches (e.g.: Guo, Kim, & Rubin, 2014; Hyunwoo, Schulzrinne, & Kim, 2016), hence are comparable.

4.6.Normality

The different variables first were checked to see whether they are normally distributed. For the reason of the low sample size (N < 200), the Shapiro-Wilk test results were used and interpreted. In case the investigated variable was not normally distributed, the nonparametric alternative of the t-tests was used, namely the Mann-Whitney test as an

alternative to the independent samples t-test and the Wilcoxon signed rank test as an alternative for the paired samples t-test.

The normality measures and their significance are presented in Table 2. In case of the intention of performing and independent samples t-test, the normality of the two different variables was calculated, while for a paired samples t-test, the normality of the difference between the two variables was calculated (Samuels & Marshall, 2019).

Variable	Statistic	df		Sig.
Retention EQ Score	.87		32	.001
Compensatory EQ Score	.86		32	.001
Difference between total post-test score and total pre-test score	.96		32	.250
Difference between retention score of pre-test and retention EQ				
score	.96		32	.267
Difference between retention EQ score and retention score of				
post-test	.91		32	.011
Difference between compensatory score of pre-test and				
compensatory EQ score	.86		32	.001
Difference between compensatory EQ score and compensatory				
score of post-test	.91		32	.010
Play time retention EQs	.89		64	.000
Play time compensatory EQs	.89		64	.000
Unique play time retention EQs	.63		64	.000
Unique play time compensatory EQs	.62		64	.000
Replay time retention EQs	.42		64	.000
Replay time compensatory EQs	.49		64	.000

Table 2. Normality of variables used in the data analysis.

4.7. Data analysis

It is expected that retention questions yield to less video engagement and contribute less to knowledge gain than compensatory questions, since they only make the students recall the content. As stated earlier, retention questions reiterate content that was already covered by the video, while compensatory questions introduce content that was not completely covered in the video. In order to answer the research question and see whether the expectation was true or not, the below described analyses were performed.

Knowledge gain

The pre- and post-test results of the participants were compared in order to see the overall knowledge gain. As the pre-test and the post-test measured the same type and amount of knowledge, the difference in total score indicated the knowledge gain. The significance was tested by a non-parametric paired samples t-test (Wilcoxon signed rank test). The score of the answers given to the retention and the compensatory questions were compared as well. It was therefore visible whether in the post-test the compensatory items were answered more correctly or not.

For this research, the main focus was to test the knowledge gain caused by the embedded questions only. Between the pre-test and the embedded questions, participants could have learned from the video itself. Therefore, taking the difference between the pre-test scores and the embedded question scores into account showed the knowledge gain caused mainly by the video. Between answering the embedded question and post-test question, there was little to no information provided by the video about the given EQ and post-test question. By comparing the scores of the two types of embedded questions to the scores of the score the given embedded question could be seen.

Therefore, the possible knowledge gain caused by the embedded questions was intended to be tested in two steps as follows: based on the categorisation presented in Figure 7, the connections between pre-test items, embedded questions and post-test items were identified. As the first step, with the help of (non-parametric) paired samples t-tests, the differences between the answers of the embedded questions and the answers of the corresponding pre-test questions were shown for both EQ types. As the second step, the difference between the embedded question score and the post-test score was discovered, by using non-parametric paired samples t-tests (Wilcoxon signed rank test). The calculations were made for both retention and compensatory EQ types.

Interaction between the two types of embedded questions

The possible interaction between the two types of questions was tested by two Spearman's rank-order tests.

Firstly, the difference in score between the retention elements of the pre-test and the retention embedded questions was calculated. The same calculation was made for the compensatory elements. Afterwards, a Spearman's rank-order correlation was run to determine the relationship between the change in scores.

Secondly, the difference in score between the retention embedded questions and the retention elements of the post-test was calculated. The same calculation was made for the compensatory elements. Afterwards, a Spearman's rank-order correlation was run to determine the relationship between the change in scores.

Video engagement

As the four videos were not the same length, the scores were transferred into percentages, in order to see the unique play time, replay time and replay time relative to the length of the videos. By combining all the video engagement data concerning retention and compensatory questions, Mann-Whitney tests were conducted to see whether there was a significant difference in how much of the video participants with retention embedded question and participants with compensatory embedded questions watched, replayed and played. It was expected that segments with compensatory questions will be replayed and played more. The reason for this expectation is that participants do not have the necessary information from the video to answer the compensatory embedded questions and will therefore spend more time looking for the missing information in the video.

4.8. Reliability

The reliability of the pre-test and post-test were tested and the data is reported in the sections below.

Pre-test

The overall reliability of the pre-test was a Cronbach's α = .57, whereas the reliability of the items related to a retention embedded question is Cronbach's α = .04. The reliability of the items related to a compensatory embedded question is Cronbach's α = -.10. The items related to the video had a reliability of Cronbach's α = .37.

Post-test

The overall reliability of the post-test was a Cronbach's α = .70, whereas the reliability of the items related to a retention embedded question is Cronbach's α = .52. The reliability of the items related to a compensatory embedded question is Cronbach's α = -.37. The items related to the video had a reliability of Cronbach's α = .64.

Test-retest reliability

The test-retest reliability was moderate (Pearson Correlation = .52; p = .002).

The overall reliability measures were relatively low. One of the reasons for that, also proven by Button, et al. (2013), can be the low sample size (N=32). Additionally, according to Ryff & Keyes (1995), despite Cronbach's alpha is the most widely used measure of test reliability, it is a conservative estimate. Another reason for the low reliability can be the small number of items per scale, meaning that there are only 8 pre-test questions and 8 post-test questions. Since these are divided into sub-groups, reliability tests of the current research consist only of 2-4 scales. Furthermore, the items of this research were created with the main purpose of covering all the different subtopics of the video and the additional topics of the compensatory embedded questions (see the rationale for item creation in the Instrumentation section), rather than to maximize internal consistency.

Inter-rater reliability

Reliability of the scores was intended to be increased by inter-rater agreement (Graham, Milanowski, & Miller, 2012). This means that a fellow researcher also scored all the answers pre-test, post-test and embedded questions to assure that scores are the same in both cases.

Using all the corrected results, an inter-rater reliability (Cohen's Kappa) was calculated. As represented in Table 3, the inter-rater reliability was high in case of the pre-test, embedded questions, and post-test scores as well. This means that the scoring criteria are transparent and objective.

Table 3. Inter-rater reliability scores.

	Ν	Карра	Sig.
Pre-test	32	0.81	. 000
Embedded Questions	32	0.89	. 000
Post-test	32	0.97	. 000

5. Results

This chapter is built up based on the sub-research questions, in order to be able to answer the main research question and hypotheses later. First the possible difference in scores of the two types of embedded questions are reported, followed by testing the overall knowledge gain. Afterwards, the tests performed to see the knowledge gain caused by both the retention and compensatory embedded questions are outlined to see the possible differences between the two types of questions. Lastly, the results of the tests are shown to see whether the two types of embedded questions interact with each other, followed by the tests concerning the video engagement.

5.1. Scores of the embedded questions

In order to see whether retention embedded questions are answered more correctly than compensatory embedded questions, the Mann-Whitney test was conducted (See Table 4). The results were significant, with $W_s = 4,954.50$, p < .001. This showed that the retention embedded question scores (M = 1.211; SD = .86) were significantly higher than compensatory embedded question scores (M = .59; SD = .64).

Table 4. Overview of EQ Scores per Type.

	Ν	Mean	SD	
Retention EQ Score	64	1.211	.86	
Compensatory EQ Score	64	.59	.64	

5.2. Overall knowledge gain

To determine whether the video with embedded questions led to knowledge gain, a paired samples t-test was conducted (See Table 5). The results were significant with t (31) = 10.30, p < .001. This showed that on average, the total post-test scores (M = 8.64; SD = 3.39) were higher than the total pre-test scores (M = 3.34; SD = 1.94). This means that participants answered more questions correctly after watching the video segments with embedded questions.

Table 5. Total Pre-test Scores and Total Post-test Scores.

	Ν	Mean	SD
Total Pre-test Score	32	3.34	1.94
Total Post-test Score	32	8.64	3.39

As mentioned in the data analysis section (4.7), the overall knowledge gain per question type was analysed as well. As the post-test and the two types of embedded questions had a different maximum score, the scores were presented in percentages.

5.3. Knowledge gain and retention embedded questions

To see the knowledge gain concerning retention type of questions, first a paired samples ttest between the total score of the retention elements of the pre-test and the total score of the retention embedded questions was conducted. The results were significant, with t (31) = 6.73, p < .001. This showed that on average, the total retention embedded question score of participants (M = 60.55; SD = 36.63) was significantly higher than the total score they got for answering the retention elements of the pre-test (M = 20.31; SD = 15.69).

