Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for naturvitenskap
Institutt for fysikk

Johan de Besche

Pictures on mathematics tests

An eye movement analysis of students solving a test with representative and decorative pictures

Masteroppgave i Lektorutdanning i realfag for trinn 8-13
Veileder: Rolf Jonas Persson
Medveileder: Berit Bungum

Mai 2023

NTNU
Kunnskap for en bedre verden

Johan de Besche

Pictures on mathematics tests

An eye movement analysis of students solving a test with representative and decorative pictures

Masteroppgave i Lektorutdanning i realfag for trinn 8-13
Veileder: Rolf Jonas Persson
Medveileder: Berit Bungum
Mai 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for naturvitenskap
Institutt for fysikk

NTNU
Kunnskap for en bedre verden

Sammendrag

Forskning på multimedia i læringssituasjoner er et veletablert felt. På lignende vis finnes det veldig mye forskning om vurdering. Kombinasjonen av disse forskningsfeltene, altså bruk av multimedia i vurderingssammenheng, er derimot lite utviklet. Noen studier som ser på multimediaeffekten i problemløsning og i testing finnes, og arbeid har blitt begynt for å utvikle en kognitive teori om multimedia for vurdering ut fra de mer velkjente teoriene kognitiv belelastningsteori og kognitiv teori om multimedia for læring.

På grunn av denne mangelen på forskning om multimedia og vurderinger, har denne oppgaven som mål å utforske hvordan inkludering av bilder påvirker elever sine lesemønstre og prestasjoner under løsing av en matematikktest. Spørsmålet var om det å inkludere enten et dekorativt eller et representativt bilde påvirker resultatene av testen eller lesemønstrene til elevene. For å besvare disse spørsmålene, bidro atten elever fra en skole i Midt-Norge ved å løse en matematikktest mens øyebevegelsene deres ble tatt opp av en eye tracker. Resultatene diskuteres i henhold til teoriene nevnt ovenfor og lignende tidligere forskning.

Konklusjonen fra resultatene var en ikke-signifikant indikasjon på at de dekorative bildene hadde en liten til medium negativ effekt på antall riktige svar, mens analyse av øyebevegelsene til elevene viste at de leste oppgaver med dekorative bilder noenlunde likt som oppgaver uten bilder. For representative bilder derimot, var lesemønstrene noe annerledes, mens testresultatene var omtrent de samme som for oppgaver uten bilder. Elevene brukte litt mer tid på å løse oppgavene med bilder, men også denne forskjellen var ikke-signifikant.

Abstract

Multimedia learning is a popular and well developed research field. Research on assessment is likewise widespread and deeply researched. The combination of multimedia assessment however, has received little attention until recently. A handful of studies exploring multimedia in testing exist, and there have been work done to develop a Cognitive Theory of Multimedia Assessment from the more well established Cognitive Load Theory and Cognitive Theory of Multimedia Learning

On the basis of this lack of research, this thesis aims to investigate how pictures affect students' reading patterns and performance on a mathematics test. The goal was to look for differences between test items that included or did not include representational and decorative pictures. To do so eighteen students from a highschool in mid-Norway participated and solved a test consisting of six exam-like problems in the subject Matematikk 1P. While they solved the test, their eye movements were recorded in order to gain insights into how they read the test items. The results of both performance and reading patterns are discussed in the context of the cognitive theories mentioned above, as well as previous similar research.

The results of the test indicated that decorative pictures might have had a small to medium negative effect on test scores, while the eye tracking analysis showed that students' reading patterns were mostly altered in the presence of representational pictures. Test items with representational pictures took slightly longer to solve than those without pictures, although no significant effect sizes on performance was found.

Acknowledgements

This thesis marks the end of my five year long education at NTNU to become a physics and mathematics teacher. In physics, mathematics and science, as well as in life in general, I believe that curiosity is an essential character trait. In choosing the topic and research method for this master thesis, I let my curiosity be my guide, and chose to investigate a new and unexplored research field in multimedia assessments, as well as a research method unfamiliar to me before this work in eye tracking.

The journey towards completing this work has been long, tiresome at times and extremely interesting, and was made possible by the invaluable help of my supervisors, Rolf Jonas Persson and Berit Bungum, providing answers to all kinds of questions I have had, and giving crucial and insightful advice throughout the year. Thanks for all your help in guiding me through this work. I would also like to thank the students who graciously participated in this study, and the help from their teachers and school in making this study possible. Finally, a special thanks to my friends, family and partner, who have been caring, supportive and understanding throughout this process.

1. Introduction

The usage of multimedia in instructional material such as textbooks and digital learning resources is widespread, and how it impacts learning is thoroughly researched (Mayer, 2002; Mayer & Mayer, 2014). Cognitive Load Theory (Sweller, 2010) and Cognitive Theory of Multimedia Learning (Mayer & Moreno, 1998) are commonly used frameworks for understanding how multimedia can positively affect learning. From these and other works, there exists a thorough understanding for how to design instructional material with multimedia.

Another large research area is that of assessment. From test theory (McDonald, 2013) to the more recent shift towards formative assessment (William, 2011), also called assessment for learning, our knowledge in these areas are vast. Despite this, research on multimedia usage in assessment is sparse, albeit growing in recent years. At present, the studies that do tackle this relation between multimedia and assessment, usually refer to cognitive learning theories as frameworks. However, this comes with the unclear assumption that the same principles apply to assessment situations as they do in learning environments. Kirschner, Park, Malone and Jarodkza (2017) has challenged this assumption, and proposed the early frameworks for a Cognitive Theory of Multimedia Assessment, whereby they suggest some of the instructional principles derived from multimedia learning should be reversed, in order to reliably measure levels of competence.

Looking deeper into the lack of research on multimedia assessment, a recent meta-analysis on multimedia effects on problem-solving (Hu et al., 2021) found only 26 studies that satisfied their inclusion-criteria, up until 2018. The analysis concluded that further primary studies were needed and noted that the Cognitive Theory of Multimedia Assessment remained empirically untested. In addition, Hu et al. (2021) mentioned that future studies should take advantage of eye tracking to allow for insights into picture processing during problem solving. Since then, more studies have been conducted (Lindner, 2020; Lindner et al., 2021; Moon et al., 2022; Erhart & Lindner, 2023;) investigating different aspects of multimedia assessments, with many of them using eye tracking to answer their research questions.

Both multimedia and assessment as terms, have definitions that span a large array of topics. To narrow it down, this present study's focus is on the usage of static pictures on written mathematics tests. Pictures are added to tests in many different forms, serving multiple functions. Therefore, a definition of the two types of pictures used in this study is warranted. Representational pictures (RP) are those that illustrate the entirety or some part of the written description of a problem, without revealing additional or unique information critical to the solution. Decorative pictures (DP) are made to be visually pleasing, but do not contain problem-specific information. These categories were chosen as they are in equivalence with relevant previous studies (Dewolf et al., 2015; Lindner et al., 2017; Lindner, 2020) and because both types of pictures are present on standardized Norwegian mathematics tests. (More information in 3.2.1 Test design)

To summarize, the purpose of this study is therefore to explore the effects of including either a representational picture (RP) or a decorative picture (DP) in text-based

mathematics test items, using the method of eye tracking. To accomplish this, an experiment was conducted where N=18 students' eye movements were recorded while they worked on a computer-based test consisting of six test items. The experiment used a within-subject design with the following three experimental conditions: Representational pictures, decorative pictures and text only. To investigate the effects on performance when adding these pictures to test items, the first research question was chosen to be:

RQ1: *"How does adding pictures to mathematics test items affect students' performance?"*

In addition, in order to explore the ways these students read and interacted with the different versions of the test items, the following additional research question was chosen:

RQ2: *"How does adding pictures to mathematics test items affect students' reading patterns?"*

As mentioned, the eye movement of participants across the computer screen was recorded during the tests. Eye tracking as a method was chosen to gain insight into how participants interacted with the different elements of test items as a function of the experimental conditions. Simply analyzing test results would not provide this insight. Traditionally, the most frequently used and important methods for probing cognitive processes are self-report measures and interviews (Rodrigues & Rosa, 2016). A big strength of eye tracking compared to these methods lies in the issues of validity, as it measures what participants actually look at, not what they report. The method is particularly useful for revealing mental representations, assessing subconscious aspects of thinking (Strohmaier et al., 2020) and investigating questions concerning the allocation of visual attention (Carter & Luke, 2020). As such, it has begun to attract the attention of educators in recent years (Rodrigues & Rosa, 2016), and was a fitting choice for this study.

Eye tracking is a new and growing method in the field of educational research (Strohmaier et al., 2020), and specific knowledge about the method is required in order to design studies, report findings and assess validity (Holmqvist et al., 2023). Therefore, a thorough review and explanation of the research methodology will be presented in chapter 3, based in part on groundwork laid by important literature reviews (Carter & Luke, 2020; Strohmaier et al., 2020; Holmqvist et al., 2023). Eye tracking has become a powerful tool for understanding learning processes (Jarodkza et al., 2017) and visual processing (Strohmaier et al., 2020), and has helped shape and strengthen instructional design derived from Cognitive Load Theory, Cognitive Theory of Multimedia Learning and Cognitive Theory of Multimedia Assessment (Mayer & Mayer, 2014; Kirschner et al., 2017).

These theories are presented below in chapter 2 Theoretical framework, along with a clarification of common terms used when discussing multimedia usage in assessments. In chapter 4, the methodology of the experiment conducted in this study is detailed, including how the test was designed and how data was collected and analyzed. After this, results are presented in chapter 5, and discussed on the basis of the two research questions presented above in chapter 6. Finally, a separate and somewhat different chapter is added at the end in chapter 7. Here I will describe what I have learned from the process of working on this thesis, how this connects to my education of becoming a mathematics and physics teacher, and what I would improve given the chance to run this experiment again with the knowledge I now have.

2. Theoretical framework

In previous research on multimedia usage in testing or problem-solving (Dewolf et al., 2015; Lindner et al., 2017; Lindner, 2020; Hu et al., 2021; Arneson & Offerdahl, 2023), cognitive learning theories are commonly used theoretical frameworks. This chapter will give an introduction to the most relevant aspects of two of these theories, namely Cognitive Load Theory (CLT) and Cognitive Theory of Multimedia Learning (CTML). In addition, the main ideas of the lesser known, but highly relevant Cognitive Theory of Multimedia Assessment (CTMA) will be outlined. First, however, a description of common terms and their definitions are given.

2.1 Term clarification

In the interest of being precise and consistent, clarification about some relevant terms are needed. First, multimedia is defined as any media that is presented in multiple ways, such as through a combination of pictures, animations, audio, text and so forth. In mathematical testing and problem-solving, the most common form of multimedia is the usage of pictures together with written text (Hu et al., 2021). Many types of pictures can be used, although the most researched type is called representational pictures. These are pictures that display task-relevant information that is already stated in the text, thereby representing the task text, either in part or fully (Lindner et al., 2017). The function of such pictures is to help construct mental models of the material.

The other type of picture relevant for this study is called decorative pictures. These are visually and aesthetically pleasing, but provide little to no task-relevant information, and are not expected to be helpful in problem solving (Mayer & Mayer, 2014). Instead, one might expect them to be a distraction, hampering the problem-solver. However, research suggests that these pictures capture very little attention, while simultaneously inducing better mood, alertness and calmness in learning situations (Lenzner et al., 2013)

Second, in much of the literature referenced above, the multimedia effect in learning is used as a term. This refers to the positive benefit of including multimedia elements in instructional material, as defined in both CLT (Sweller, 2010) and in CTML as the multimedia principle (Rudolph, 2017). Some studies discuss to which degree the multimedia effect is transferable to problem-solving (Hu et al., 2021) or testing (Lindner et al., 2017; Lindner, 2020) as well. The only theory currently that specifically tries to explain the role of multimedia in these settings is CTMA (Kirschner et al., 2017). The theory was proposed in 2017 and still lacks the depth of empirical foundation present in the more established cognitive theories (Hu et al., 2021). When discussing the general possible effects of the presence of multimedia in this paper, effects of multimedia will be used.

Third, the term assessment is key. As the name suggests, CTMA is focused on explaining multimedia in assessments. Assessments consist of a wide selection of methods with the intent of assessing learning, with testing being one of these methods. In standardized tests, such as the Norwegian written math exams, students are usually presented with test items consisting of a problem or task to be solved. As such, many factors are similar between problem-solving, assessment and testing. This study will use the term test for the present experiment, but also uses the terms assessment and problem-solving when referring to literature on these topics.

Fourth, when encountering a problem to be solved, the term representation is key. A division is made between external representations of a problem, and internal representations. The external representation of a problem is how the problem is presented (Hu et al., 2021). In order to solve this problem, one has to internalize the presentation into an internal representation, also referred to as a mental representation (Smith, 1998) or a mental model (Johnson-Laird, 2005). Finally, it is also necessary to externalize this internal representation in order to show the solution to the presented problem.

2.2 Cognitive theories

In this section, an overview of the cognitive theories used as a framework to discuss the results of the present experiment, is presented. In order, these are Cognitive Load Theory (CLT), Cognitive Theory of Multimedia Learning (CMTL) and Cognitive Theory of Multimedia Assessment (CTMA). The first two are well established theories explaining how to design instructional material that facilitate efficient learning. CLT is more general, but does define relevant effects about multimedia design. Meanwhile, CTML is, as the name suggests, more specific to the domain of multimedia and builds upon CLT amongst other core tenets of human cognition and perception. Theory on multimedia learning currently does not integrate motivational and affective factors, but this is stated as an important future research goal (Mayer & Mayer, 2014). As this study’s intent is to investigate effects pertaining to testing, a common form of assessment, the more specific CTMA is particularly relevant. As the names suggest, this theory builds on CTML and is presented last.