Secondly, a Wilcoxon signed rank test was conducted to see the difference between the total retention embedded question score and the total score participants got for answering the retention elements of the post-test. The results were not significant with T = 193.50, p = .087. This showed that the total score of the retention elements of the post-test (M = 75.00; SD = 29.78) was not significantly higher than the total retention embedded question scores of participants (M = 60.55; SD = 36.63).

5.4. Knowledge gain and compensatory embedded questions

The knowledge gain concerning compensatory type of questions, first a Wilcoxon signed rank test between the total score of the compensatory elements of the pre-test and the total score of the compensatory embedded questions was conducted. The results were significant, with T = 118.00, p = .007. This showed that the total compensatory embedded question score of participants (M = 29.69; SD = 24.13) was significantly higher than the total score they got for answering the compensatory elements of the pre-test (M = 17.19; SD = 19.51).

As the next step, another Wilcoxon signed rank test was conducted to see the difference between the total compensatory embedded question score and the total score participants got for answering the compensatory elements of the post-test. The results were not significant with T = 137.50, p = .130. This showed that the total compensatory embedded question scores of participants (M = 29.69; SD = 24.13) was not significantly higher than the total score of the compensatory elements of the post-test (M = 21.88; SD = 15.88).

5.5. Interaction between retention and compensatory questions

The interaction between the two types of questions was tested by two Spearman's rank-order tests.

The first Spearman's rank-order correlation showed that there was a moderate, positive correlation between the change in scores of retention and compensatory elements, which was statistically insignificant ($r_s = .339$, p = .058).

The second Spearman's rank-order correlation showed that there was a very weak, positive correlation between the change in scores after the compensatory embedded questions and the change in scores after the retention embedded questions, which was statistically insignificant ($r_s = .106$, p = .564).

It can therefore be stated that the interaction between the two types of embedded questions was not proven.

5.6.Video engagement

All the video engagement data was converted into percentages because the four videos were not the same length. This way, the relative video engagement measures were analysed. The mean video engagement was calculated for all the variables, namely, play time, unique play time and replay time. The detailed description of what these variables stand for is included in the Data analysis section (4.5).

As shown in Table 6, on average, participants played 97% of the videos and replayed 9.52%. It can be stated that most participants did not watch the video in full length (M = 81.65). The analysis was conducted using the data of all the retention questions (N = 64) and all the compensatory embedded questions (N = 64).

Table 6. Total Video Engagement.

	Ν	Mean	SD
Play Time	128	97.00	50.06
Unique Play Time	128	81.65	32.41
Replay Time	128	9.52	24.47

The results of the Mann-Whitney tests for play time, unique play time and replay time are presented in Table 7 and interpreted below.

Engagement measure	EQ Type	Ν	Mean (%)	SD	Ws	Sig.	
Dlay Timo	Retention	64	96.36	49.30	4 047 50	.701	
Play Time	Compensatory	64	97.65	51.19	4,047.50		
Unique Play Time	Retention	64	81.63	31.88		.852	
Unique Play Time	Compensatory	64	81.68	33.18	4,090.50		
Boplay Timo	Retention	64	9.54	26.37	1 1 1 2 00	022	
	Compensatory	64	9.50	22.62	4,115.00	.333	

Table 7. Video engagement differences.

In case of the Mann-Whitney test performed for the play time, the results were not significant, with $W_s = 4047.50$, p = .701. This showed that the percentage of the video that participants with compensatory embedded questions played was not higher than (M = 97.65; SD = 51.19) the percentage of the video that participants with retention embedded questions played (M = 96.36; SD = 49.30).

In case of the Mann-Whitney test performed for the unique play time, the results were not significant, with W_s = 4090.50, p = .852. This showed that on average, the percentage of the video that participants with compensatory embedded questions played once, was not higher than (M = 81.68; SD = 33.18) the percentage of the video that participants with retention embedded questions played once (M = 81.63; SD = 31.88).

In case of the Mann-Whitney test performed for the replay time, the results were not significant, with W_s = 4113.00, p = .933. This showed that on average, the percentage of the video that participants with compensatory embedded questions replayed was not higher than (M = 9.50; SD = 22.62) the percentage of the video that participants with retention embedded questions replayed (M = 9.54; SD = 26.37).
6. Discussion and conclusion

The aim of this research was to see the potential differences between two types of embedded questions concerning knowledge gain and video engagement. In this chapter, the research question and sub-questions get answered, including the three hypotheses. It is followed by a list of limitations. Suggestions of future research are included as well, then a brief explanation of how this research can be utilised in science and practice, ending with a conclusion of the study.

6.1. Answering the research question

Scores of the embedded questions

Proving the first hypothesis and answering the first sub-question, participants scored higher on retention embedded questions than compensatory embedded questions. This finding also proves that both retention and compensatory embedded questions worked as intended. Since retention and compensatory questions are types of questions and not levels, they cannot be compared directly to Bloom's levels. However, it can be stated that compensatory questions are harder than retention questions, since for the latter the answers are already in the video while for compensatory questions they are not. If retention questions are looked at as lower level questions, this finding is in line with the statement of Bing (1982), namely that lower level embedded questions are more effective than higher level embedded questions.

Knowledge gain

Post test scores were higher than pre-test scores, meaning that learning occurred due to the video with the embedded questions. This finding provides an additional proof for the positive effect of embedded questions on knowledge gain, which has already been proven in many studies (Rothkopf, 1966; Rothkopf & Bisbicos 1967; Rickards & DiVesta, 1974; Felker & Dapra, 1975; Bing, 1982). As stated by Bridges, Stefaniak and Baaki (2018), most of these studies still investigate embedded questions in a text, while the current study provides evidence for the effect of such questions on knowledge gain in the context of video lectures.

Contrary to the expectations, that were also included in the second hypothesis, there was no significant difference between the knowledge gain caused by the two types of embedded questions. This study distinguished embedded questions based on types and disregarded the possible effect of the content of the correct answer to the embedded questions. However,

the content of the answer might have had an effect on the knowledge gain caused by the questions. Rothkopf & Bisbicos (1967) stated that the more specific phrases a correct answer contains, the more effective the embedded question in promoting learning. The lack of significant knowledge gain in the current research could therefore have been affected by the too generic correct answers required to the embedded questions.

As mentioned, based on the level of difficulty, retention questions could be looked at as lower level questions and therefore, compensatory as higher level. Andre & Thieman (1988) highlighted that higher level questions only increase knowledge gain when not combined with lower level questions. This might mean that the effectiveness of compensatory questions was lowered by the inclusion of the retention questions. The fact that the two types of questions did not interact with each other, could also be explained by the unsuccessful combination of question types suggested by Andre & Thieman (1988). It might be that compensatory questions alone would have caused higher knowledge gain.

Felker & Dalpra (1975) stated that higher level embedded questions are the most effective if followed by a complex problem solving task. This was not the case in the current research, which might have lowered the amount of knowledge gain resulting from compensatory embedded questions.

Video engagement

As of video engagement, the fact that most of the participants did not watch the whole video is a realistic indication for the engagement during an educational video, which can be part of a MOOC or flipped classroom assignment. The engagement in the current study was higher than the 66% found by other researches, such as Guo et al. (2014). This is interesting, since participants did not get any reward for completing the experiment. The explanation for the relatively high engagement can be the intrinsic motivation to help the researcher, meaning that the participants had the intention to do well on the experiment and thereby making the research successful. The high engagement can also be explained by the embedded questions themselves. In the study of Guo et al. (2014), the educational videos did not contain embedded questions. This means that the inclusion of the type of embedded questions. This is in line with the findings of the studies proving the positive effect of embedded questions on learner engagement (Gilardi, Holroyd, Newbury and Watten, 2015; Gilboy, Heinerichs and Pazzaglia, 2015). The aforementioned researches were investigating embedded questions in physical texts, while the current research specifies the potential benefits of embedded questions on learner engagement in video lectures.

According to Conrad and Donaldson (2011), such high learner engagement (as in the current research) should result in higher level critical thinking. That should mean that compensatory questions are more effective, since they are assumed to be of higher difficulty and hence might require more critical thinking, that could show in an increased knowledge gain caused by them. As shown in the results, compensatory embedded questions did not result in significantly more knowledge gain, when compared to retention embedded questions. A possible explanation for why the finding of Conrad and Donaldson (2011) was not reconfirmed by this research is that compensatory embedded questions were not complex enough to enhance higher order thinking.

The last hypothesis was not confirmed, meaning that retention questions did not yield to significantly less video engagement than compensatory questions. This also answers the last sub-question. Video engagement did not have an interaction separately with the types of embedded questions. However, as explained above, the general positive effect of embedded questions on video engagement has been proven. The spreading activation theory of Quillian (1962) might have shown its effect between the retention and compensatory questions. In other words, compensatory questions might have had an effect on retention questions as well, thereby making the knowledge gain and video engagement similar between the two types of questions.

Possible explanations of the results

As indicated in the sections earlier, the compensatory embedded questions did not act according to the expectations neither in case of knowledge gain, nor in video engagement. Since neither types of embedded questions have been tested before, the assumptions regarding how they will act in an educational video were purely based on theory.

As stated, the literal answers to compensatory embedded questions cannot be found in the video, therefore one might ask, why they were even expected to be beneficial for the learning

process. The argument of curiosity and spreading activation have provided a logical explanation for the usefulness of such type of embedded questions (See 2.6 in the Theoretical Framework) (Quillian, 1962; Kang, McDermott, & Roediger III, 2007; Guo et al., 2014; Kolås et al., 2016).

A possible explanation for the results might have a motivational root. As mentioned earlier, the current research disregarded any other factors than knowledge gain and video engagement concerning learning. It might have been the case that the different questions had a different effect on aspects such as motivation. As mentioned earlier, videos have a positive influence on motivation (Núñez, 2017). However, motivation was not included in this paper, since adding another aspect would have made the study too complex and as such a new research in the field, simplicity was of high importance.

Additionally, the difference between the retention and compensatory questions in difficulty might have been too big. In other words, it might be that the compensatory embedded questions were seen as too complex or too difficult. According to Tweissi (2016), difficult questions can negatively affect the self – efficacy of participants, even if only followed by relatively easy ones. Furthermore, Vancouver & Kendall (2006) stated that negative self-efficacy can lower both the motivation and knowledge gain of participants. The combinations of the aforementioned two statements suggest that the high difference in the difficulty could indeed have resulted in lower general performance.

It might also be the case that the participants simply did not see the added value or function of the compensatory embedded questions, and hence did not take them as seriously as the retention ones.

To close, the fact that the effectiveness of the compensatory questions was not proven in the current research, should be noted. However, a research concentrating solely on the earlier mentioned factors, (e.g.: motivation) could provide different results. Furthermore, the potential activation spread effect of the compensatory questions was mentioned, however could be further tested. For the reason of the spreading activation theory of Quillian (1962), it might be that the retention embedded questions were answered better in this research than in a similar research without compensatory embedded questions.

The main aim of this study was to find out whether different types of embedded questions might have a different effect on knowledge gain and on learner engagement. Although the study included a thorough experiment, during the research period certain limitations came to light. It is expected that these points will help future researchers to avoid facing the same shortcomings. The following limitations are important to measure for future researchers: the duration of the experiment, the delivery mode of the experiment, the scope of the analysis and sampling bias.

First of all, the duration of the experiment might have influenced the motivation of the participants and – indirectly – the results at the end of the study. The respondents were volunteers and were doing the experiment in their free time. It is important to mention that the respondents did not get anything in return for participating. They did not participate in the research because they got rewarded. The respondents all came out of the (in)direct network of the researcher which implies that their motivation to help out was personal. Therefore, a 45-minute-long experiment might have been too long. However, it was important to simulate an online learning environment and there, 45 minutes is not an unusual length.

The second limitation, the delivery mode of the experiment, could be connected to the foregoing argument. The experiment was delivered to the participants via an online research environment. In other words, participants could fill in the experiment from anywhere they liked. Therefore, there was no direct control of the behaviour of participants and the circumstances (time, environment etc.) in which the experiment was completed. On one hand, this could harm the reliability of the measurement. On the other hand, this put the researcher in the same situation as the average provider – and the participant in the situation of the average user – of MOOCs and/or other online learning environments.

Thirdly, no interviews have been conducted with the participants, which limited the scope of the analysis to the factual results. This research does not include an analysis of the motivational aspects behind the results. Therefore, this research stayed at macro level and was not able to provide explanations of individual behaviour.

The last limitation is that since the respondents of the experiment were all in the direct or indirect network of the researcher, a sampling bias was inevitable. Within the group of respondents, respondents with a completed vocational education (3.1%) were underrepresented, while more than half of the respondents had a university degree or higher. This would have made it harder to generalize the results, however, according to Christensen, et al. (2013), out of the 34,779 MOOC participants, 79.4% had at least a Bachelor's degree, which suggests that the current sample represented the population accurately.

6.3. Future research

The aforementioned limitations lead to recommendations for future research. These are explained below.

First of all, all the respondents got treated with both two retention and two compensatory questions. Although this made it possible to test whether the two types of embedded questions were interacting, testing the pure effect of the two types of embedded questions therefore became highly complex. Therefore, it is recommended for future research create two groups of respondents. The first group should only get retention questions while the other group should only get compensatory questions.

Secondly, it is recommended to rethink the process of recruiting participants. For this research, respondents were recruited from the (in)direct network of the researcher and they did not receive any reward. Therefore, the main reason why these respondents actually participate came out of internal motivation. For future research, it might be interesting to repeat the current research but offer something in return for participation (e.g. respondents get a small present for participating).

Thirdly, it is recommended to include interviews in the data analysis. The added qualitative data would give insight in the motivational aspects of the participants and therefore deepen the understanding and interpretation of the qualitative results of the current research.

Additionally, motivation could be a significant factor in interacting with the embedded questions, however could not fit into the scope of the current research. Building further on the statement on Núñez (2017), that videos have a positive effect on the motivation of

students, the potential effects of the different types of embedded questions on motivation would be a new research area.

Lastly, as mentioned earlier, the amount of retention and compensatory elements were not equal in neither the pre-test nor the post-test. Therefore, it is recommended to replicate the current study with as many retention elements in both tests as compensatory.

6.4. Implications

Practical Implications

One of the main complaints of teachers is that the quality of video lectures is not good enough or are missing important information (Herreid & Schiller, 2013). For this research, the Component Display Theory (CDT) by David Merrill (1983) was useful for the video analysis. CDT has provided a good overview of what still needed to be included in the video to make the understanding easier. Therefore, for teachers and other video-lecture creators, CDT can provide a way to analyse the educational videos and see where improvements are needed.

Moreover, by adjusting the CDT to serve as a video audit and creating the structure of the experiment, the current research can lay the groundwork for the upcoming similar researches. As the area has not yet been researched, the structure of the experiment had to be created from scratch, while from now on, other researches can save a lot of time by just using this structure.

Theoretical Implications

In education, video (lectures) are being used more frequently. Although it has been proven in several researches that embedded questions in general are beneficial for knowledge gain (Strouse, O'Doherty & Troseth, 2013), knowledge about embedded questions in video lecture still needs more attention. The behaviour of embedded questions in this different medium is a valuable scientific addition.

Moreover, even less research was done about whether and how different types of embedded questions (might) have a different effect on knowledge gain. This research therefore should be seen as a first exploration into a new research area within the wider topic of embedded questions.

Lastly, this research can be seen as a response to the recommendations by Twessi (2016), to include a pre-test in the experiment. By including a pre-test, it was better possible to measure the pure effectiveness of the embedded question alone.

6.5.Conclusion

To conclude, the final answer to the research question, *What is the effect of retention and compensatory embedded questions on knowledge gain and video engagement?*, would be the following: In this video lecture with embedded questions, embedded questions were proven to enhance knowledge gain and resulted in a higher than average video engagement. However, despite the initial expectation, there was no difference in the effect of the two types of embedded questions.

No comparable research is known about the topic of retention and compensatory embedded questions in video lectures, while (online) lessons make use of interactive elements more and more. Although the difference between these two types of embedded questions has not been proven, the research is highly valuable from an innovative point of view.

This research should be seen as a first exploration about the effect of different types of embedded questions in video lectures on knowledge gain and learner engagement. As stated, this direction is new in this research field. Therefore, the current research can be looked at as the pioneer of researches about the effectiveness of different types of embedded questions and as a recommendable fundament for any research to come in the same topic.