2.2.1 Cognitive Load Theory

Cognitive load theory's basic premise revolves around how human processing is constrained by working memory. Working memory, in contrast to long term memory, can only store a limited amount of information at a time (Sweller et al., 2019), and its capacity is further constrained when information has to be processed (Sweller et al., 1998). The effort required to process and complete a task is defined as the task's cognitive load. If the load becomes too high, the learner becomes overloaded, and unable to learn effectively. To be able to discuss how the total load changes based on the task at hand, three types of load need to be defined, namely intrinsic, extraneous and germane.

A useful way of formulating the various cognitive load effects is by element interactivity. Element interactivity is defined as the amount of elements of information that require processing at the same time (Sweller, 2010). In this way, intrinsic load is defined by the natural complexity of a given task. A high level of element interactivity results in a high intrinsic load, while the opposite low element interactivity results in low intrinsic load. A second consideration when determining intrinsic load is the learner's knowledge level. For example, when learning to read, each letter counts as an element, and they have to be processed together to form words, sentences and meaning. However, as one acquires expertise in the form of mental schemas, entire words or sentences can be represented as single elements. Therefore, intrinsic load becomes lower when expertise is higher.

Extraneous load is defined as load created by material outside of the content, two examples being background music or noise and irrelevant information (Rudolph, 2017).

This means extraneous load is imposed when faced with nonoptimal instructional procedures (Sweller, 2010). Therefore, if element interactivity can be reduced without altering what is learned, the associated load is extraneous. On the other hand, if changing the element interactivity does alter what is learned, the load is intrinsic.

Extraneous and intrinsic load are dependent on the element interactivity of the task material, with the exception of what constitutes an element based on the learner's knowledge levels. Meanwhile, germane load is defined only by learner characteristics and refers to the resources allocated to processing the intrinsic load of a task (Sweller, 2010). Outside of motivational considerations, the learner has no control over germane load. As an example, a task with high intrinsic load and low extraneous load, will require the learner to allocate a large proportion of their working memory, resulting in a high germane load. Now, by increasing extraneous load for the same task, the germane load is reduced, as the learner has to prioritize their working memory resources to dealing with the extraneous details of the instructional material (Sweller, 2010). A consequence of this definition is that germane load is positively correlated with learning, and as such the goal should be to reduce extraneous load so germane load can be kept high.

In summary, the total cognitive load imposed by instructional material can be because of element interactivity associated with both intrinsic and extraneous load (Sweller, 2010). To deal with these elements, the learner must allocate their working memory resources. The usefulness in this formulation is that it combats a possible contradiction when it comes to total load. As mentioned, when extraneous load goes down, the theory states that germane load goes up. However, countless experimental results show that total load does not remain constant in this situation (Sweller, 2010). As element interactivity determines the total load to come solely from the addition of intrinsic and extraneous load changes, the contradiction is eliminated. An increase in extraneous load also increases total load, while germane load is reduced.

In order to facilitate learning, a multitude of effects that relate to changes in element interactivity associated with extraneous load have been proposed and tested. This constitutes one of the main goals of CLT (Sweller, 2010). As the theory applies to all instructional material, only a selection of the most relevant effects for this study will be presented here. These include the redundancy effect, the multimedia effect, the split-attention effect and the expertise reversal effect.

The redundancy effect states that the exact same information should not be presented multiple times using different formats, as the learner will try to integrate all information presented (Jarodzka et al., 2017). This leads to higher element interactivity and total load, without having provided more information, thus hampering germane load and learning. In contrast, the multimedia effect states that presenting a subject matter in different formats, such as pictures or animations accompanying a text, leads to an ease of understanding.

The split-attention effect happens when learners are faced with multiple sources of related information that must be integrated together. If these sources are unintelligible to the learner in isolation, and at the same time are presented independently or far apart from each other, either spatially or temporally, working memory becomes overloaded and unavailable for learning (Sweller et al., 2011). Meanwhile, the expertise reversal effect

indicates that effective instructional material for novices can be less effective or even ineffective at instructing experienced learners (Paas et al., 2003).

2.2.2 Cognitive Theory of Multimedia Learning

The Cognitive Theory of Multimedia Learning asserts that people learn more deeply from words and pictures than from words alone (Mayer, 2014). This assertion is called the multimedia principle in learning. However, simply adding a picture to text does not create an automatic increase in learning, as the way these two elements are integrated matters. Therefore CTML aims to explain how to use pictures in combination with words in order to improve human learning (Mayer, 2014). In this section, the underlying assumptions of the theory is explained, and an overview of some of the derived instructional design principles for learning material is given.

CTML is based on three cognitive science assumptions (Mayer, 2014). The first is that the human information processing system has dual channels for processing of visual and verbal material. The second is that each channel has a limited amount of information it can process. This second assumption is shared with CLT, and forces the learner to choose which pieces of information to pay attention to (Mayer, 2014). Finally, the third assumption is the requirement for active processing, meaning to select, organize and integrate the given information with prior knowledge, in order to ensure meaningful learning (Rudolph, 2017).

Based on these three assumptions, CTML presents a model of information processing where said processing is divided into two systems, one visual and the other verbal. These systems work in tandem during the first two of three cognitive stages of learning, namely selecting and organizing incoming information to create both a visual and a verbal mental model. In the third stage, called integrating, connections between corresponding parts of the two models are made in order for the learner to integrate it with prior knowledge (Rudolph, 2017). The model explains the role of the three different memory stores, and the way in which different forms of information are processed in the dual channels, as explained in figure 2.1.

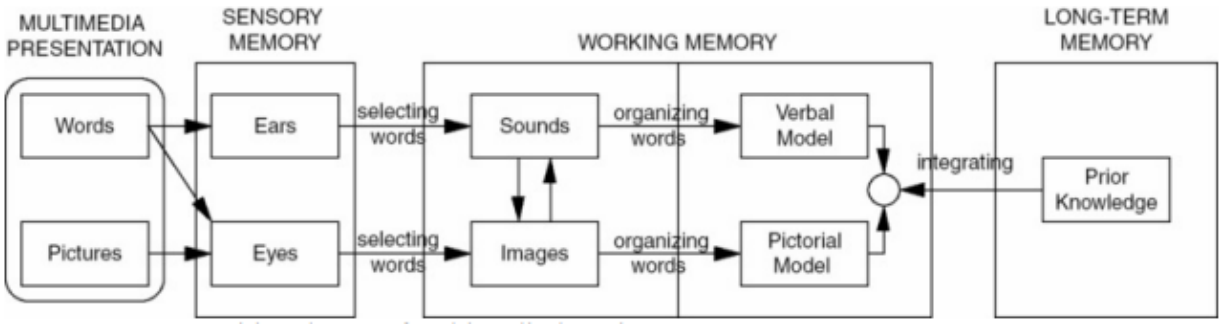


Figure 2.1 CTML's model of information processing. The leftmost box is the multimedia element presented to the learner, consisting of both words and pictures. This presentation is sent to the sensory memory as sensory inputs, as indicated by the arrows. Then a selection of relevant words and pictures are picked up in the working memory. The working memory continues processing by organizing these words and pictures into both a verbal and pictorial mental model. The final stage is that of integrating these two mental models with prior knowledge from long term memory. (Figure from Mayer & Mayer, 2014)

This model was initially tested in a series of experiments, the results of which yielded five major principles (Mayer & Moreno, 1998). Since then, several more principles have been added (Mayer & Mayer, 2014; Rudolph, 2017). In addition, Mayer (2014) has proposed three instructional goals following the structure and ideas of three types of load from CLT. The goals are therefore to manage essential processing, analogous to intrinsic load, to minimize extraneous processing, analogous to extraneous load, and finally to foster generative processing, analogous to germane load. A useful way of grouping the principles of CTML is by which of these three instructional goals they are helpful for achieving. The most relevant principles in the context of this study will be presented below.

Under management of essential processing are the multimedia and the pre-training principle. The multimedia principle states that instructional material containing both words and pictures are more effective for learning than material that only uses words (Rudolph, 2017). In presenting material by using multiple representations, the learner is able to build both a verbal and visual mental model, as well as connection between them (Mayer & Moreno, 1998). This results in an increase in learning, and is often referred to as the multimedia effect in learning, analogous to the multimedia effect as described in CLT (See section 2.2.1). The pre-training principle states that a description of names and characteristics of key topics prior to the lesson will lessen the essential processing needed (Mayer & Mayer, 2014).

Under ways of minimizing extraneous processing, we find the spatial and temporal contiguity, signaling and coherence principles (Mayer & Mayer, 2014). The spatial contiguity principle states that multimedia material should be presented with words and pictures placed near each other rather than spaced far apart. Similarly, the temporal contiguity principle states that pictures and words should be presented at the same time (Rudolph, 2017). The signaling and coherence principles state that highlighting of essential material is useful, and that extraneous material should be eliminated, respectively. Multiple other principles exist, many of them grouped under ways of fostering generative processing. However, these are less relevant to the present experiment, as they explain concepts such as usage of audio and design of learning sessions.

2.2.3 Cognitive Theory of Multimedia Assessment

Both the aforementioned theories concern themselves with designing optimal instructional material for learning. When designing test items, there is clear evidence against the usage of these instructional guidelines (Kirschner et al., 2017). A fitting example of this is the expertise reversal effect from CLT, where material that is highly effective with novices, becomes negative when used with experts (Paas et al., 2003). In addition, while there are plenty of guidelines on creating tests based on test item design theory, there are no principles for multimedia-based items (Kirschner et al., 2017). Because of this, Kirschner et al. have begun to work towards a theory of multimedia assessment (2017), where they take a deep look into the main differences between multimedia learning and multimedia assessment. This theory uses the base concepts of CLT and CTML such as information processing and cognitive load, but calls into question their principles of instructional design, because facilitating learning is no longer the main goal. As CTMA is the most relevant theory in regards to the multimedia test in the present experiment, the theory will be the main framework used in discussing the results of this study. An overview of the theory and its implications is presented below.

According to Kirschner et al.: "...the ultimate goal of assessment is to reliably and validly distinguish someone who knows something or can do something from someone who cannot / does not and / or determine who is a novice, who is an expert, and where someone is on the continuum between the two." (2017) A consequence of this goal is that assessments need to meet the three qualities objectivity, validity and reliability. Objectivity entails that the same results are to be reached by different coders. Validity means that subjects of similar skill or knowledge should come to similar outcomes, and finally reliability is that when repeating the assessment under similar circumstances, similar outcomes should be reached. Validity can be further divided into two aspects, ecological and criterion validity. Ecological validity is defined as to what extent the assessment demands are similar to typical tasks in the given field. Meanwhile, criterion validity is defined as the relationship between assessment score and the broader assessment criterion of the field or situation.

In CLT and CTML, extraneous load is avoided at all costs to make room for germane load and as a result, effective and efficient learning. However, as mentioned above, the goal of summative assessment is not to facilitate learning, but rather to separate the expertise level of subjects. In addition, there is a key difference between learning and problem solving that needs addressing. Finally, an important pedagogical note here is that assessment for learning, also called formative assessment, does share goals with the aforementioned learning theories. Using assessments for formative purposes has important consequences for classroom practices (William, 2011) and should therefore be considered when designing these for pedagogical use.

Coming back to the uniqueness of summative assessments, a test that uncritically adopts the multimedia learning design principles, can quickly run into problems of criterion validity (Kirschner et al., 2017). Criterion validity requires a distinct difference in performance between experts and novices, in favor of the experts. Therefore, one might have to increase extraneous load when designing test items, to be able to measure the amount of knowledge and skill that subjects possess. The optimal level of extraneous load would be that which both fosters instructional understanding and preserves validity, both criterion and ecological (Kirschner et al., 2017).

As explained in 2.1.1 about intrinsic load, experts with better schemata can chunk together elements in their knowledge structures. Because of this prior knowledge advantage, the element interactivity of intrinsic load is lower for experts. However, for a task to uncover this advantage, the total load needs to be such that an expert is able to perform it, while a novice struggles. Therefore, for less demanding tasks, an optimal level of added extraneous load can help support assessment (Kirschner et. al, 2017).

From CLT and CTML principles, we know that learning should facilitate the creation of a coherent mental representation, which is fostered by decreasing incoherence in the instructional material. But then the question becomes whether assessment material should be coherent or not, if it is to assess to what extent the subject has managed to acquire a coherent mental representation in the task or domain. CTMA hypothesizes that incoherences in assessment material might be beneficial for increasing criterion validity, as dealing with inconsistencies are an opportunity to show competence (Kirschner et al, 2017). This is an opposite principle from that of the aforementioned learning theories.

As mentioned previously, the second important factor when discussing differences between assessment and learning, is that of problem solving. When solving problems, the solver is expected to retrieve information from long-term memory and apply it to the given problem. This differs somewhat from a learning situation, where the goal is to process information from sensory memory and integrate it with prior knowledge from long-term memory.

These theoretical assumptions, as well as empirical evidence supporting the assumptions (Kirschner et al., 2017), makes up the argument that the uncritical transfer of guidelines for multimedia learning is problematic for the design of multimedia assessment. The lack of easy transferability from the learning theories over to assessment theory means that careful testing of the principles needs to be made in assessment situations, in order to expand our understanding of multimedia assessment.

Another factor of multimedia assessment that needs addressing is the differences between paper-based assessments and computer-based assessments. From the angle of paper-based assessments, there are strong theoretical underpinnings and empirically based principles (Kirschner et al., 2017). Computer-based assessments however, might require its own principles along with or instead of those derived from paper-based applications. As such, using multimedia in computer-based assessments should be done with caution and be justified by validity. In particular, the idea of visual distractors is key in eye tracking research (Gaspelin & Luck, 2018), and a question then becomes whether computers can increase the number of possible distractors.