References

- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A
 Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659-701.
 doi:10.3102/0034654316689306
- Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger III, H. L., & McDermott, K. B. (2008). Examining the Testing Effect with Open- and Closed-Book Tests. *Applied Cognitive Psychology*, 22(7), 861-876. doi:10.1002/acp.1391
- Algoufi, R. (2016). Using Tablet on Education. *World Journal of Education, 6*(3), 113-119. doi:10.5430/wje.v6n3p113
- Anderson, R. C. (1972). How to Construct Achievement Tests to Assess Comprehension. *Review of educational research*, *42*(2), 145-170. doi:10.3102/00346543042002145
- Andre, T., & Thieman, A. (1988). Level of adjunct question, type of feedback, and learning concepts by reading. *Contemporary Educational Psychology*, 13(3), 296-307. doi:10.1016/0361-476X(88)90028-8
- Bell, B. A. (2010). Pretest–Posttest Design. In N. J. Salkind, *Encyclopedia of Research Design* (pp. 1087-1091). Thousand Oaks: SAGE Publications, Inc. doi:10.4135/9781412961288
- Benton, T. (2015). An empirical assessment of Guttman's Lambda 4 reliability coefficient. *Quantitative psychology research*, 301-310. doi:10.1007/978-3-319-07503-7_19
- Bing, S. B. (1982). Role of adjunct questions and reading ability levels on rote and conceptual learning from prose. *Instructional Science*, *11*(2), 129-138. doi:10.1007/BF00154883
- Black, P., & William, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21(1), 5-31. doi:10.1007/s11092-008-9068-5
- Brar, J., & van der Meij, H. (2017). Complex software training: Harnessing and optimizing video instruction. *Computers in human behaior, 70*, 475-485.
 doi:10.1016/j.chb.2017.01.014

- Bridges, J., Stefaniak, J., & Baaki, J. (2018). A Formative Design Examining the Effects of
 Elaboration and Question Strategy with Video Instruction in Medical Education. *Journal of Formative Design in Learning*, 2(2), 129-143. doi:10.1007/s41686-018-0025-5
- Butcher, K. R. (2014). The Multimedia Principle. In R. E. Mayer, *The Cambridge Handbook of Multimedia Learning* (pp. 174-205). Cambridge: Cambridge University Press.
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M.
 R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. doi:10.1038/nrn3475
- Callender, A. A., & McDaniel, M. A. (2007). The Benefits of Embedded Question Adjuncts for Low and High Structure Builders. *Journal of Educational Psychology*, *99*(2), 339-348. doi:10.1037/0022-0663.99.2.339
- Choi, H. J., & Yang, M. (2011). The effect of problem-based video instruction on student satisfaction, empathy, and learning achievement in the Korean teacher education context. *Higher Education*, *62*(5), 551-561. doi:10.1007/S10734-010-9403-x
- Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. (2013, November 6). The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why? Pennsylvania. doi:10.2139/ssrn.2350964
- Conache, M., Dima, R., & Mutu, A. (2016). A Comparative Analysis of MOOC (Massive Open Online Course) Platforms. *Informatica Economică, 20*(2), 5-14. doi:10.12948/issn14531305/20.2.2016.01
- Connell, J. P. (1990). Context, self, and action: A motivational analysis of self-system processes across the life span. In D. Cicchetti, & M. Beeghly, *The Self in Transition: Infancy to Childhood* (Vol. 8, pp. 61-97). Chicago: University of Chicago Press.
- Conrad, R., & Donaldson, J. A. (2011). *Engaging the online learner: Activities and resources*. John Wiley & Sons.
- Crestani, F. (1997). Application of Spreading Activation Techniques in Information. *Artifical Intelligence Review*, *11*(6), 453-482. doi:10.1023/A:1006569829653

- Evans, B. J., & Baker, R. B. (2016). MOOCs and persistence: Definitions and predictors. *New Directions for Institutional Research*, 2015(167), 69-85. doi:10.1002/ir.20155
- Felker, D. B., & Dapra, R. A. (1975). Effects of question type and question placement on problem-solving ability from prose material. *Journal of Educational Psychology*, 67(3), 380. doi:10.1037/h0076616
- Felton, L. A., Keesee, K., Mattox, R., McCloskey, R., & Medley, G. (2001). Comparison of Video Instruction and Conventional Learning Methods on Students' Understanding of Tablet Manufacturing. *American Journal of Pharmaceutical Education, 65*(1), 53-56. Retrieved from https://pdfs.semanticscholar.org/6b4a/e88b488b9698906f3e34f3532a0dbbcceaf6.p df
- Fiedler, M. L. (1975). Bidirectionality of influence in classroom interaction. *Journal of Educational Psychology*, *67*(6), 735. doi:10.1037/0022-0663.67.6.735
- Frase, L. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology*, *59*(4), 244-249. doi:10.1037/h0025947
- Frase, L. T., Patrick, E., & Schumer, H. (1970). Effect of question position and frequency upon learning from text under different levels of incentive. *Journal of Educational Psychology*, 61(1), 52. Retrieved from https://files.eric.ed.gov/fulltext/ED030546.pdf
- Gilardi, M., Holroyd, P., Newbury, P., & Warren, p. (2015). The effects of video lecture delivery formats on student engagement. *Science and Information COnference*, (pp. 791-796). London. doi:10.1109/SAI.2015.7237234
- Gilboy, M. B., Heinerichs, S., & Pazzaglia, G. (2015). Enhancing student engagement using the flipped classroom. *Journal of nutrition education and behavior, 47*(1), 109-114. doi:10.1016/j.jneb.2014.08.008

- Gill, T. G. (2008). The Single Client Resonance Model: Beyond Rigor and Relevance.
 Informing Science: the International Journal of an Emerging Transdiscipline, 11, 281-310. Retrieved from
 https://www.researchgate.net/profile/T_Gill/publication/241869240_The_Single_Cli
 ent_Resonance_Model_Beyond_Rigor_and_Relevance/links/0a85e53359e38ba3270
 00000.pdf
- Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings.* Center for Educator Compensation and Reform. Retrieved from https://files.eric.ed.gov/fulltext/ED532068.pdf
- Guo, P. J., Kim, J., & Rubin, R. (2014). How Video Production Affects Student Engagement:
 An Empirical Study of MOOC Videos. *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 41-50). ACM.
- Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of educational research*, *56*(2), 212-242. doi:10.2307/1170376
- Herreid, C. F., & Schiller, N. A. (2013). Case studies and the flipped classroom. Journal of College Science Teaching, 42(5), 62-66. Retrieved from https://sciencecases.lib.buffalo.edu/cs/pdfs/Cases_Flipped_Classroom.pdf
- Hyunwoo, N., Schulzrinne, H., & Kim, K. H. (2016). *Youslow: What influences user abandonment behavior for internet video?* New York: Columbia University.
- Johnson, C. S., & Delawsky, S. (2013). Project-based learning and student engagement. *Academic Research International, 4*(4), 560-570. Retrieved from http://www.savap.org.pk/journals/ARInt./Vol.4(4)/2013(4.4-59).pdf
- Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T. Y., & Camerer, C. F. (2006). The hunger for knowledge: neural correlates of curiosity. Unpublished manuscript. Division of Humanities and Social Sciences, California Institute of Technology. Retrieved from https://www. researchgate. net/profile/Joseph_Wang6/publication/242568025_The_Hunger_for_Knowledge_Ne ural_Correlates_of_Curiosity/links/0046352ce084a8d668000000. pdf.

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19(4-5), 528-558. doi:10.1080/09541440601056620

- Koenigs, S. S., Fiedler, M. L., & Decharms, R. (1977). Teacher Beliefs, Classroom Interaction and Personal Causation. *Journal of Applied Social Psychology*, 7(2), 95-114. doi:10.1111/j.1559-1816.1977.tb01332.x
- Kolås, L., Nordseth, H., & Hoem, J. (2016). Interactive modules in a MOOC. 2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 1-8). IEEE.
- Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice,* 41(4), 212-218. doi:10.1207/s15430421tip4104_2
- Limniou, M., Schermbrucker, I., & Lyons, M. (2018). Traditional and flipped classroom approaches delivered by two different teachers: the student perspective. *Education and Information Technologies, 23*, 797-817. doi:10.1007/s10639-017-9636-8
- Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American educational research journal, 37*(1), 153-184. doi:10.3102/00028312037001153
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494-513. doi:10.1080/09541440701326154
- Merrill, M. D. (1983). Component Display Theory. In C. M. Reigeluth, *Intructional-design theories an Models: An overview of their current status* (pp. 253-296). New Jersey: Lawrence Erlbaum Associates.
- Merrill, M. D. (1987). The new Component Design Theory: instructional design for courseware authoring. *Instructional Science*, *16*(1), 19-34. doi:10.1007/BF00120003
- Michas, I. C., & Berry, D. C. (2000). Learning a Procedural Task: Effectiveness of Multimedia Presentations. Applied Cognitive Psychology, 14(6), 555-575. doi:10.1002/1099-0720(200011/12)14:6<555::AID-ACP677>3.0.CO;2-4

- Nojavanasghari, B., Baltrusaitis, T., Hughes, C. E., & Morency, L.-P. (2016). The Future Belongs to the Curious: Towards Automatic Understanding and Recognition of Curiosity in Children. *Proceedings of the Interspeech Workshop on Child Computer and Interaction*, (pp. 16-22). doi:10.21437/WOCCI.2016-3
- Núñez, J. C. (2017). Testing Audiovisual Comprehension Tasks with Questions Embedded in
 Videos as Subtitles: A Pilot Multimethod Study. *The EuroCALL Review*, 25(1), 36-60.
 doi:10.4995/eurocall.2017.7062

Quillian, R. M. (1962). A revised design for an understanding machine. *Mechanical translation*, 7(1), 17-29. Retrieved from https://pdfs.semanticscholar.org/9259/8a2ec14db01614cbcc911a93b94b762c31d6. pdf

- Rickards, J. P., & Di Vesta, F. J. (1974). Type and frequency of questions in processing textual material. *Journal of Educational Psychology, 66*(3), 354. doi:10.1037/h0036349
- Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, *3*(4), 241-249. doi:10.2307/1162033
- Rothkopf, E. Z., & Bisbicos, E. E. (1967). Selective facilitative effects of interspersed questions on learning from written materials. *Journal of educational psychology*, *58*(1), 56. doi:10.1037/h0024117
- Ryff, C. D., & Keyes, C. M. (1995). The Structure of Psychological Well-Being Revisited. Journal of Personality and Social Psychology, 69(4), 719-727. Retrieved from http://midus.wisc.edu/findings/pdfs/830.pdf
- Samuels, P., & Marshall, E. (2019). *Checking normality for parametric tests in SPSS.* Retrieved from The University of Sheffield Web site: https://www.sheffield.ac.uk/polopoly_fs/1.579181!/file/stcp-marshallsamuels-NormalityS.pdf
- Sorva, J., & Sirkiä, T. (2015). Embedded questions in ebooks on programming: useful for a) summative assessment, b) formative assessment, or c) something else? *Proceedings*

of the 15th Koli Calling Conference on Computing Education Research, (pp. 152-156). doi:10.1145/2828959.2828961

- Storme, T., Vansieleghem, N., Devleminck, S., Masschelein, J., & Simons, M. (2016). The emerging pedagogy of MOOCs, the educational design of technology and practices of study. *Journal of Computers in Education*, 3(3), 309-328. doi:10.1007/s40692-016-0070-5
- Strouse, G. A., O'Doherty, K., & Troseth, G. L. (2013). Effective Coviewing: Preschoolers' Learning From Video After a Dialogic Questioning Intervention. *Developmental Psychology*, 49(12), 2368-2382. doi:10.1037/a0032463
- Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 33*, 161-164. doi:10.1016/j.jarmac.2014.02.001
- Tweissi, A. (2016). *The Effects of Embedded Questions Strategy in Video among Graduate Students at a Middle Eastern University* (Doctoral dissertation, Ohio University).
- Vancouver, J. B., & Kendall, L. N. (2006). When Self-Efficacy Negatively Relates to Motivation and Performance in a Learning Context. *Journal of Applied Psychology*, *91*(5), 1146 – 1153. doi:10.1037/0021-9010.91.5.1146
- Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology, 24*(8), 1183-1195. doi:10.1002/acp.1630
- Wu, W., Yang, Y., Zhang, H., & Huang, S. (2014). Designing video teaching for postgraduate engineering education. *World Transactions on Engineering and Technology Education*, *12*(3), 539-544. Retrieved from http://www.wiete.com.au/journals/WTE&TE/Pages/Vol.12,%20No.3%20(2014)/38-Wu-W.pdf

Appendices

Appendix A – Transcript of the video including screenshots

Voorbeelden van gesloten vragen – Examples of closed

questions

In deze instructievideo neem ik een aantal tips voor het construeren van gesloten toetsvragen met je door en geef ik voorbeelden van gesloten vragen. In een andere instructievideo genaamd "constructie van gesloten vragen" is ingegaan op de kwaliteitscriteria, de type vraagvormen en de richtlijnen voor het maken van een besloten toets. Voorbeelden uit deze instructievideo komen onder andere uit de boeken "Toets kwaliteit in de praktijk: hoe maak ik goede toetsen met gesloten en open vragen" van Teelen Kennismanagement en "Toetsen in het hoger onderwijs" van van Berkel.

In de instructievideo over toetsmatrijzen als ook in de video over constructie van meerkeuzevragen heb je gezien dat naast bepalen van de inhoud ook wordt gekeken naar een moeilijkheidsgraad van de vraag.

Bij het construeren van vragen is het dan ook belangrijk om na te gaan op welk niveau de vraag moet worden gesteld.

Ga dat je met name om de **reproductie** van kennis, dan moet je vooral vragen naar informatie die studenten zichzelf moet herinneren. De student heeft dit letterlijk geleerd, en de toetsvragen bevragen dit. Denk bijvoorbeeld aan vragen over feiten en gebeurtenissen. Of een nog concrete voorbeeld: Waar op dit plaatje zitten de nieren?

Gaat het je er met name om de studenten **inzicht** tonen, dan moet een student op basis van de geleerde theorie bijvoorbeeld een bepaalde casus of gebeurtenis kunnen verklaren. Het gaat dan niet alleen om de kennis van bepaalde definities maar meer om die definitie te kunnen gebruiken om tot verklaringen te komen. Denk bijvoorbeeld aan het voorspellen van een vervolg of verklaren van de gebeurtenis. Een nog concreet voorbeeld is: De klachten van mevrouw A wijzen op ziektebeeld a b of c.

Moeilijkheidsg	raad van de v	ragen (nive	au)	
Reproductie		Waar op de	dit plaatje zitten e nieren?	>
				-
 EDV/TRB 			_	
 Weight and the second se				
● ● ● EportNet				

	Waar op dit plaatje zitten
Reproductie	de nieren?
	De klachten van mw A wijzen

Of ga dat je met name om dat de student de kennis **toe kan passen**. De kandidaat kan bijvoorbeeld in een toets een medisch dossier van de patiënt raadplegen en daarbij een vervolg advies geven.

Voordat je begint met de constructie van de toets met gesloten vragen moet je dus je toetsmatrijs hebben gemaakt zodat je weet over welk onderwerp je vragen gaat formuleren en op welk niveau je dat gaat doen.

In de instructievideo over de constructie van gesloten vragen heb je kunnen zien dat een vraag bestaat uit **een stam, en afleider en de sleutel**, het goede antwoord. In deze instructievideo neem ik een aantal tips met je door die van invloed zijn op de stam en geef daarbij telkens een voorbeeld.

Bij de voorbeeldige integreren we steeds aan de **kwaliteitseisen** bij gesloten vragen. We kijken allereerst of het zinvol is de vraag voor te leggen. Daarvoor moet de vraag **valide** zijn.

Als tweede kijken we of er sprake is van **betrouwbaar**heid. Is de vraag objectief?

Vervolgens moeten vragen voldoende **transparant** en onderscheidend zijn en de vierde kwaliteitseis is **bruikbaarheid**. Daarbij gaat het om de organisatorische aspecten die ook van invloed zijn op de toets. Mocht je hier meer over willen weten, kijk dan naar de instructievideo over kwaliteitseisen.

De eerste tip is een goede gesloten vraag bevalt een actueel, concreet en duidelijk probleem of kennisaspect. Een voorbeeld dat niet voldoet aan één aspect per vraag is bijvoorbeeld: *In Nederland is dit is het oprichten van de CITO in 1967 de trend ontstaan leerlingen op basisscholen regelmatig met geobjectiveerde schooltoetsen lastig te vallen.*

Is het juist of onjuist? In dit voorbeeld kun je zien dat het bevragen van een aspect per vraag belangrijk is want wat wordt hij nu precies gevraagd?

Is de CITO in 1967 opgericht? Bestond basisscholen allen 1967? Gaat het om lastigvallen? Of zijn school toetsen geobjectiveerd?



Voerbeelden van gesieten vinget	Gesloten vragen	· • •
Stam		
🗖 Een afle	eider	
🗖 De sleu	tel (het goede antwoord)	
• • • • • • • • • • • • • • • • • • •		B (#
Voorbeelden van gesloten vaagen		0 *
	Kwaliteitseisen	

Kwaliteitse	eisen
- Valida	



In Nederland is sinds het oprichten van de CITO in 1967 de trend ontstaan leerlingen op basisscholen regelmatig met geobjectiveerde schooltoetsen lastig te vallen **Vraag:** juist / onjuist



Gegeven In Nederland is sinds het oprichten van de CITO in 1967 de trend ontstaan leerlingen op basisscholen regelmatig met geobjectiveerde schooltoetsen lastig te vallen Vraag: juist / onjuist Het is duidelijk dat de betrouwbaarheid van deze vraag onvoldoende is, aangezien een verschillende interpretaties mogelijk zijn. Ook de validiteit is in het geding aangezien de tekstbegrip van studenten ook een grote rol kan spelen bij het beantwoorden van deze vraag. Je weet dus niet wat je wilt meeten.

Hoewel informatie soms overbodig is, kan het ook juist de raadzaam zijn om genoeg context te geven. Met name bij citaat is het belangrijk om een kader of stroming aan te geven. Anders kunnen studenten vanuit verschillende invalshoeken redeneren. Bijvoorbeeld bij de stelling het geheel is meer dan de som der delen. Geef de context bij de citaat weer, zoals: *"In de gestaltpsychologie"* puntje puntje puntje of *"Freud bespreekt het geval van"* puntje puntje puntje. Deze gegeven context maakte de vraag meer transparant en betrouwbaar.