To conclude, varying load to ensure ecological validity, but doing so carefully. Integrating the principles from CLT and CTML, but not by following them uncritically, rather by ignoring some of them to ensure appropriate load given the assessment situation. Keeping in mind the aims of the assessment and level of expertise of the group, and that presenting material on a computer screen is different from doing so on paper.

3. Eye tracking as a research method

Eye tracking is an experimental method used to assist researchers in understanding visual attention (Bergstrom & Schall, 2014). This is done by detecting eye movement and gaze placement on a visual target during a given time and task (Carter & Luke, 2020). In eye tracking research, the eye-mind connection, correlating eye movements to cognitive processes is usually assumed. The details, implications and validity of this correlation is discussed below in section 3.2. In addition, this section will include a description of how the eye functions, including a brief overview of the biological structure, as well as definitions of the different forms of eye movement. Further follows a review of the history of eye tracking technology. Finally, a discussion of the current state of existing technology, detailing possibilities, advantages and disadvantages of the method, and research agendas in the field of eye tracking in education research is given.

3.1 Eye function and movement

The eye's function is to gather and focus light and convert it into electrical signals called nerve impulses (Carter & Luke, 2020). Light is passed through the pupil, which can contract and expand to adjust image intensity. The cornea and lens focuses the light into a clear, inverted image on the retina. In the fovea centralis, a high concentration of cones works to create details and colors. Because of this, the fovea captures a detailed image of what the eye is pointed at. An overview of the anatomy of the eye is shown in figure 3.1, and of the visual field in figure 3.2 below.

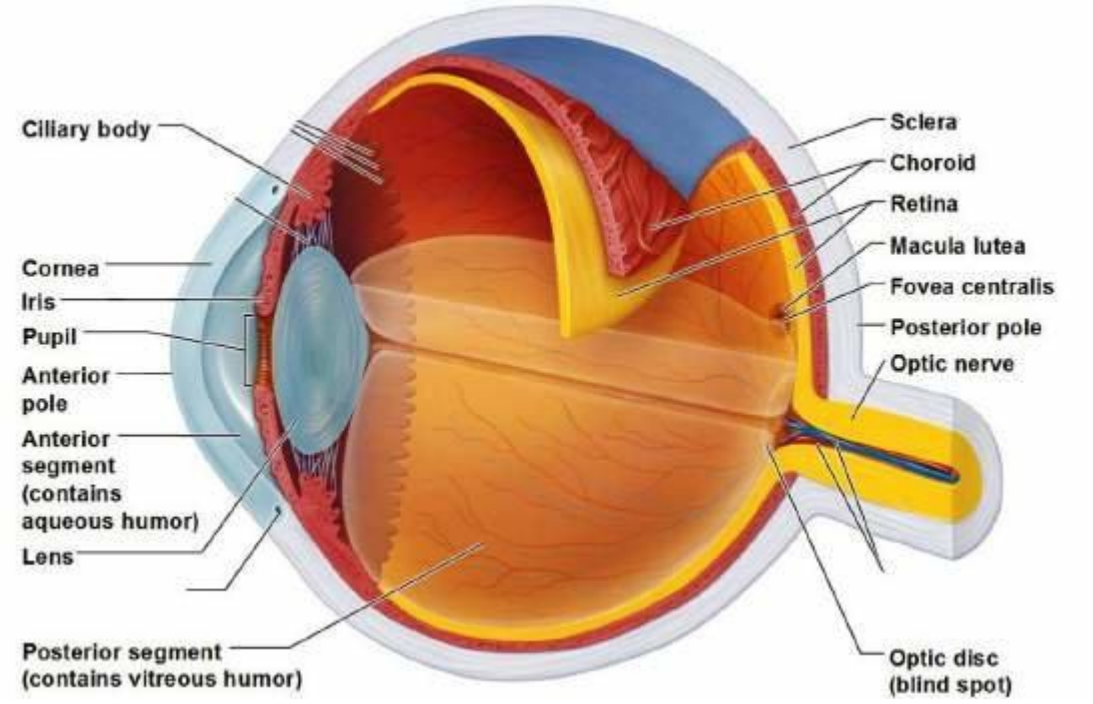


Figure 3.1 An overview of the eyes anatomy, showing the position of the pupil where light is passed through, the cornea and lens that is used to focus the light, as well as the retina and fovea centralis at the back of the eye where a high concentration of cones makes it possible to create detailed and colorful images (Figure from Sadek, 2014).

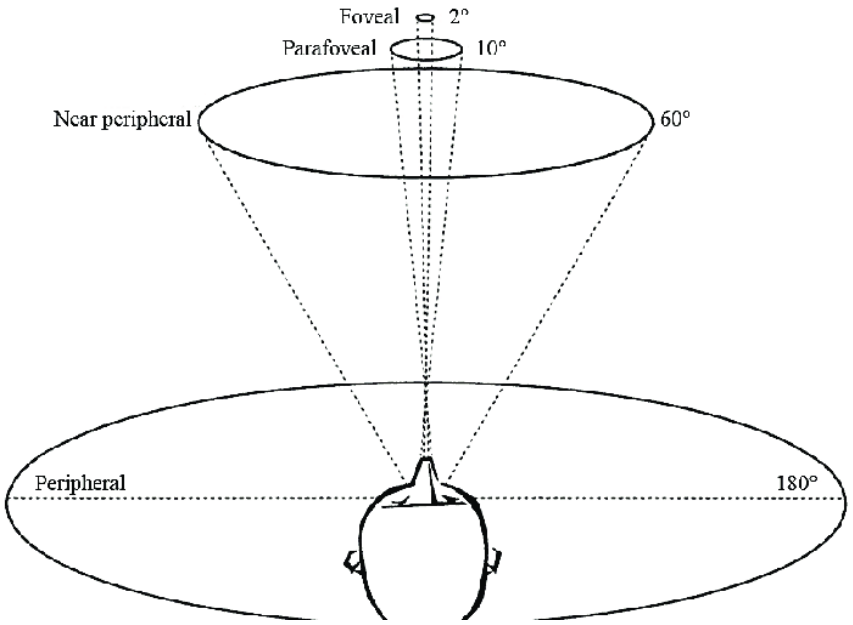


Figure 3.2 Schematic showing the visual field of the human eye, divided into foveal, parafoveal, near peripheral and peripheral. Foveal vision accounts for about half of the visual information sent to our brain, but consists of less than 8% of our visual field (Bergstrom & Schall, 2014). Eye trackers only track what is registered by foveal vision. (Figure from Ivančić Valenko et al., 2020)

A simple, but useful model of eye movement assumes that the eye either stands still, called a fixation, or moves rapidly between fixations, called a saccade (See figure 3.3). This model is used by most eye tracking algorithms when capturing eye orientation (See figure 3.4). Fixations are a pause of the eye movement on a specific area of a visual target, during which perception is stable and visual information is gathered (Carter & Luke, 2020). These typically last between 150 to 300 ms. Saccades are defined as rapid eye movements from one fixation point to another. These happen several times per second as the visual target is scanned (Rodrigues & Rosa, 2016). Saccade velocity and duration are dependent on the distance traveled (Carter & Luke, 2020), and are reported to take around 30 ms during reading, with about a 2 degree rotation of the eye, and between 40 to 50 ms for scene perception, with around 5 degrees of rotation. During saccades, new information is not acquired (Rayner, 1998).

This model does not take into account that the eye is capable of moving in other ways, for instance by smooth pursuit of a moving target, and by vergence, bringing the eyes closer together or further apart as the target moves closer or further from the viewer (Carter & Luke, 2020). All of the above motions are subject to voluntary control, unlike other ocular movements such as pupil diameter, optokinetic response and vestibulo-ocular reflex. In addition, during fixations, the eye is not perfectly still, but rather has both tremor, drift and micro-saccades that are filtered away by the eye tracking algorithm (Carter & Luke, 2020).

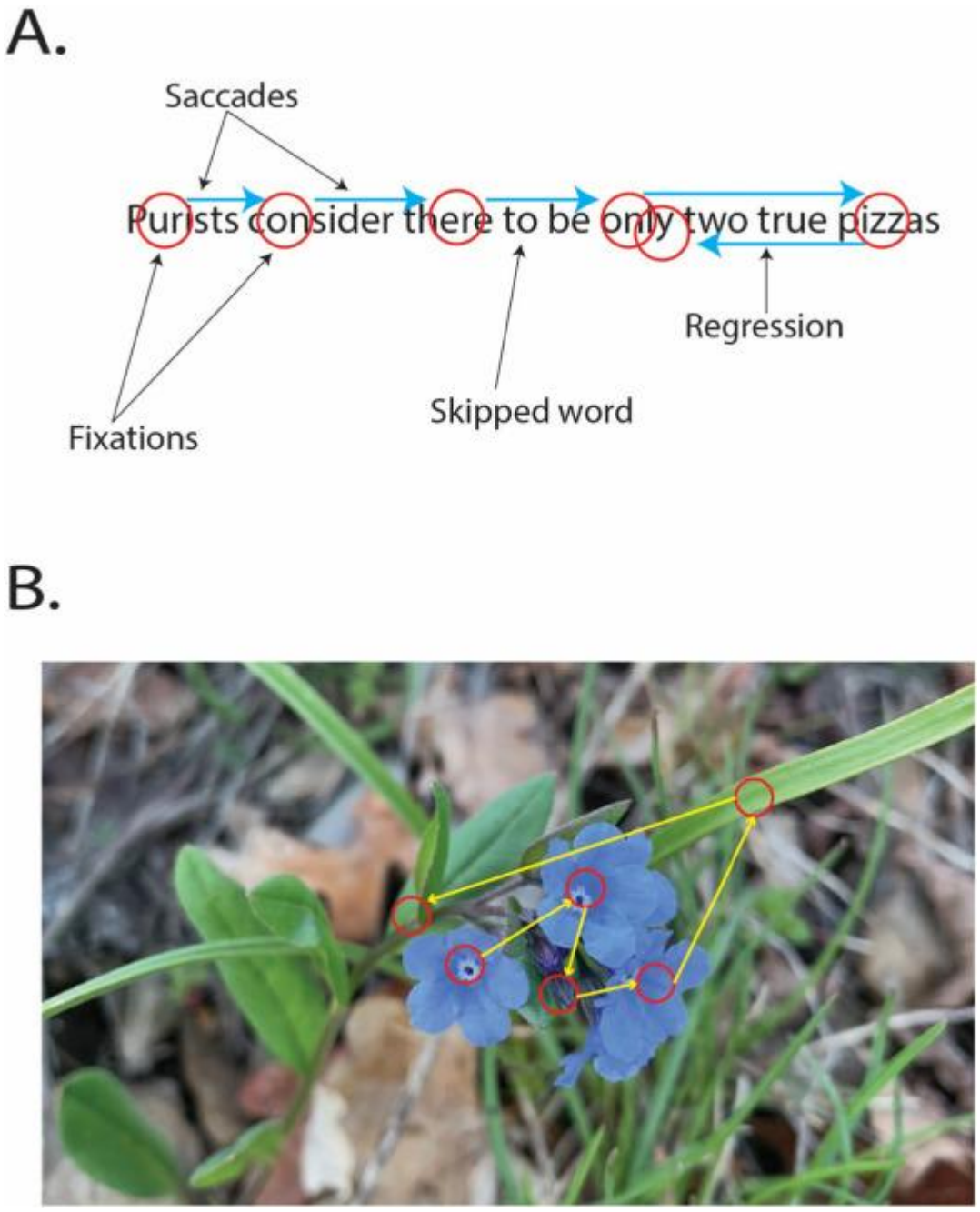


Figure 3.3 Fictive examples showing how gaze patterns can be visualized by capturing and displaying saccades and fixations on both text (A.) and a picture (B.) In the text (A.) fixations are shown as red circles, while saccades are drawn as blue arrows. In the picture (B.), fixations are again shown as red circles, while saccades are represented by yellow arrows. (Figure from Carter & Luke, 2020)

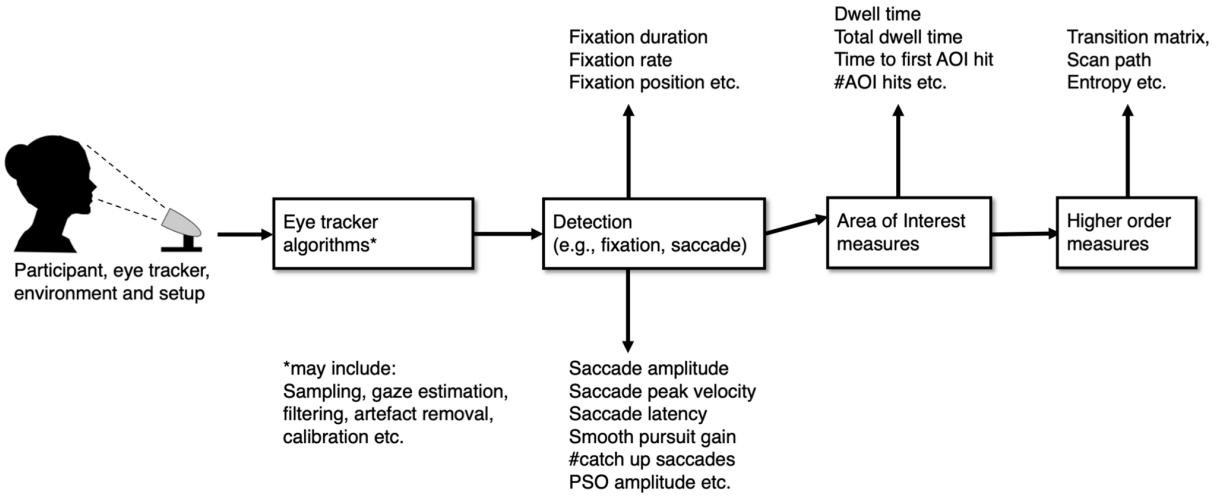


Figure 3.4 The steps of eye tracking analysis, from capturing eye orientation, to feeding the raw data into the eye tracker algorithm, which detects fixations and saccades, and finally the usage of software analysis to allow for Area of Interest and other measures. (Figure from Holmqvist et al., 2023)

3.2 Eye-Mind assumption

The correlation between eye movement and cognitive processes was first proposed by Just and Carpenter (1980) as the "eye-mind" assumption, and has since also been referred to as the "eye-mind" link (Rayner 2009, Reichle & Reingold, 2013) or hypothesis (Anderson et al., 2004). This assumption states that eye movements provide a history of where visual attention is being directed (Rodrigues & Rosa, 2016) by following the direction of which the fovea is pointed. As mentioned above, high acuity vision is limited to the fovea, a small portion of the visual field, because of the way the eye is structured. This strongly motivates us to move our eyes to point the fovea directly on whatever stimulus we are processing (Carter & Luke, 2020).