Afleiden informatie maakt vragen minder efficiënt en vaak ook minder valide. De volgende tip is dus om afleidende informatie te vermijden. Een voorbeeld met afleidende informatie is de volgende:

Gegeven: Mulisch stelt in zijn roman Het stenen bruidsbed de zoektocht naar wat werkelijk gebeurd is aan de orde.

Vraag: De uitspraak De leugen regeert is afkomstig van:

- a. De hoofdfiguur en de roman
- b. Koningin Beatrix
- c. Harry Mulisch

Overbodige informatie in de vraag leidt studenten af, en kans op het verkeerde been zetten. In deze vraag worden dingen door elkaar gehaald dat werkt afleidend. Het gegeven in de vraag is overbodig en kost tijd en energie bij studenten om erachter te komen of de informatie van belang is, ja of nee. Ook hiermee maken de vraag minder valide en bovendien minder transparant.

De volgende tip is gesloten vragen te toetsen zonder strikvragen. Probeer niet de student in bijvoorbeeld een valkuil te laten lopen











middels een strikvraag. Een voorbeeld van een strikvraag is: de volgende: Wanneer een vliegtuig neerstort in de Noordzee tussen Engeland en België, moet volgens de Zeewet (1918) de helft van de overlevenden in Engeland worden begraven en de andere helft in België.

Vraag: juist of onjuist?

Nu worden overlevenden uiteraard niet begraven. Dit soort strikvragen dragen niet bij aan het meten van dat gene wat je met een toets wilt nagaan. De vraag verlaagt dus de validiteit.

Een ander voorbeeld van een strikvraag of vragen waarin te sterk naar details wordt gevraagd is de volgende: Wat is de hoofdstad van Denemarken?

- a. Kopenhagen
- b. Stockholm
- c. Oslo
- d. Geen van drieën

Deze vraag doet eerder een beroep op nauwkeurig lezen dan op je kennis van topografie. Ook hier meet je dus niet wat je wilt meten. Dus ook deze vraag verlaagt de validiteit.

Een tip om te voorkomen dat de taalvaardigheid van studenten gaat toetsen in plaats van de stof waar het om gaat, negatieve formuleringen kan je beter vertalen in eenvoudige positieve vraagstellingen. Een voorbeeld is: Welke van de volgende landen ligt niet in Afrika? Wanneer je vragen negatief formuleert, vraag je van studenten dat deze eerst een taaltrip toepast. Hiermee maak je de vraag moeilijker maar op verkeerde gronden.

Dit komt de validiteit niet ten goede, maar mocht je toch negatieve termen gebruiken, zorg er dan in ieder geval voor dat deze opvallen. Bijvoorbeeld door te onderstrepen of dik te drukken.









Nog een tip om te voorkomen dat je taalvaardigheid van de studenten gaat toetsen in plaats van de stof waar het om gaat is het vermijden van dubbele ontkenningen. Een voorbeeld van een dubbele ontkenning is Om te voorkomen dat er geen problemen ontstaan met de begripsvaliditeit van de toets is het raadzaam om een toetsmatrijs te maken. Juist of onjuist?

Een dubbele ontkenning werkt verwarrend en je meet voor een deel ook de taalvaardigheid van de student.

Dat zal meestal niet de bedoeling zijn en maakt de vraag daarmee minder valide.

Een dubbele ontkenning kan vaak gemakkelijk worden omgezet in een positieve zin waarbij soms wel de antwoord sleutel moet worden aangepast.

Ook van belang is het vermijden van absolute of vage formuleringen. Absolute formuleringen zijn formuleringen waarin woorden als niet, nooit en allemaal voorkomen. Door deze absolute termen alleen is het alternatief al om juist. Bij vage formuleringen worden woorden als vaak, enige of nauwelijks gebruikt. Deze formuleringen zijn discutabel omdat ze geen duidelijke grenzen aangeven.

Alleen door de toevoeging van vage formulering is het antwoord dus juist.

Een paar voorbeelden:

Alle mannen kunnen beter kaart lezen dan vrouwen. Het woord alle een absolute formuleringen in de stam, geeft hier al aan dat deze stelling, zeer waarschijnlijk, niet juist zal zijn. Je kan deze vraag ook stellen door of iemand goed kan kaartlezen wordt bepaald door zijn sekse.

Door het woord alle te vermijden maak je de vraag specifieker en meer valide.



Juist/onjuist







Voorbeelden van gesloten vragen			0 🖈
geen absolute of va	ige Z		
- LAN	T		
Alle mannen kunn	enb	🔊 dan vrouwen. De	ze stelling is
A. juist	VaXde		
B. onjuist			
Of ismand good ke	n kanatlanan wa	بالمعمماء المحممط فأن	selve Dere
stelling is	II Kaar tiezen wor	ut bepaalu uoor ziji	i sekse. Deze
iuist	Valide		
A juist	-		
origoise		7. Carlos and a second s	

Een voorbeeld van een vage formulering is: *regenwolken hebben een licht gewicht*.

Licht is een vage aanduidingen en maakt de vraag discutabel. Want wat is licht of zwaar? Als je met deze vraag wil meten of studenten weten dat vochtig lucht zwaarder is dan droge lucht, kan je de vraag beter in een vergelijking opstellen. Zoals: Vochtige lucht heeft een lager soortgelijk gewicht dan droge lucht.

Vage bewegingen maken een vraag minder betrouwbaar en zijn niet transparant.

Bij gesloten vragen is het ook van belang om te zorgen voor eenduidigheid van de gestelde vragen. Hiermee zorg je voor een goed fundament en een betrouwbare toets. Wanneer er achteraf een discussie kan zijn over het juiste antwoord, dan is het betrouwbaarheid lager. Vragen die opinies bevatten zijn subjectief. Bij deze vragen zijn deskundig het vaak oneens over het correcte antwoord.

Een voorbeeld is er bij: *De buitenspelregel bij voetbal verlaagd aantrekkelijkheid van het spel*. Bij dit voorbeeld is het beter een neutrale formulering te gebruiken of de citaat met bronvermelding te benoemen. Hiermee wordt het antwoord minder discutabel en dus betrouwbaar. Dit heeft betrekking op het 100% regel. Check bijstellingen of deze echt honderd procent juist of 100% onjuist zijn.

Dit waren de tips met voorbeelden die betrekking hebben op de stam.

Nu volgen een aantal voorbeelden die met de alternatieven te maken hebben.



Regenwolken hebben een licht gewicht Juist of onjuist

Vochtige lucht heeft een lager soortelijke gewicht dan droge lucht Juist of onjuist?



Vochtige lucht heeft een lager soortelijke gewicht dan droge lucht Juist of onjuist?



orhoekten van gesloten vragen.		· O
>~~		
Vermiid subjectiviteit		
De buitenspelregel bij voetbal verlaag: spel.	t de aantrekkelijkhei	d van het
	juist / onjuist	
'volgens Johan Cruiiff verlagat'		\frown
	juist / onjuist	100%
(Betrouw)		100%
baar		
▶ ▶ • • • • • • • • • • • • • • • • • •		a 🦊 ::

Workbolden wa politien regen	0 ^
Stam	
Een afleider	
De sleutel (het goede antwoord)	
Een afleider	
Een afleider	
▶ N 4) 95671635 ×	a 🧔 🔅

Omdat het lastig is om korte bondige antwoorden te formuleren, is de lengte van het antwoord vaak al een aanwijzing voor de juistheid ervan. Uit ervaring weten studenten vaak dat correcte antwoorden meer woorden vereisen dan de afleiders. Een tip is dus om plausibele antwoorden te formuleren van ongeveer gelijke lengte. Dan verlaag je de raatkans en gaan studenten de alternatieve serieus overwegen.

Een voorbeeld: *Hoe moet een heftruck worden geparkeerd als deze tijdelijk niet wordt gebruikt?*

Je kan bij deze vraag ook wanneer de stof niet goed beheerst uit de lengte en gedetailleerdheid van alternatief C al afleiden dat dat het juiste antwoord is.

Dat maakt de vraag dus niet valide.

Het is van belang dat de sleutel en afleiders qua lengte en formulering vergelijkbaar zijn.

Bijvoorbeeld tot deze formulering met kortere vergelijkbare antwoorden. Als een heftruck tijdelijk niet wordt gebruikt moet deze buiten de transportroute worden geparkeerd. Aan welke eis moet bij het parkeren nog meer worden voldaan?

Door de vraag met afleiders een sleutel zo te formuleren, wordt de vraag valide.





```
    In de buurt van de plek waar men bezig is met de vorken naar beneden
    Op een daarvoor bestemde plaats buiten de transportroute met de vorken plat
    op de grand, de handrem aangetrokken en de contactsleutel eruit
    Op een plaats buiten de transportroute
```





De volgende tip is om de alternatieve niet te laten overlappen. De afleiders moeten zodanig afwijken van de sleutel, dat er geen discussie ontstaat over het juiste antwoord. Met deze vraag: *Welke provincie is het grootst?* kan bijvoorbeeld discussie ontstaan omdat Gelderland het grootst is qua oppervlakte, maar Zuid-Holland het grootst is qua inwonersaantal.

Dat maakt de vraag minder objectief en betrouwbaar.

Ik heb het bij de voorbeelden bij de stam al gehad over absolute en vage formuleringen. Ook voor de alternatieve geld dat absolute en vage formuleringen niet bijdragen aan de validiteit. Bijvoorbeeld bij de vraag: *Wat is een kenmerk van een koe?* Door het gebruik van het woord altijd in de afleiders, weet je je bij voorbaat dat het juiste antwoord A moet zijn, waarin de vage term meestal is gebruikt.

De vraag kan beantwoordt worden zonder kennis van de koe te hebben, waardoor de vraag niet valide is.

Betere formulering van alternatieven, zonder absolute of vage termen zijn bijvoorbeeld:

- A. Ze heeft vier magen
- B. Ze is een carnivoor
- C. Ze is onevenhoevig

De volgende tip.

De stam en alternatieven moeten inhoudelijk en grammaticaal op elkaar aansluiten. In dit voorbeeld wordt in de stam gevraagd waar een verpleger in kan raadplegen, niet bij wie.

Aleen antwoord C sluit dus aan bij de vraag, omdat hier aanwijzingen worden gegeven die het juiste antwoord verraden, is de vraag niet valide genoeg.

Nog een tip bij de alternatieven.

Rangschik, deze in een logische volgorde. Om te voorkomen dat de sleutel stelselmatig op dezelfde plaats staat, of een patroon













in de antwoorden ontstaat kan je het best een vaste logische volgorde aanhouden. Denk aan een alfabetische volgorde, maar ook eenheden van tijd, of hoeveelheden vragen om een logische volgorde. Je maakt het de student bovendien onnodig moeilijk wanneer de alternatieven in een niet logische volgorde staan.

Logische volgorde maakt vragen transparant en meer valide.

Het alternatief, *geen van bovenstaande antwoorden* is berucht in de toetspraktijk. Deze vragen zijn niet per definitie slecht maar wel moeilijk, omdat je als student alle drie de antwoordmogelijkheden van A tot en met C moet kennen, in het geval van antwoord D. De laatste tip, die ik je hier bij geef is dit soort vragen te vermijden.

Bijvoorbeeld deze vraag: De Vries trekt uit de bewering van De Swaan ten aanzien van het geven van colleges, de conclusie dat docenten meer entertainer moeten zijn. Welke filosoof zal deze conclusie ondersteunen?

- A. Hegel
- B. Kierkegaard
- C. Poppel
- D. Geen van deze drie

Deze vraag is op zich geen slechte vraag, maar wel erg moeilijk, omdat van elk alternatief aangegeven moet worden of deze correct is, ja of nee. Het is transparanter om meerdere vragen te construeren uit deze ene vraag. Formuleringen zoals "*alle van bovenstaande*" of "geen van bovenstaande" zijn nooit het directe antwoord op mijn vraag. Je kan beter het enige juiste antwoord als sleutel gebruiken.

Dit waren een aantal voorbeelden om de kwaliteitseisen. validiteit, betrouwbaarheid, en transparantie toe te lichten. Ik zal de voorbeelden samenvatten in een aantal tips.

De stam en de vraag moeten een duidelijk probleem bevatten en geen overbodige informatie.









Vage en absolute aanduidingen kan je beter vermijden.

Dubbele ontkenningen kun je beter vermijden.

Vermijd ook subjectiviteit door het toepassen van de 100% regel en geef bij citaten ook een denkkader of context mee.

De tips met betrekking tot de alternatieven zijn:

Zorg voor alternatieve van vergelijkbare lengte, woordgebruik, en formulering.

Alternatieven moeten plausibel zijn en inhoudelijk, en grammaticaal aansluiten bij de vraag.

De afleiders moeten zodanig afwijken van de sleutel dat er geen discussie kan ontstaan over het juiste antwoord. Ook hier kan je denken aan de 100% regel.

Zet de alternatieven in een vaste logische volgorde, bijvoorbeeld alfabetisch

En construeren liever meerdere vragen dan dat je vaak werkt met het alternatief geen van bovengenoemde antwoorden.

Hiermee werdt aan het einde gekomen van de instructievideo "voorbeelden van gesloten vragen".

[Muziek]



Appendix B – Example of the video audit

Appendix C – Embedded questions and answers

Embedded questions and answers

Segment 1

Compensatory

Dutch

V: Wanneer stellen we dat een vraag transparant is?

A: Een vraag is transparant als het duidelijk is voor de student wat moet worden

beantwoord en wat de student moet doen om de vraag te beantwoorden.

English

Q: What is the definition of the transparency of a closed ended question?

A: A question is transparent if it is obvious for the student what is the question that needs to

be answered and what to do in order to answer

1 p: Obvious for the student WHAT is the

Total: 2 points (ALSO IN POST-TEST)

question

1p: Obvious for the student WHAT TO DO to

Retention

Dutch

V: Wat wordt er bedoeld met de validiteit van een gesloten vraag?

A: Een vraag is valide als het zinvol is om de vraag voor te leggen.

English

Q: What is the definition of the validity of a closed-ended question?

A: A question is valid if it makes sense to ask the question.

Total: 2 points

1p: If it is about something you want to investigate

2p: If it makes sense to ask the question

Segment 2

Retention

Dutch

V: Hoe kun je context aan de volgende vraag toevoegen ?

"Mannen leren zelfs wanneer ze zelf het zelf onderwijzen "

Juist/Onjuist

A: Een voorbeeld van context aan de vraag te geven kan de volgende zijn:

De volgende verklaring werd voor het eerst uitgesproken door Seneca: "Mannen leren zelfs wanneer ze iets zelf onderwijzen "

Juist/Onjuist

English

Q: How can you give context to the following question?

"Even while they teach, men learn"

True/False

A: E.g.: According to Seneca, even while they teach, men lea

E.g.: The following statement was first said by Seneca.

Total: 2 points

1p: Answer only mentions that context has to be given

2p: Answer includes a good example

Compensatory

Dutch

V: Waarom wordt een vraag meer transparant wanneer er meer context wordt gegeven?

A: Omdat door context te geven het voor de student gemakkelijker zal zijn om te begrijpen en te weten wat te doen en hoe te antwoorden.

English

Q: Why does giving context make the question more transparent?

A: Because it makes it easier for the student to understand and know what to do and how to answer.

Total: 2 points

1p: Makes the question clearer for the student

2p: Makes it clearer for the student what to do and how to answer

Segment 3

• Retention

Dutch

V: Hoe kunnen de volgende set van alternatieven anders worden gesteld zodat overlap tussen de alternatieven wordt voorkomen?

Hoe lang leeft een Orchidee gemiddeld?

- A. Het gaat na één jaar dood
- B. Het leeft heel lang
- C. Het leeft meer dan een jaar
- D. Het moet elk jaar herplant worden

A: Een voorbeeld van de alternatieven kan de volgende zijn:

Hoe lang leeft een Orchidee gemiddeld?

- A. Minder dan een jaar
- B. Een jaar
- C. Twee jaar
- D. Meer dan twee jaren

English

Q: What is an option for a different set of answers to the following question in order to avoid overlap?

How long does an orchid generally live?

- A. It dies after the first year.
- B. It lives for many years.
- C. It lives for more than one year.
- D. It needs to be replanted each year.

A:

- A. Less than one year
- B. One year.
- C. Two years.
- D. More than two years.

Total: 2 points (ALSO IN POST-TEST)

1p: Answer only states what the problem is with the alternatives (overlap).

2p: Gives a good example for the

Compensatory

Dutch

V: Wat kan je zeggen over de bruikbaarheid van deze vraag?

Welke van de volgende antwoorden is de beste indicatie voor een hoge moraal in het team van een supervisor?

- A. Wanneer werknemers zelden overuren hoeven te maken.
- B. De werknemers zijn bereid om de eerste prioriteit te geven aan het bereiken van groepsdoelen, waarbij zij hun persoonlijke verlangens ondergeschikt maken.
- *C.* De supervisor blijft graag laat op kantoor om alvast de volgende dag te plannen.
- D. De werknemers geven dure verjaardagscadeaus aan elkaar.

A: Het is niet efficiënt geformuleerd, omdat de antwoorden niet dezelfde lengte hebben. Dat vermindert de bruikbaarheid.

English

Q: What can you say about the usability of this question?

A: It is not efficiently formed, because the answers are not the same length. That decreases the usability.

Total: 2 points

1p: Answer states that the usability is relatively low

1p: Answer talks about the length of the alternatives

2p: Answer states that the usability is relatively low

AND the reason why it is low. Reason: because the

Segment 4

Compensatory

Dutch

V: Waarom wordt een vraag meer valide wanneer de opsomming van de vragen via een logische volgorde gebeurt?