Researchers are widely in agreement that tasks involving complex information gathering and processing, such as reading, eye movements and attention are linked (Rayner, 1998). Also in mathematics education research, the assumption is feasible for interpreting most eye movement data, particularly so when participants are working on visual problems with a time restriction (Strohmaier et al., 2020). However, some experimental results are inconsistent with the assumption (Rodrigues & Rosa, 2016). For example, attention can move ahead of the eyes when reading (Underwood & Everatt, 1992), and memory retrieval can successfully be modeled independent of eye movements (Anderson et al., 2004).

More specifically to mathematics education and problem-solving, Schindler and Lilienthal (2019), showed that during a case study of geometry problem-solving, some eye movement patterns correlated well with data from a stimulated recall interview, while other eye movements were not in accordance with the interview data. They suggested that the eye-mind assumption does not always hold during geometry problem-solving and that even when the link between the two was there, some ambiguity of its interpretation still exists. This was only a single person case study, but as an initial guide, they mapped the eye movement patterns to interpretations given by the interviewee shown in Table 3.1 below.

Table 3.1 Interpretations of eye movements during geometry problem-solving, from a stimulated recall interview. (Quoted directly from Schindler & Lilienthal, 2019)

EMA holds	Fixations or small eye movements within a certain area	Processing corresponding information, in particular summing up the adjacent angles in a corner of the diagram
		Noticing a previous mistake
	Focusing on new areas of interest, differing from the preceding ones	Switching attention to an earlier approach (backward direction)
		Starting with a new approach to tackle the problem (forward direction)
		Getting an idea to approach the problem
	Looking back and forth between two corners	Double-checking the symmetry of the two angles
		Thinking how to use the symmetry of the angles for proceeding further
		Envisioning an auxiliary line mentally
		Considering the two adjacent areas of an envisioned line with peripheral vision
		Attention of the inscribed figure: considering whether to use the figure in his approach
EMA does not hold	Looking up left (outside the task sheet)	Thinking how to proceed next within a given approach (potentially to avoid distraction from the task sheet)
		Trying to remember a procedure
		Mental calculation
		Thinking about a new approach to solve the problem
		Thinking/processing information
	Quick eye movements on non-meaningful elements	Excitement about a discovery
		Stress because of time pressure
		Panicking, noticing a mistake

3.3 History of eye tracking technology

Carter and Luke (2020) ascribes Charles Bell to originating eye tracking research. He did so by the classification of eye movements and by linking these to control by the brain as early as 1823. According to Płużyczka, the earliest eye tracking experiments were done using observation of subjects reading with a mirror placed on the pages of the book read. This allowed for the observation of the reader's eye movements (2018). Already in 1879, Lamare concluded that reading was done by saccadic movements between fixations, rather than smooth, continual movement of the eyes. These conclusions were based on experiments using a blunt point placed on the upper eyelid connected to a microphone,

resulting in a sound each time the eye moved. Similarly invasive methods were tested by others during these early times, including attaching pens or sticks to plaster-of-paris rings placed upon the cornea. In the 1960s, Yarbus developed a corneal lens attached by suction to carry out experiments on perception by letting people view art paintings (Jarodzka et al., 2017).

The first non-invasive and precise experiments were done in 1901, starting the phase of optical eye tracking (Płużyczka, 2018). In these experiments, Dodge and Cline used light reflecting on the surface of the cornea and directed it on a photosensitive plate. In 1935 another researcher, Buswell, used prisms to capture the reflection of a light beam directed on the cornea onto film (Carter & Luke, 2020). Buswell's eye tracker was the first non-contact device (Płużyczka, 2018). Breakthroughs in design of eye trackers continued throughout the century, but researchers who studied eye movements prior to the 1970s generally stayed away from cognitive factors such as learning, attention and memory. Past this, two phenomena brought eye tracking to new possibilities: First, the establishment of cognitive psychology and second, the use of computer and television technology (Płużyczka, 2018).

This led to linguists and cognitive psychologists being some of the first researchers to take advantage of the possibilities of eye tracking. Reading and text comprehension is therefore a large part of the applied fields of eye tracking research. With the development and spread of personal computers, usability and human-computer interaction researchers also got significant value from the method (Bergstrom & Schall, 2014). Historically therefore, eye tracking research has its biggest theoretical and experimental basis in these three fields.

3.4 Modern eye tracking

The biggest drawbacks of eye tracking historically were the laboriousness of data analysis and expensiveness of precise light-sources and video recording devices. During the last few decades however, advancements in cheap technology and powerful software tools have opened up possibilities for more and more researchers to use eye tracking as a method in experiments. This has led to an exponential rise in publications where the method is featured, as shown in figure 3.5.

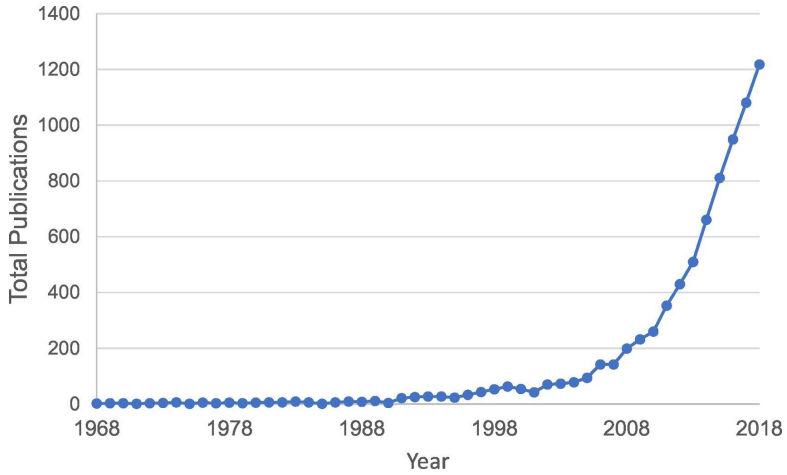


Figure 3.5 Number of publications per year in Web of Science where eye tracking is a topic. (Figure from Carter & Luke, 2020)

Currently, eye tracking is used in a diversity of research fields, from psychopathology, to neuroscience, ophthalmology, animal research and optometry to name a few (Holmqvist et al., 2023). In addition to these applied uses, there is also important research being done on how various aspects of studies using eye trackers, such as the participant, environment or instrument, affect data quality.

Eye tracking is a highly technical research method. Therefore, a thorough understanding is required in order to design studies and report sufficient information when publicizing research. Sufficient reporting is critical in order to assess validity and be able to reproduce studies. This need has been addressed by multiple literature reviews in the past few years, creating guidelines for usage and reporting (Carter & Luke, 2020; Fiedler et al., 2019; Strohmaier et al., 2020; Holmqvist et al., 2023). In this section, an explanation for how modern eye trackers function and an overview of the aforementioned reporting guidelines is given.

Most modern, camera-based eye trackers, including the one used in this study (Tobii x2-60), are so-called P-CR eye trackers (Holmqvist et al., 2023). These eye trackers send out infrared light that reflects off the cornea and uses a camera to capture the reflection as well as the pupil position. The P in P-CR stands for the pupil center in the camera image, while the CR refers to a reflection center in the cornea from the infrared illuminator. These two positional measurements are used to estimate gaze direction. To ensure accuracy and validity, a calibration of gaze direction is performed. Here, the participant is asked to look at multiple known locations across the visual target, and the data is compared with expected results based on the locations being known (Carter & Luke, 2020). Some issues regarding this technology may impair data quality, such as pupil occlusion, pupil color and mascara making it harder to detect the pupil's position (Holmqvist et al., 2023). Despite this, P-CR trackers dominate the eye tracker market, and are easier to operate than alternative technologies.

In their comprehensive meta-review, Holmqvist et al. (2023) found that guidelines from previous literature reviews on eye tracking differed in their suggestions and do not match with current reporting practices. This lack of regularity led them to review how aspects of eye tracker studies affect data quality, in order to create an empirical foundation for reporting guidelines. The results of the review were the following minimal reporting

guidelines. First, details about the eye tracker that was used, the setup and geometry, software version and exclusion criteria post-recording , as well as the sampling frequency (provided in 3.2.2). Second, a description of the recording environment, the instructions given to the participants and participation criteria (provided in 3.2.4). Third and finally, a measure of data-quality, a description of data processing and analysis (provided in 3.3).

As mentioned earlier, there has been a huge rise in eye tracker studies across all research fields in the last few years. A majority of the studies are from the field of psychology (Carter & Luke, 2020), but the method is also becoming increasingly popular elsewhere, such as in mathematics education (Strohmaier et al., 2020). In this research field, Strohmaier et al. (2020) reviewed the literature up until 2018. Here they found that eye tracking was used to study a wide range of topics, with the majority looking at numbers, arithmetics and fundamental processes such as counting and basic operations. Like multiple other reviews (Carter & Luke, 2020; Holmqvist et al., 2023), Strohmaier et al. (2020) noted that the reporting of eye tracker methods had large inconsistencies.

In terms of how to interpret results and what conclusions can be drawn from eye tracking studies, Strohmaier et al.(2020) noted that there were possibilities for qualitative, explorative and quantitative analysis. The usual cited advantage of eye tracking over other methods is the insights into cognitive processes that are otherwise not observable. This brings up the challenge of linking eye movement to said cognitive processes and the discussion of the eye mind assumption (section 3.2). Research has shown that the relation between mental representations and eye movements is often very strong.

A proposed measure to help clarify and interpret results is to have a plan on data interpretation before the study is conducted (Hessels et al., 2016; Carter & Luke, 2020; Strohmaier et al., 2020). In addition, relying on previous research and comprehensive guides when choosing what measures to report, can make it easier to compare studies and interpretations. This is especially important as there are a vast amount of different metrics when analyzing eye tracking data, and these are sometimes highly correlated (Strohmaier et al., 2020).

3.5 Relevant insights from eye tracking research

As mentioned, eye tracking as a method is used in a vast amount of different research fields and for many different purposes. For the purposes of this study, research using eye tracking to explore the effects of multimedia in problem solving and testing is most relevant. Therefore, a selection of highly relevant studies and research agendas in educational science will be presented here. Beginning with research agendas, an important contribution of eye tracking research currently is to examine the instructional principles of CLT and CTML under ecologically valid conditions (Jarodzka et al., 2017). More research is needed on this, as there are multiple unanswered questions about these theories' applicability to realistic learning environments. Two examples of studies where eye tracking shed light on some unexpected findings in regards to instructional principles are the following.

First, Jarodzka et al. (2015) tested the split-attention effect, using multimedia material on the topic of arts, where they introduced Dutch secondary school pupils to two versions of a test. One version presented the test items with pictures split far left off the rest of the

task, while the second version used an integrated approach, presenting the pictures within the text. Their findings, contrary to the predictions of CTML through the split-attention effect, was that pupils did not spend a lot of time integrating the information in the split version of the test. Rather, pupils chose to ignore that which they did not deem mandatory for solving the task, and actually performed slightly better on the split-version than the integrated version (Jarodzka et al., 2017).

Second, investigating the multimedia effect on the topic of vector calculus, Ögren et al. (2016) had university physics students solve eight true or false tasks. Half the tasks presented a representational picture along with the task text. According to the multimedia effect, the tasks including a representational picture should allow the students to build richer mental models and perform better. However, this was not the conclusion of the study. Instead, a bias towards believing tasks with representational pictures to be true was found. Using eye tracking, they also found that fixation time on the pictures did not relate to performance, but that fixation time on the task test did. Additionally, saccades transitioning between the picture and task text was also positively related to performance (Jarodzka et al., 2017).

As explained in section 2.2.3 about CTMA, insights into how experts differ from novices in a given field can be a huge benefit in designing appropriate assessment tasks for revealing levels of expertise. Also in this regard, eye tracking offers some unique benefits, and a second research agenda in educational research is therefore to investigate visual expertise as well as the effects of expertise on eye movements in visual domain such as when reading and solving problems (Jarodzka et al., 2017).

Furthermore, eye tracking research has been used to explore how humans interact with pictures in general, in learning situations and in problem solving. From these findings, it is known that eye movements are influenced by factors such as complexity and saliency of the stimulus (Carter & Luke, 2020). Eye-catching and complex stimuli receive more attention, and dark, blurry or low-quality stimuli also require more viewing time. Size of the stimuli also affects viewing times, with larger stimuli being viewed for longer. Many other factors can also affect eye movements, such as familiarity, novelty, meaningfulness and different task instructions on the same stimuli (Carter & Luke, 2020). Because of this, when making comparisons between stimuli in two experimental conditions, it is important to control for these factors.

When choosing what measures to analyze and report for this study, a review of measures used in similar research was done. Two prominent examples are presented here, together with their respective choices of measures. In the first example, Dewolf et al. (2015) looked at whether higher education students attended to representational illustrations while solving realistic word problems, so-called P-items. In this analysis, they used the measure number of fixations on the pictures, both in raw numbers and relative to the total number of fixations on the task. In addition, they analyzed the cases in which students fixated on a picture at least once to see if the fixation duration was long enough for pictorial processing to happen. Rayner et al. (2009) has shown that fixations need to be a minimum of 150 ms long to process a visual scene.

The second example is from research done by Lindner et al. (2017), where they looked at the use of representational pictures in a multiple choice science test. In this study, total fixation times on areas of interest (see 4.2.3 Areas of Interest) was used in the analysis.

Lindner has also been a contributor to a number of other studies looking at different facets of multimedia testing using eye tracking (Lindner et al., 2014; Strobel et al., 2016; Sass et al., 2017; Strobel et al., 2018, Strobel et al., 2019). In these studies, multiple choice assessment, graph reading, the split-attention effect and seductive details were investigated, and the studies all used total fixation time on areas of interest as the eye tracking metric for their analysis. Some also included other measures, such as transitions between areas of interest

4. Methodology

This chapter is divided into two parts, the first being section 4.1 Experimental design, detailing how the test used was created, what apparatus was used for data collection, what kind of participants were sampled and the procedure that was followed when conducting the experiment. The second section, 4.2 Analysis, explains how the data was used to present results and help answer the research questions from the introduction.

4.1 Experimental design

This study set out with the intent to answer what effects adding pictures to mathematics test items had on students reading patterns and performance. The usage of pictures, both decorative and representational, is common in testing environments such as the Norwegian standardized mathematics exams. However, their effects are not well known, making this an important and somewhat unexplored topic of research. Exploring these effects in physics was also considered, however, in their standardized testing material, less variation in picture types is present. In addition, high school physics pupils are often the same people that begin their academic careers as students. University students are often the subject of research, while first year (of high school) practical mathematics pupils are much less represented in multimedia assessment research (Hu et al., 2021).

When it comes to the research method chosen, measuring performance through an experiment consisting of a test with different types of pictures, seemed like a natural place to start. Following this, the usage of eye tracking was chosen as it is a uniquely beneficial method in investigating the possible reasons behind differences in performance, as well as uncovering insights into how the pupils interact with the test. Eye tracking as a method has been used in much of the existing research on multimedia in assessments and problem solving (Hu et al., 2021), and was a good fit considering the goal of this study. These factors led to the choices of research topic and method of this study.

To reach this study's goal, N=18 students were asked to solve a test consisting of six exam-based questions. The questions were presented on a computer screen that simultaneously recorded the test takers eye movements across the screen. The following sections detail how the test was designed, specifications of the apparatus that was used, sampling and procedure of the experiment, and finally, the way in which the gathered data was analyzed.

4.1.1 Test design

Six test items were selected from the standardized written exams in the Norwegian mathematics courses Matematikk 1P (Practical mathematics 1 for high school) and Matematikk 10. trinn (Mathematics for final year of secondary school) . Items that featured a picture necessary to solve the problem were not chosen, as it would be impossible to create a text-only version of such an item without altering the task text. All but one of the chosen items had a representational picture attached to them. A preference towards picking these items was made as high-quality representational pictures are more difficult to produce and benefit from going through quality assurance by user testing or feedback from experienced educators and assessment creators. One of the chosen test

items had a decorative picture attached, and for this item, a representational picture was created. For the other five items, decorative pictures were added.

Three versions of the test were created in order to follow a within-subject design (Charness et al., 2012) with the following experimental conditions: Representational picture, decorative picture and text-only test items. In the first version of the test, the first two items were presented with a representational picture, the middle two with a decorative picture, while the last two were without any pictures. The same distribution was used in the two other versions, creating a 3x3 Latin square (See table 4.1). This means that all subjects experienced each of the three conditions on two of six test items.

Table 4.1 Within-subject design of the test. Each group solved the same test items, but with different experimental conditions for each pair of items. RP=Representational pictures, DP=Decorative pictures, TO=Text-only.

	Item 1 and 2	Item 3 and 4	Item 5 and 6
Group 1	RP	DP	TO
Group 2	DP	TO	RP
Group 3	TO	RP	DP

The test items were originally open-ended problems or in one case a fill-in answer. For the purpose of this experiment, five multiple choices were created for each test item. The multiple choice format was chosen as it eliminates the need for interpretation of the test results, as well as allows for comparisons and a quantitative analysis, because the answers all follow the same format. A third advantage of the format is that for open-ended response formats, participants would have to spend more time writing down their answers, during which eye tracking data can not be collected, as the eye tracker used in this experiment only tracks eye movements across the computer screen.

To eliminate variance in the formatting of the test items, all versions of the items had the exact same task text and multiple choices, in the same position of the screen. When a picture was presented together with the task text and multiple choices, the picture was placed above the task text. An example of how this looked is shown in figure 3.7 below. All three versions of all six test items are included in Appendix A. The test items were presented on the screen at a height of 28 cm and width of 19.5 cm, corresponding to a viewing angle of about 25° vertically and 18° horizontally given the typical viewing distance of 60 cm.

En bensintank har form som et rett, firkantet prisme. Tanken er 40 cm bred, 90cm lang og 30 cm høy. Hvor stort volum har tanken?

a) 108000 cm^3

b) 10800 cm^3

c) 180 cm^3

d) 1800 cm^3

e) 18 m^3

En bensintank har form som et rett, firkantet prisme. Tanken er 40 cm bred, 90cm lang og 30 cm høy. Hvor stort volum har tanken?

a) 108000 cm^3

b) 10800 cm^3

c) 180 cm^3

d) 1800 cm^3

e) 18 m^3

Figure 4.1 Decorative picture and text only versions of test item 6, showing the formatting of all test items, with the picture presented on top, the task text in the middle, and the multiple choices listed vertically below. The task text translated to english reads: "A tank of gasoline has the shape of a straight, four-sided prism. The tank is 40 cm wide, 90 cm long and 30 cm high. How large is the tank's volume?"

4.1.2 Apparatus

In this section, specifications of the eye tracker hardware and software, as well as the monitor that was used, is given. These specifications were chosen in order to adhere to minimal reporting guidelines in eye tracking research (Holmqvist et al., 2023). Sample rate refers to how many times a second the eye tracker records gaze position, while the stated accuracy and precision measures are the ones reported by Tobii (2014), measured under controlled conditions. Both eyes were recorded. In addition, a description of the apparatus setup is given below, including a picture of the setup in figure 4.2.

Eye tracker hardware:
Model: Tobii X2-60 Compact
Sample rate: 60 Hz
Accuracy: 0.4 degrees
Precision: 0.34 degrees

Software:
Version: Tobii Studio 3.4.3.1267 Professional edition
Calibration: 9-point regular calibration at medium speed
Fixation filter: I-VT filter
Eye selection: Average of both eyes
Minimum fixation duration: 60 ms

Monitor:
Model: DELL U2414H
Resolution: 1920x1080
Aspect ratio: 16:9
Size: 21.2x14 inch
Refresh rate: 60 Hz

The monitor was positioned 60-65 cm away from participant's eyes and height about level with the middle of the screen. An adjustable chair was used to standardize head position and viewing distance of participants. Freedom of head movement is reported in the Tobii User's Manual as 50 cm in width and 35 cm in height, with an operating distance from eye tracker to participant of 45-90 cm (Tobii, 2014). No chin support was used.

Figure 4.2 Picture of the eye tracker setup. Included in the picture is the eye tracker hardware, consisting of the camera and IR light (1), the monitor (2), mouse (3) and keyboard (4), as well as the eye tracker processing unit (5) and computer (6). Only gaze positions on the monitor were recorded.

4.1.3 Sampling, validity and generalizability

The participants in the present experiment were all students attending the course Matematikk 1P (Practical mathematics for first year high-school students). The selection of participants was done using convenience sampling, as participating in the experiment was voluntary, and only two classes, a total of about 40 students, were asked. Out of these, N=18 students (age 16 to 17, 61% female) agreed to participate, by signing an informed consent sheet. Both classes were represented in the participant group, with eight coming from one class, and ten from the other. According to their teachers, the classes both performed about average when compared to the usual Matematikk 1P class. The sheet was constructed after NSD guidelines (Personvernerklæring, 2020). NSD stands for the Norwegian Center for Science Data, and is an organization that handles research data in a legal and safe way, and assists research projects with getting legal access to personal data. NSD approved the handling and collection of data for this study. No other participation criteria was set, however none of the participants were wearing glasses or reported any visual impairments.

When discussing whether the eighteen participants were representative for students in Matematikk 1P, one has to take into account the low sample size, the fact that all participants came from the same school and finally that participation was voluntary. The low sample size means that natural variance between individuals can affect results to a higher degree than with sample sizes suitable for quantitative analyses of statistical significance. In addition, a possible bias is present in that all students are from the same school, having the same teacher and using the same text-book, creating a less diverse mathematical background than one can expect to find in the general population of students. This can also be viewed as a strength, as it creates a more homogenous participant group, which might rule out certain variables that could affect the results of the study. Another possible bias occurs from the fact that students that had a reason not to want to participate likely did not do so. This could for example be test-averse students, or students that valued not missing any of the regular classroom activities happening simultaneously with the experiment.

In terms of the test design, standardized tests were used to pick test items, and as such the validity of the test should be solid (Kirchner et al., 2017). One aspect that might influence the ecological validity of the test in a negative direction is the environment in which the test was given for this experiment. As explained in detail in the next section (4.1.4 Procedure), participants sat in a room with only the researcher present, the test was presented on a computer screen, and the participants were informed that the test was strictly for research purposes and would not affect their assessment in the subject Matematikk 1P. These factors all differ from the usual testing environment, as most written mathematics tests are presented to these students on paper in the classroom with all students present, and the tests have an impact on the final grade of the subject.

A fourth factor worth mentioning is the participants' awareness of having their eye movements recorded. The first effect this may have is that participants can become more conscious of where they look, especially at the beginning of recording. In addition, participants were informed that large head movements would make eye tracking data collection difficult. Having to sit upright and not move their head too far from the initial

seating position could potentially be different than how these students usually sit when taking tests, which could reduce the validity.

The test item's original form on the standardized tests were mostly open response formats, meaning they posed a question that the test taker was tasked to answer in writing. One question was an exception to this, as it asked the test taker to fill in a numerical answer. However, for the test used in this experiment, multiple choices was used as the response format for all test items. This required the creation of suitable multiple choice options for all the test items. Therefore, the usage of multiple choices, with their advantages and disadvantages, should be addressed.

When creating multiple choices, there are some key concepts to be aware of. In order to design good multiple choices, an understanding of plausible distractors are needed (Angell et al., 2019). This requirement makes the design of multiple choice questions more difficult and time-consuming than other types of questions (Jovanovska, 2018). Another disadvantage of multiple choice questions is that they are susceptible to guessing, thus reducing their reliability (Jovanovska, 2018). This is offset by increasing the number of choices available, and they are less susceptible to guessing than true/false questions. At the same time, multiple choice questions are easier to score, as they dont require subjective interpretations of partial answers (Jovanovska, 2018). Another benefit is that answering these questions takes less time than when using open response formats, which allows for testing over a broader course material in the given time of the test (Jovanovska, 2018).

When it comes to eye movement data, the terms accuracy and precision are central in determining quality. Accuracy refers to whether the measured eye position corresponds to the actual eye position, while precision is how consistent the measurements of eye position are. Both accuracy and precision values of the hardware used are reported above. These values describe the capabilities under ideal conditions, and are subject to variation depending on a number of factors (Carter & Luke, 2020). In particular, accuracy is also highly dependent on the participant, the calibration and the setup (Holmqvist et al., 2023). As mentioned when discussing data quality in section 3.4, for P-CR eye trackers, these factors include partially occluded pupils because of sleepy participants, dark eyelashes or mascara (Carter & Luke, 2020; Holmqvist et al., 2023). Multiple studies (Holmqvist et al., 2011; Rayner et al., 2012) have shown that age matters as a variable affecting eye movements. This means the generalizability of eye tracking results across age groups might be limited (Strohmaier et al., 2020).

4.1.4 Procedure

Before participants began the test, a short introduction to the test content was given. This included participants being informed that the test consisted of six multiple choice items, that the test items were based on exam questions and that they had a fifteen minute time limit to complete the test. The participants were not explicitly informed about the experimental conditions, but knew that the usage of pictures was a focus of the research. In addition, they were informed about how the eye tracker worked, that it only recorded eye movements across the computer screen and how calibration would be performed.

Participants were asked to find a comfortable seating position with appropriate viewing distance from the screen. The chair had an adjustable height, making it possible for all

participants to view the screen from the same angle. They were also informed that large head movements could result in poorer eye tracking data quality. In addition, a 9 point calibration of the eye tracker was conducted before starting data collection. They worked on the test individually, with only the researcher present in the room, at their highschool in mid-Norway. The room had no visually distracting elements behind the computer screen, and was well lit.

All participants were assigned one of the three versions of the test at random. The experiment was conducted during their mathematics education hours. A 15 minute time limit to complete the test was set, and the participants were allowed no aids other than a pen and paper to scribble and to write down their answer from the multiple choice options in a pre-existing table found on the paper. All participants signed an informed consent form and answered all test items, except for one participant that skipped one task.

The decision to assign a time limit was to simulate a more realistic test situation, increasing validity, while the intent of disallowing aids was to decrease the amount of time spent looking away from the computer screen. This also had the second benefit of possibly reducing the motivation for body and head movements during the experiment, which could lead to a loss in data quality. In addition, they were asked to rate the perceived level of difficulty of the test as a whole, when compared to the usual tests they encounter in the subject on a scale from 1 to 10, where 1 was defined to be much easier, and 10 to be much harder than usual tests.