A: Omdat op de vraag op deze manier de kennis van de studenten test over de bewuste vraag zelf, in plaats van te testen of iemand de verschillende alternatieven juist heeft geinterpreteerd. Op deze manier is het zinvoller om de vraag te stellen.

English

Q: Why does the logical order of the alternatives make the question more valid?

A: Because in that way the question tests the knowledge of the students about the question itself and not about how to make sense of the different answers. This way, it is meer zinvol om de vraag voor te leggen.

1p: It makes more sense to ask the question like this.

2p: It makes more sense to ask the question like this because this way it tests the

Retention

Dutch

V: Welke tips hebben baat bij het denken aan de 100% regel?

A:

- 1) Vermijd subjectiviteit van de stam
- 2) Alternatieven mogen niet tot discussie leiden

English

- Q: In case of which tips can the 100% rule help?
- A: 1.) Vermijd subjectiviteit van de stam
 - 2.) Alternatieven mogen niet tot discussie leiden

Total: 2 points

1p: Vermijd subjectiviteit van de stam

1p: Alternatieven mogen niet tot discussie leiden

Appendix D – Pre-test questions with answers

Welke stad is de hoofdstad van Nederland?

- A. Den Haag
- B. Rotterdam
- C. Amsterdam
- D. Utrecht

1. Hoe noem je het omcirkelde deel van het bovenstaande vraag?

Antwoord: Stam

Answer	Point(s)	Total: 2
De vraag	1	points
Stam	2	

Stel dat je een professor in wiskunde bent die wil bepalen of jij met jouw les de kennis van studenten over 'hypothese testen' significant hebt laten verbeteren. Daarom besluit je de scores van een test, die zij vóór de instructie hebben afgelegd, te vergelijken met hun scores op een andere test die zij nadien aflegden. Welke van de volgende t-tests is geschikt om in deze situatie te gebruiken?

2. Wat is het grootste probleem met de bovenstaande vraag?

Antwoord: Afleidende informatie OF

Overbodige informatie in de vraag

Answer	Point(s)	
Has to be read several	1	
times to be understood		May 2 painta ayan if
Too long	1	more are mentioned
Afleidende informatie	2	
Overbodige informatie	2	

Wat dient er te worden gedaan na het oplopen van een prikwond op de hand om infectie te voorkomen?

- A. Ga altijd direct naar de dokter voor een Tetanus-prik.
- B. Gebruik slechts een antibioticum als de wond pijnlijk is.
- C. Veeg de wond nooit af met alcohol tenzij deze nog steeds bloedt.
- D. Zorg ervoor dat er geen vreemde voorwerpen in de wond zijn achtergebleven.

3. Wat moet er in de antwoorden van de bovenstaande vraag worden veranderd om ze beter te maken?

Antwoord: Vermijd absolute formulering

Answer	Point(s)	
Only the	1 p	
words (altijd,		Total: 2 points
nooit) are		Even if both
mentioned		
Absolute	2 p	are mentioned
formulering is		
mentioned		

4. Aan welke vier kwaliteitseisen moet een gesloten vraag voldoen?

Antwoord:

- 1. Validiteit
- 2. Betrouwbaarheid
- 3. Transparantie
- 4. Bruikbaarheid

Answer	Point(s)	
Validiteit	0,5 p	
Betrouwbaarheid	0,5 p	Total: 2 points
Transparantie	0,5 p	
Bruikbaarheid	0,5 p	
Antwoord: Als het zinvol is de vraag voor te leggen OF

Gaat alleen over de opgegeven leerstof OF

Is niet moeilijker of anders dan in de voorafgaande lessen

Answer	Point(s)	
Meet wat je wilt meten	1 p	
Gaat alleen over de	1 p	Total: 2
opgegeven leerstof		points
Is niet moeilijker of anders	1 p	Even if
dan in de voorafgaande		three are
lessen		mentioned
het zinvol is de vraag voor	2 p	
te leggen		

Venus Williams is de beste atleet van dit decennium.

Juist/Onjuist

6. Waarom is de bovenstaande vraag niet betrouwbaar?

Antwoord: Subjectief

Geen context gegeven

Answer	Point(s)	
Subjectief	2 p	
Not using the word	1 p	Total: 2 points
"subjective" but		Even if both are
describing it		mentioned
Missing a source	1 p	
Geen context	2 p	

7. Wanneer zeggen we dat een vraag bruikbaar is?

Antwoord: Vragen zijn efficient geformuleerd, zonder franje, en zijn binnen de tijd te beantwoorden

Answer	Point(s)	
Mention	1 p	
efficiency		
Mention that	1 p	Total: 2 points
there should		
not be		
irrelevant info		

8. Waarom wordt een vraag door een "geen van de bovenstaande" antwoordoptie minder transparant?

Antwoord: Dat maakt de vraag moeilijker omdat van elk alternatief aangegeven moet worden of deze correct is ja, of nee.

Answer	Point(s)	
Because it makes the	1 p	
question harder		Total: 2 points
Because they need to decide	2 p	
about each given alternative		
whether it is true or not		

Appendix E – Post-test with answers

Hoeve	el provincies heeft Nederland?
(A.	10
в.	11
C.	12
D.	13

1. Hoe noem je het omcirkelde deel van het bovenstaande vraag?

Post-test

Antwoord: Afleider OF Alternatief

Answer	Point(s)	
Antwoord	1	Max 2
Afleider	2	points
Alternatief	2	

De Verenigde Staten moeten altijd een buitenlands beleid aannemen op basis van welke van de volgende?

2. Wat is het grootste probleem met de bovenstaande vraag?

Antwoord: Geen context gegeven OF Absolute formulering "altijd"

Answer	Point(s)	
Mention	1	
"altijd"		Max 2
Absolute	2	points
formulering		
Geef context	2	

Hoe lang leeft een Orchidee gemiddeld?

- A. Het gaat na één jaar dood
- B. Het leeft heel lang
- C. Het leeft meer dan een jaar
- D. Het moet elk jaar herplant worden
- 3. Herformuleer de set van bovenstaande antwoordopties zodat overlap tussen de alternatieven wordt voorkomen.

Antwoord:

- a) Minder dan een jaar
- b) Een jaar
- c) Twee jaar
- d) Meer dan twee jaren

Answer	Point(s)	
Answer only states what the problem	1	
is with the alternatives (overlap).		Max 2 points
Gives a good example for the	2	
alternatives, without overlap		

4. Aan welke vier kwaliteitseisen moet een gesloten vraag voldoen?

Antwoord:

- 1. Validiteit
- 2. Betrouwbaarheid
- 3. Transparantie
- 4. Bruikbaarheid

Answer	Point(s)	
Validiteit	0,5 p	
Betrouwbaarheid	0,5 p	Total: 2 points
Transparantie	0,5 p	
Bruikbaarheid	0,5 p	

Wat is, volgens David Attenborough, de belangrijkste reden waarom zoveel mensen in 1849 naar Californië zijn verhuisd?

- A. In centraal Californië werd goud ontdekt.
- B. Ze wilden religieuze nederzettingen vestigen.
- C. Het oosten bereidde zich voor op een burgeroorlog.
- D. Het land van Californië was vruchtbaar en goedkoop.

5. Wat kun je zeggen over de validiteit van de bovenstaande vraag?

Antwoord: De validiteit is relatief hoog, geen probleem met de validiteit.

Answer	Point(s)	
If the participant states in any	2	
form that the question is valid		Max 2 points
or that the validity is high, max		
points are awarded		

Kabeljauw smaakt beter dan zalm.	
----------------------------------	--

Juist/Onjuist

6. Waarom is de bovenstaande vraag niet betrouwbaar?

Antwoord:

- "Wanneer er achteraf een discussie kan zijn over het juiste antwoord, dan is de betrouwbaarheid lager" (Quote from the video) OR
- The questions is subjective OR
- The question needs context

Answer	Point(s)	
Participants describe	1	
subjectivity or the lack of context		
but do not use the words		
"subjective" or "giving context"		Max 2 points
Participant mentions that there	2	Max 2 points
can be a discussion about the		
good answer		
Subjective	2	
Geef context	2	

7. Wanneer stellen we dat een vraag bruikbaar is?

Antwoord: Vragen zijn efficient geformuleerd, zonder franje, en zijn binnen de tijd te beantwoorden

Answer	Point(s)	
Well	1 p	
formulated		
alternatives		
Mention	1 p	Total: 2 points
efficiency		
Mention that	1 p	
there should		
not be		
irrelevant info		

8. Waneer stellen we dat een vraag transparant is?

Antwoord: Een vraag is transparant als het duidelijk is voor de student wat moet worden beantwoord en wat de student moet doen om de vraag te beantwoorden.

Answer	Point(s)	
Obvious for the student	1	
WHAT is the question		Max 2 points
Obvious for the student	1	
WHAT TO DO to answer		