4.2 Analysis

When analyzing the data from this experiment, a combination of methods were used. The data consisted of test results as well as eye tracking data. An overview of what was done to analyze both test results and eye tracking data is presented below, with added explanations of how eye tracking data analysis works and can be done.

4.2.1 Performance

When analyzing the test results in order to answer the research question about performance, two metrics were considered. The first was the test score, which was calculated both per item and per experimental condition. In addition, effect sizes of test scores per experimental condition were calculated using Cohen's delta, which is done by taking the difference between means and dividing them by the pooled standard deviation. The test scores were also compared in a dependent t-test in order to check for significance, using a two-tailed hypothesis and a 0.05 significance level. The second metric used was time on task (ToT), meaning the time a participant spent working on a particular test item. The same calculations of Cohen's delta and significance were made for ToT. These two metrics together are usually what is measured when investigating performance of a given test or problem-solving situation (Hu et al., 2021).

4.2.2 Eye tracking analysis

Eye tracking data consist of a series of samples containing coordinates for points on the visual target that was tracked (Carter & Luke, 2020). The number of samples are determined by the sampling rate of the eye tracker. Unless necessary, working with this raw sample data is cumbersome and undesirable. Instead, an eye tracking algorithm processes the data to identify fixations and saccades (as explained in 3.1.1). When multiple temporally ordered samples are spatially close to each other, these get assigned as a single fixation. Meanwhile, if two samples are temporally adjacent, but spatially far enough apart, this is registered as a saccade.

This classification gives the user options for viewing and presenting data in a number of ways. One option is to replay the presentation of visual targets with fixations and saccades overlaid in real time. Fixations are usually represented by dots that expand the longer the fixation lasts, while saccades are shown as straight lines moving between fixation dot centers. A gaze-plot uses a similar presentation, but instead of viewing the continuous evolution in real time, the user can choose a given time-frame in which to view and present all fixations and saccades that occurred during this time. To make sense of the temporal order in gaze-plots, fixations are numbered, as can be seen in the gaze-plots presented in the results (Figure 4.1 and 4.2). Another solution is to have saccades be presented as arrows rather than lines (As shown in figure 3.3).

In the analysis of this study, a combination of gaze-plots and the area of interest (AOI) metric called total fixation duration were used as the eye tracking data to identify key patterns and differences between experimental conditions and participants. AOIs are explained in detail below in section 4.2.3. Total fixation duration was chosen as the main metric, as it is a commonly used metric in previous eye tracking studies (Ögren et al., 2016; Lindner et al.,

2017; Lindner et al., 2021), making it easier to compare results (Strohmaier et al., 2020). Total fixation duration measures the sum of time spent fixating on a particular area of interest, and is sometimes also referred to as dwell time. Gaze-plots were used to look for reading patterns during the initial reading of the task text, and to further examine outliers in AOI metrics, such as one participant having a much higher total fixation duration on non-AOI areas than the rest. In addition, gaze-plots were investigated in order to assess the data-quality. On some later test items, fixation positions would skew somewhat off-center from the bottom multiple-choice AOIs, indicating a possible loss in accuracy during the data collection period. Two participants were excluded from the analysis because of large amounts of data loss, with full test items being answered without eye tracking data being recorded. The usage of areas of interest is detailed below in section 4.2.3.

4.2.3 Areas of Interest

Areas of interest (AOI), sometimes also referred to as interest area or regions of interest, is a method used to analyze how participants interact with a particular part of the stimulus (Carter & Luke, 2020). The researcher selects areas of a given stimulus for example by overlaying a rectangle over a text thread. By having AOIs defined it is possible to gather and analyze metrics on how participants interact with the defined areas. An example of AOIs is shown in figure 4.3. These areas can have custom shapes, sizes and positions on the visual stimuli, and subjective choices when creating AOIs can make comparisons between studies difficult (Hessels et al., 2016). The spatial accuracy of collected data determines how small one can make an AOI in order to capture eye movement towards an area. In some reported cases (Hessels et al., 2016), there are statistical differences between AOI creation methods. However, using large AOIs rather than small ones erases these differences, and can be beneficial for other reasons as well, such as reducing noise (Carter & Luke, 2020). The disadvantages of using larger AOIs instead of precise ones are minimized as long as there is clear spatial distinction between relevant areas.

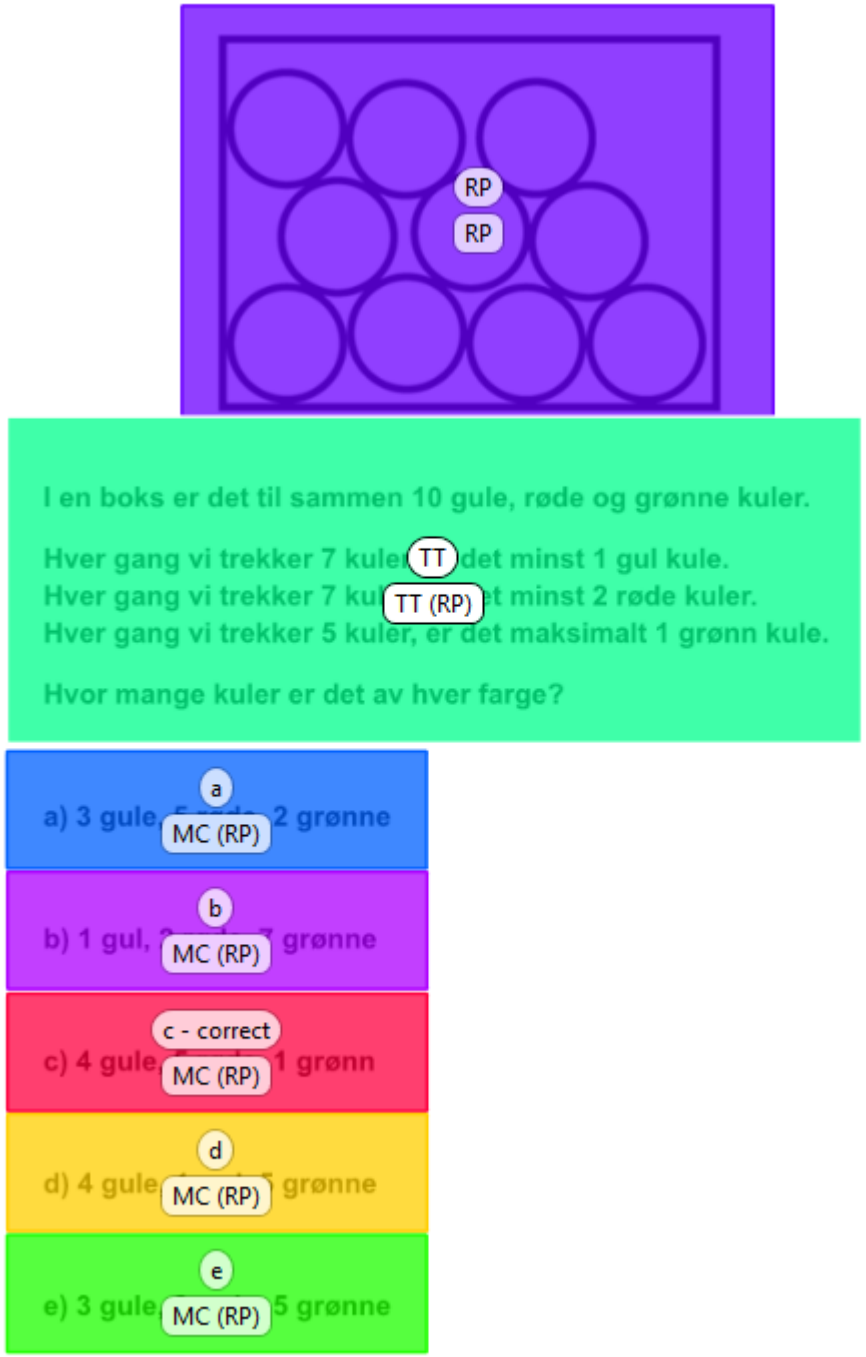


Figure 4.3 Test item 2 with a representational picture (RP), task text (TT) and multiple choice (MC) areas of interest (AOI). The colored boxes are drawn overlays on the stimulus, labeled as explained above. Any fixation within these areas or boxes is registered as a fixation within the AOI, and any saccades in or out is also registered. This makes comparisons between AOIs, participants and tasks convenient. As an example, one can compare how many fixations the task text got with how many the picture got or the time to first fixation for the different multiple choices.

In general, when trying to answer questions about how participants interacted with different specific parts of a stimulus, AOIs should be created (Carter & Luke, 2020). AOIs are particularly useful for quantitative research, as an analysis of individual reading samples, such as by looking at gaze plots, is much more time consuming than comparing the metrics calculated from the software of AOI interactions. Variables that can be gathered from an AOI analysis include, but are not limited to, time to first fixation, first fixation duration, number of visits and total fixation duration.

The AOIs created in this analysis were "hand-drawn" rectangles, meaning that the size and position of the rectangles were chosen by the researcher to fit over the relevant parts of the stimulus. AOIs were made to be larger rather than precisely fitting the relevant areas, as to compensate for possible low accuracies and prevent false negatives. As there was plenty of white space between pictures, task texts and multiple choices, false positives were considered to be less likely than negatives, even when adopting larger AOIs. When multiple choices and task texts were of the same length, identical AOIs were used. The representational and decorative picture versions of tasks often had different sizes of AOIs on their pictures, as the picture size differed from task to task, as well as by picture type.

5. Results

In this chapter, the main results of the experiment are presented. In keeping with the research questions, the results are divided into performance and reading patterns. Under performance, both test scores and time on task (ToT) measures are reported. In the section on reading patterns, an overview of the fixation distribution between AOIs is given, and differences in reading patterns between experimental conditions is explored. Finally, some examples of outliers that are interesting to discuss are presented.

5.1 Performance

In this section, test scores and ToT for each item, as well as effect sizes of test scores and ToT and significance levels are presented. The proportions of correct answers for each test item is shown in table 5.1. Effect size and significance of the test results are shown in table 5.2, and in table 5.3 effect size and significance of ToT is shown. On answering the question about the perceived level of difficulty of the test on a scale from one to ten, the mean answer was 5,6 with a standard deviation of 1,1. A higher number means a higher level of perceived difficulty. Some of the students that answered with a higher number noted that the test items were different from the kind of problem solving they were used to.

Table 5.1 Test score measured in proportion (P) of correct answers for each test item and time on task (ToT) for each item measured in seconds. Mean test score for the entire test was P=0,411.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
P	0,000	0,563	0,400	0,375	0,875	0,188
ToT	57,13	74,25	59,20	84,31	46,25	77,06

A comparison of the test scores was made between all combinations of the three experimental conditions text-only (TO), representational pictures (RP) and decorative pictures (DP), the results of which are shown in table 5.2. Results show that there was a small to medium difference in effect size between DPs and the two other conditions, but these were not significant at p <0,05. In short, students scored slightly higher in the RP and TO conditions than in the DP condition.

Table 5.2 Test scores: Effect size (Cohen's d) and significance from t-test (p-value) of test outcome differences between experimental conditions.

	RP vs TO	DP vs TO	RP vs DP
Cohen’s d	−0,0794	−0,379	0,309
p-value	0,935	0,403	0,465

Using the ToT measure, the same comparisons between conditions were made, as shown in table 5.3. Here, results show a very small, not significant effect size difference between all combinations of conditions. The mean ToT and standard deviation (σ) for each conditions was ToT_{RP}=68,75s and σ_{RP}=39,87s, ToT_{DP}=66,81s and σ_{RP}=39,87s, ToT_{TO}=68,75 and σ_{TO}=31,96s.

Table 5.3 Time on Task: Effect size (Cohen's d) and significance from t-test (p-value) of ToT differences between experimental conditions.

	RP vs TO	DP vs TO	RP vs DP
Cohen’s d	0,138	0,091	0,052
p-value	0,627	0,718	0,505

5.2 Reading patterns

In this section, typical and unique reading patterns are documented, and comparisons between experimental conditions on two main metrics are presented. The first metric is mean total fixation duration on each area of interest, divided into experimental conditions. Figure 5.1 and figure 5.2 show these numbers in two different ways, the first visualizing the summation of the areas, and the second visualizing the standard deviations together with the mean durations. Following this, an exploration of the typical and some unique reading patterns is presented.

First, however, a look at how much attention each of the pictures used in the test is presented. Effect size when comparing fixations on the two picture types was medium at Cohen’s d=0.63, with a barely non-significant p-value of p=0.06 given a significance level of $p<0.05$. When looking at fixations on the pictures from table 5.4, some participants had zero fixations at all on the pictures, and most had total fixation durations of less than a second for three of the representational pictures and five of the decorative pictures.

Table 5.4 Mean total fixation duration on the decorative (DP) and representational (RP) picture-AOIs used in the experiment. Fixation duration on the decorative picture in test item 1 is multitudes larger than the rest of the decorative pictures. In addition, three of the representational pictures have fixation times comparable to the decorative pictures. Test item 1 had the largest mean total fixation duration on the picture for both picture conditions. A possible reason for this could be that test item 1 was an especially difficult task, with none of the participants answering it correctly. The differences in fixation duration shown here are particularly interesting and will be discussed further in chapter 6.

	Test item 1	Test item 2	Test item 3	Test item 4	Test item 5	Test item 6
RP	4,94	2,58	0,40	0,35	2,47	0,43
DP	3,92	0,49	0,22	0,45	0,27	0,70

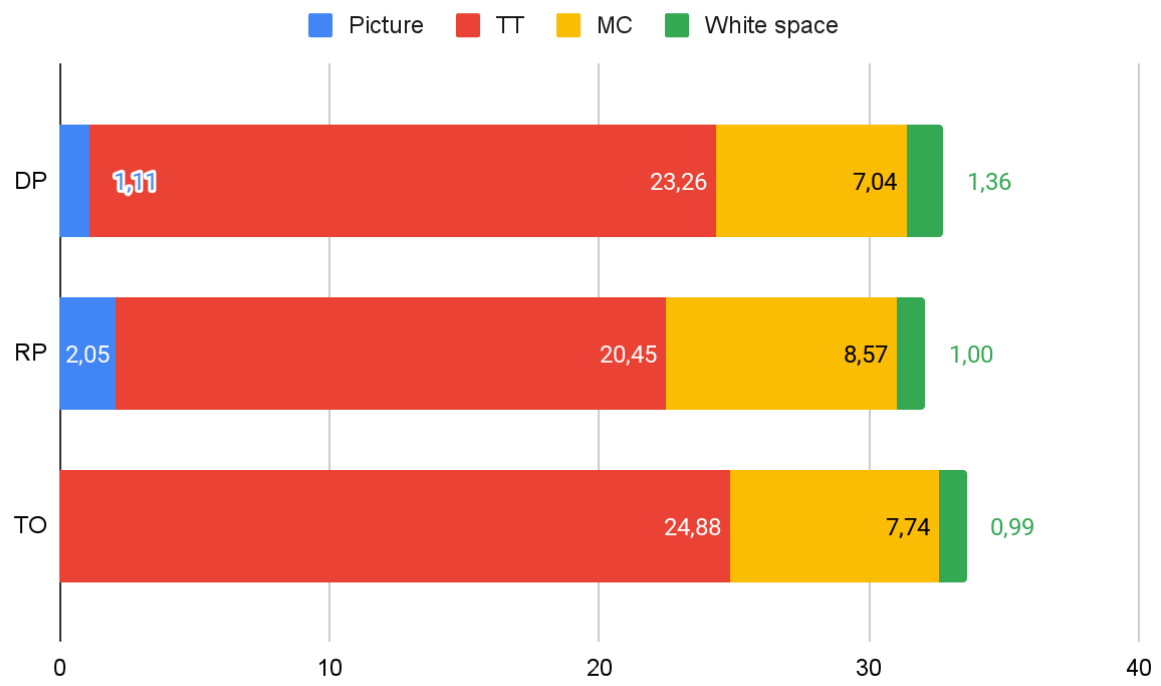


Figure 5.1 Stacked bar graph showing the breakdown of mean total fixation durations for all three experimental conditions: Decorative picture (DP), representational picture (RP) and text-only (TO). Breakdown consists of the defined areas of interest: Task text (TT), pictures (RP or DP), multiple choices (MC) and finally white space, which consists of all fixations outside of any of the defined areas of interest. All measurements are mean time spent fixating on the given area of interest in seconds. The task text is fixated on the least when a representational picture is present, and most when there is no picture to look at. The differences of the sums of total fixation duration on all areas of interest between experimental conditions are very small.

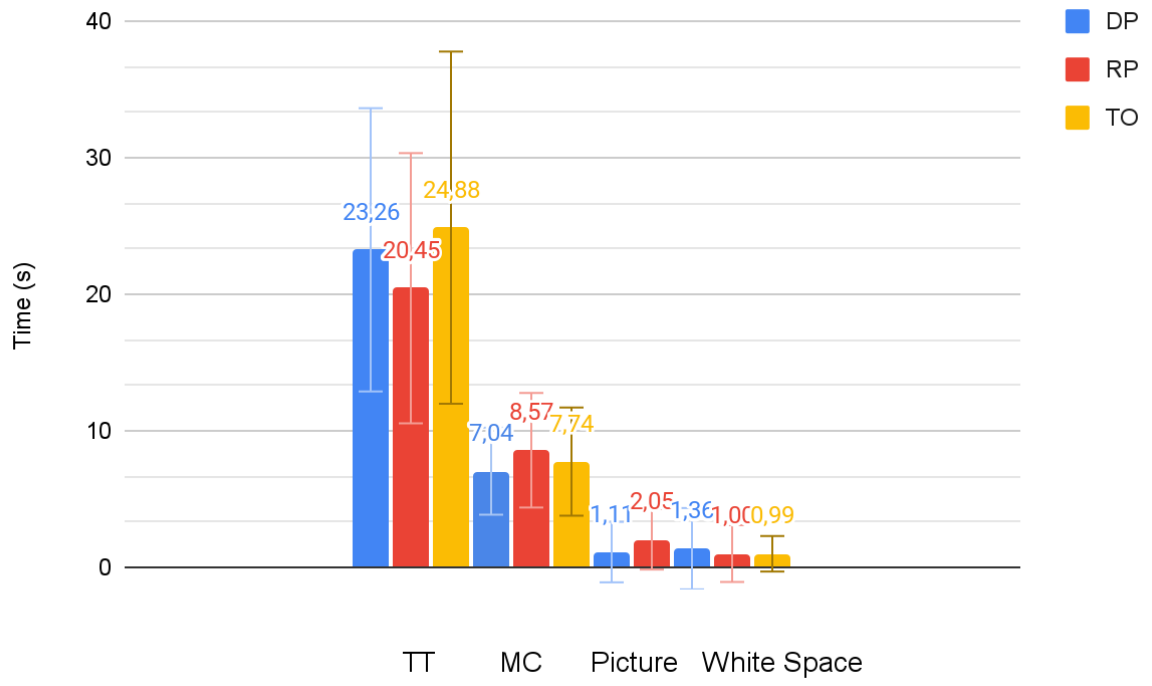


Figure 5.2 Graph showing the mean total fixation duration for each condition on the different areas of interest: Task text (TT), multiple choices (MC), pictures and white space. White space means all fixations that were outside of the defined areas of interest. Created from the same data as in figure 5.1, except for the inclusion of error bars representing standard deviation of the means. The conditions decorative picture (DP), representational picture (RP) and text-only (TO) are represented by blue, red and yellow colors respectively. Numbers on top of the bars are the time in seconds spent fixating on the different areas of interest. On test items with a representational picture, participants spent less time reading the task text and more time looking at the multiple choices when compared to test items with no pictures. On test items with decorative pictures, participants spent less time reading both the task text and the multiple choices than on test items with no pictures. However, differences are small and not significant at $p<0,05$.

With sixteen participants reading six test items each, where one participant skipped a test item, there were a total of 95 readings of the test items. In 68 of these readings, participants fixated on the entire task text before moving their gaze towards the multiple-choice AOIs or, when included, picture AOIs. Broken down by experimental condition, 20 of these were in the RP condition, 21 in the DP condition and 27 in the TO condition. Around half of these 65 readings followed the task text more or less chronologically, with no regressions back to previously read words, making it the most common reading pattern. Figure 5.3 shows an example of a typical initial reading.

In 13 out of 95 readings (5 from RP, 3 from DP and 5 from TO), participants fixated on one or more of the multiple-choices before having read the entire task text. 6 of these 13 were from test item 1. 20 out of 63 readings had participants fixating their gaze on the picture, 11 on RP, 9 on DP, before reading the entire task text. In figure 5.4, an example of a reading pattern where the participant is moving back and forth between the picture and the task text is shown.

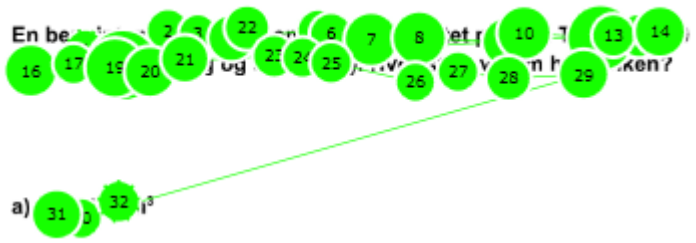


Figure 5.3 Gaze pattern from a participant reading the text only version of test item 6. The placement of fixations and their order (as numbered), shows the initial reading of the entire task text before fixating on the first multiple choice alternative. This was the most common way of reading the test items, with fixations on all of the task text before the multiple choices or pictures were fixated upon. About two thirds of the gaze patterns revealed this type of reading pattern.

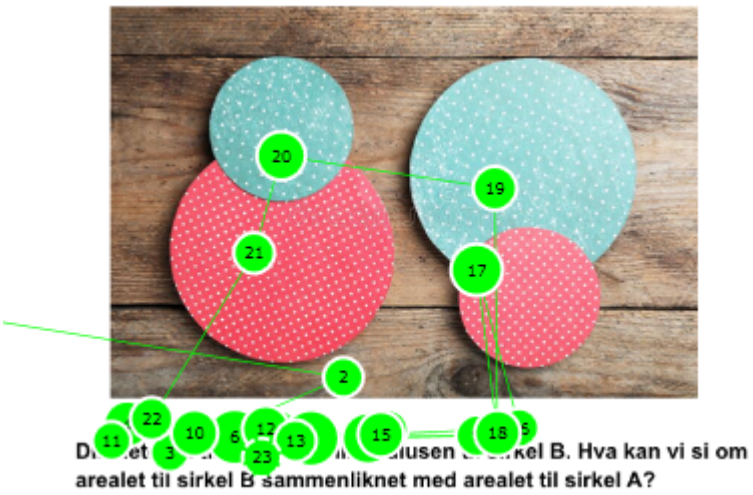


Figure 5.4 Gaze pattern from a participant reading the decorative picture version of test item 1. Here, the participant interrupted reading of the task text to look at the picture. The task text of this test item reads: "The diameter of circle A is equal to the radius of circle B. What can we say about the area of circle B when compared to the area of circle A?" The participant moves their gaze from the text to the picture at the words "circle B", possibly indicating that they are looking for connections between the circles described in the task text and the circles displayed in the picture.

A comparison between the performance measure time on task and the eye tracking measure total fixation duration was made to ascertain the relative amount of time participants spent on reading the test item. The time spent not looking at the screen usually meant participants were instead looking down at their notebook and making calculations by pen and paper. The mean relative time spent looking at the screen was 49,4%, with a standard deviation of 15,0%. Broken down by experimental conditions, the mean time spent looking at the screen relative to time on task was 46,3% for RP, 49,3% for DP and 52,5% for TO.

During exploration of reading patterns, it was discovered that some of the multiple choice answer options lacked fixations completely. In fact, 12 out of 18 test items (6 items x 3 conditions) had some participants skip on one or more of the multiple choices. The most commonly skipped was the bottom option. Gaze plots were investigated in order to rule out decreases in accuracy on later test item recordings explaining this effect, meaning that even if a fixation cluster was slightly off-target for the last answer option AOI, it was still counted as the participant having read the answer option. As an example, see figure 5.5, where fixations are not directly on the last two answer options as a result of poor accuracy.

One participant's gaze pattern had uniquely many fixations outside of any AOIs and even outside of the visual target. This participant was the main contributor to the total fixation duration on white space, and had gaze-patterns that were unlike any of the other participants. Two examples are shown below in figure 5.5. P03 spent 39,8% of their time on task fixating on the screen, which is less than the average participant. They also had a slightly below average test score of 2 out of 6 correct answers (Mean test score=41.1%) and had the longest average time on task (113 seconds compared the mean of 66 seconds).

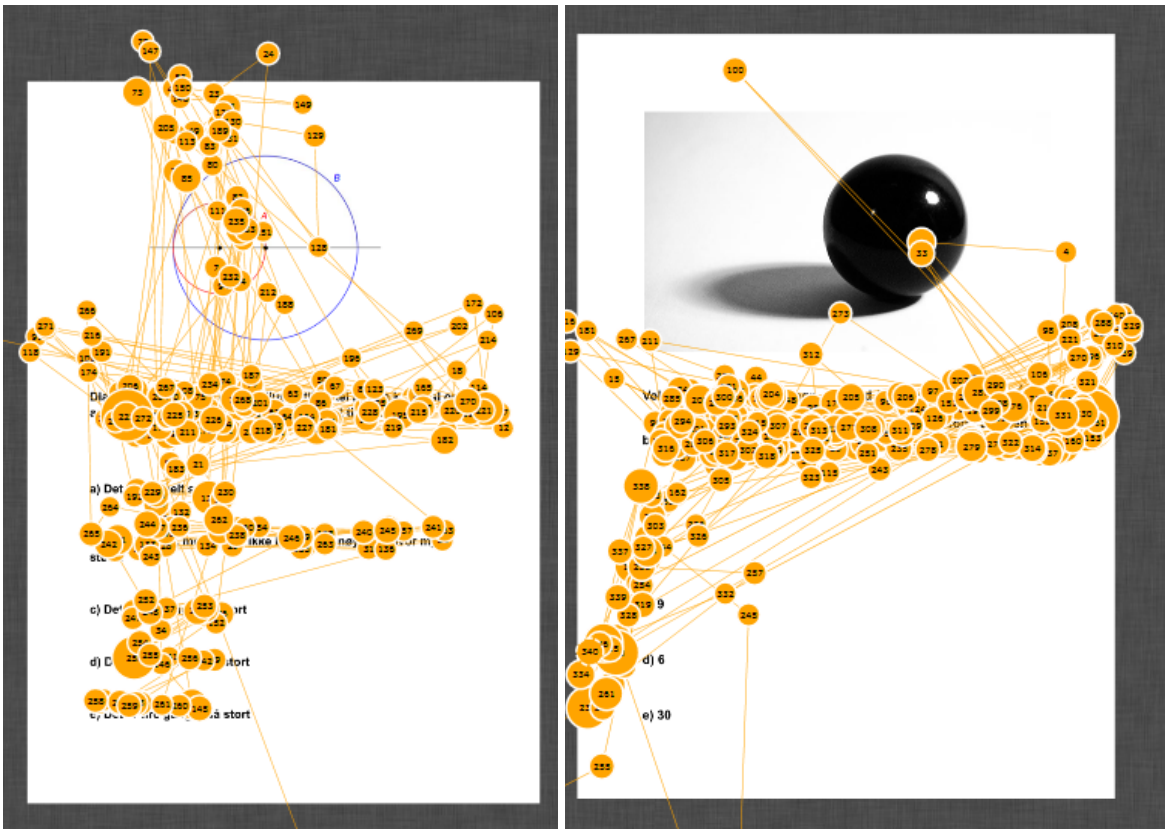


Figure 5.5 P03's gaze-patterns on test item 1 and test item 3 respectively, showing multiple fixations on white space. This pattern of reading was unique to this participant, who exhibited similar fixation patterns on all six test items. As a note, there is a possibility that the accuracy has become poorer in the bottom left of test item 3.

6. Discussion

This discussion is divided into three sections. The first section, 6.1, attempts to answer RQ1 based on the results about performance presented above and knowledge from theory and previous research. In the second section, 6.2, a similar structure is used to answer RQ2, with the addition of a discussion of some important reading pattern observations that are not directly related to the research questions. In the third and final section, 6.3, implications of the research done in this study is discussed, with suggestions for possibilities of future research.

6.1 Performance

The first research question of this study is: *"How does adding pictures to mathematics test items affect students' performance?"* In order to answer this, test scores and time on task was analyzed, as presented in section 5.1. Before discussing the results from this section, an overview of predicted results from theory and previous research is given, for both representational pictures (RPs) and decorative pictures (DPs).

The cognitive learning theories presented in chapter 2 states that material consisting of more information demands more from our cognitive processing systems. This might result in poorer performance in the picture conditions, especially in the time on task measurement, as students need to spend more time processing more information. However, as explained by the dual channel assumption of CTML, displaying information in both pictures and words might lessen these demands, as this kind of a display takes advantage of both channels processing capacity in the working memory (Mayer & Mayer, 2014).

For DPs, CTML predicts that performance might drop somewhat on the basis of the coherence principle, stating that extraneous material, meaning material that is not directly relevant to the learning concept or task, should be eliminated (Rudolph, 2017). Empirical evidence for effects from including DPs in problem solving is not sufficient for a reliable interpretation (Hu et al., 2021). In learning situations, they are shown to induce an increase in alertness, calmness and mood (Lenzner et al., 2013)

When it comes to RPs, the multimedia effect described by CLT and the multimedia principle described by CTML both agree that multimedia material is more efficient for learning than if the material was presented as written text only. If this effect is carried over to assessments, testing and problem-solving is unclear in the current research literature, with some results pointing to RPs having a positive effect on performance (Lindner et al., 2017; Lindner 2020) and a meta-analysis concluding that the use of multimedia in these domains remains yet to be fully explored, but showing a small to medium significant average multimedia effect on response accuracy and no significant multimedia effect on response time (Hu et al., 2021).

The test results of the present experiment revealed a non-significant small to medium negative effect size (Cohen's d=-0,379, p=0,406 at p<0,05) when comparing test scores between DPs and text only. When looking at time on task between DPs and text only, the effect size was only Cohen's d=0.091 and not significant. Given the significance levels of these results, it is likely that differences between experimental conditions are because of natural variance, especially considering previous research on effects of DPs.

Comparing test results between RPs and text only conditions did not reveal much of a difference, with effect sizes being small to tiny, and not significant for both test scores and time on task. This calls into question whether the inclusion of RPs were actually useful for the students. Some research has shown that students only gain an advantage in using RPs on tests when they have practiced using these kinds of pictures while learning (Lindner et al., 2021). A deeper discussion of how the participants of this study interacted with the picture is provided in section 6.2 below.

In summary, the effect of adding pictures on performance is unclear, with results showing an indication that test items with DPs caused a decrease in test score when compared to both RP and TO test items. In terms of time on task, participants spent slightly longer time on both picture conditions than the text only condition, but none of the effect sizes were significant. Larger sample sizes would be needed in order to draw clearer conclusions to this research question.

6.2 Reading patterns

The second research question of this study is: *"How does adding pictures to mathematics test items affect students' reading patterns?"* Answering this question was made possible by using eye tracking, which revealed where on the test items students fixated at each given time. According to the mind-eye assumption (Just & Carpenter, 1980), these fixations strongly correlate to where the attention of the student is at a given time, meaning that how they read and where they fixate could be seen as parallel concepts.

Some limitations to this assumption exist, with a relevant example being from the case study conducted by Schindler and Lilienthal (2019). They found that quick eye movements across non-relevant parts of the stimuli corresponded with strong emotions in the participant, such as excitement, stress or panic. In addition, fixations up and to the left of the task text were made in situations where the participant was thinking about how to solve the problem. However, during the information gathering stages of reading the test items, fixations are much more likely to correspond to the mental processes of participants (Carter & Luke, 2020). Therefore, the initial reading of the test items, meaning the time in which the entire task text was fixated on for the first time, was of particular interest. In addition, the breakdown of which AOIs participants looked at and for how long from figure 5.1 and figure 5.2, as well as table 5.4 showing how much the different pictures were looked at, gives insight into how the reading patterns changed as a function of the experimental conditions.

The eye movement data revealed that about two thirds of the students fixated on all parts of the task text before averting their attention towards the multiple choices or pictures. The other third that did move between areas of interest during the initial reading of the task text, often spent brief moments fixating on either the top multiple choice options or the accompanying picture. This could be a sign that the placement of the multiple choice elements and pictures distracted the students from reading the task text in full. Very likely it could also be a strategy that seeks to integrate relevant information from outside the task text while reading it.

In some gaze-plots, saccades between the picture and corresponding text were quite clearly made by participants during initial reading, as shown in figure 5.4. By corresponding text is meant the part of the text that explains the same concepts that the picture is a representation of. Interestingly, the picture from figure 5.4 was classified as a DP, however this was one of the pictures that students viewed the most. As mentioned in the results, a reason for this outlier might be the difficulty of this particular test item, as none of the participants answered it correctly. One could also argue that the pictures exhibit a degree of being representational, as the task text informs the reader about two overlapping circles, and the picture displayed has overlapping circles in them. In addition, the picture is one of the larger ones used in this experiment, and its saliency is high compared to some of the other DPs (See appendix A for an overview of all test items with pictures). Both these factors are known to affect fixation duration (Carter & Luke, 2020).

This movement between picture and task text has been shown in a previous study to be correlated with a better test score (Ögren et al., 2016). In the same study, fixation duration on the picture alone was not related to test scores. A possible explanation of this pattern by CTML is that these eye movements are evidence for the participant integrating their pictorial mental model with their verbal model, as shown in figure 2.1.

As only three of the RPs, as well as the one outlier DP described above, had mean fixation times larger than a second, it is likely that the other pictures were not as useful to the students, given that they have not looked at them nearly as much, if at all. For the DPs this is as expected given previous research (Lenzner et al., 2013; Lindner, 2020). A relevant note here is that the RP on test item six was simply a representation of a rectangular prism, and as such would not require large viewing times in order to integrate with the task text asking participants to calculate the volume of said prism.

In terms of overall reading patterns and interactions with AOIs, in both picture conditions, students spent slightly less time (Total fixation duration for: RP=31,83 seconds, DP=32,92 seconds, TFD=33,49 seconds) fixating on the test items than in the control condition. However, as described in 6.1 about performance, they spent more time in total trying to solve the test items, the so-called time on task metric, when a picture was present. These differences are rather small and can be because of natural variance. They could also be explained on the basis of cognitive-affective and motivational effects (Hu et al., 2021), i.e. that test items with pictures can be experienced as more motivating, resulting in students using more time trying to solve them without giving up.

Looking deeper at how long participants interacted with the different AOIs, figure 5.1 and 5.2 offers two different visualizations of this breakdown. From figure 5.1, one can see that in the RP condition, students spent less time fixating on the task text and less time fixating on the test item as a whole. This is in keeping with similar previous research (Lindner et al., 2017; Hu et al., 2021), and explained by CTML and CLT as the multimedia effect, which states that processing a picture together with text is easier than processing just text, and therefore can happen more quickly. However, the differences are small when compared to the standard deviations, as shown in figure 5.2, meaning one has to be careful about making definitive conclusions here. For the DP condition, the breakdown in figure 5.1 looks similar to the TO condition. As explained above, the DP of test item 1 was the main contribution to the mean time spent fixating on DPs as a whole, and excluding this item would lower this number. making the bars for DP and TO in figure 5.1 even more similar.

To summarize the discussion of how adding RPs and DPs to test items affected reading patterns in this study, the main result is that RPs changed the reading patterns more than DPs. This is to be expected, as RPs are considered useful when answering a test item, and students have been shown in previous studies to be good at ignoring irrelevant information (Lenzner et al., 2013; Jarodzka et al., 2015). According to the cognitive learning theories, extraneous material should not be presented, according to the coherence principle, but this does not take into account motivational factors (Mayer & Mayer, 2014).

In the following paragraphs of this section, some interesting observations that were made during explorations of reading patterns are presented. These are not directly related to any of the two research questions, but were considered worth mentioning non the less. The first pertains to how the participants interacted with the multiple choice format, and in the second, a unique reading pattern exhibited by one of the participants is discussed, as it could have implications for the design of this type of experiment.

Interestingly, some of the participants did not look at all the multiple choices available on the test items. The most commonly disregarded option was the last on the vertical list. One possible explanation for this observation is that they had already found the option that corresponded with their calculated answer, and thus did not feel the need to read further down the page. A second explanation is that participants actually did process the answer options, either by using non-foveal vision, or without the eye tracker registering it because of data loss or accuracy problems. If neither of these explanations are true, learning better techniques for reading multiple choice questions might be beneficial for the students. Especially for difficult test items, considering all options is a good idea, and it is therefore surprising that the last option was looked at as little as it was.

Given the analysis done of gaze-plots, data loss seems unlikely, but some examples of possible decrease in accuracy throughout the data collection process, i.e. towards the later test items, were found. As an example, see figure 5.5, where the position of fixation clusters over the MCs skew to the left on the lower parts of the page.

When analyzing the gaze-plots of participants, some unique and interesting patterns were discovered, exhibited by specifically one student. In addition to the unique reading pattern, this student spent less time fixating on the screen relative to time on task (39,8%) than the average student (mean=49,4%, σ=15,0%). In terms of performance measures, this participant scored slightly below average with two out of six items answered correctly (Mean test score P=0.411) and had the longest average time on task. As shown in figure 4.4, this student had many and long fixations on parts of the visual stimuli that were entirely lacking in detail.

A likely reason for these fixations is that the student was not gathering information, but rather thinking about something else, such as how to solve the given problem, while still fixating on the screen (Schindler & Lilienthal, 2019). From CTMA (Kirschner et al., 2017), we know that a difference between learning and problem solving, is that for problem solving, the solver has to retrieve from long term memory and apply it to a problem, whereas in learning, a focus on integrating the new knowledge with prior knowledge is made. As mentioned in 3.2 about the Eye-Mind assumption, memory retrieval can be modeled independent of eye movements (Anderson et al., 2004), creating both empirical

evidence and a theoretical explanation for the Eye-Mind assumption not holding when thinking about how to solve a problem.

As the students were all provided with pen and paper in order to scribble down their ideas and calculations during problem solving, it is likely that most other students looked down and away from the observable area of the eye tracker while making mental calculations, as revealed by the average time spent fixating on the screen relative to time on task. Of course, these other students could also have been looking at the computer screen, in the same manner as the previously mentioned student, just that all their fixations happened to coincide with some area of interest independent of their current mental processing. This would in that case create false positives, registered as more fixation time on an area of interest than actual processing time spent on the area. This is an important possible weakness of these kinds of experiments, and solutions to this potential problem will be discussed further in 6.3.

6.3 Implications and directions for future research

As explained in section 2.2.3 on CTMA, understanding the effects of adding pictures to test items is crucial if one is to design multimedia tests that are suitable for assessing learning outcomes in students. The purpose of this study was to be a contribution to this understanding, by exploring the effects of adding representational and decorative pictures to mathematics tests. This section will begin by discussing the implications of the main findings from the study and how it relates to the current research field on multimedia in testing and problem solving. The transferability of the research will be discussed, as well as possible directions for future research based on current knowledge and understanding.

In terms of performance the main finding was that for items with decorative pictures, when compared to both representational and text only items, saw a decrease in test scores, with a small to medium effect size of Cohen's d=-0.379. Although the effect size was not significant, meaning there is a likelihood that the differences are because of natural variance, this was the largest effect size found in performance. Previous research on the effects of decorative pictures in problem solving is sparse and inconclusive (Hu et al., 2021). Therefore, it might be a good idea to be precautionary when deciding whether to include these kinds of pictures on tests for now.

A second main finding worth highlighting was the lack of interactions with multiple of the representational pictures. As discussed above, this could be a result of the students being overloaded by difficult test items, but it could also be because these students are not used to working with representational pictures when solving problems. Research suggests that students who use representational pictures while learning, do better on tests with representational pictures than students who have not practiced integrating these pictures (Lindner et al., 2021). Since all but one of these test items were picked from the database of standardized exam questions in practical high school mathematics (Matte 1P) and the final exam for secondary school mathematics (Matematikk 10. trinn), having students be able to fully utilize these pictures would be an advantage for them on an eventual exam. An important note here is that during data-collection, the participating students were in the middle of their curricula for the year, and by the time exams become relevant, this competence might be higher than it was at the time of this experiment.

The final implication to be mentioned before discussing transferability, has less to do with the specific findings of this experiment, and more to do with the state of the research field that is multimedia assessment. As argued for in chapter 2.2.3, assessments, hereunder testing, exhibit some key differences from learning. Therefore, in order to ensure ecological validity when assessing learning, an integration between what we know about multimedia learning and multimedia assessment, as well as a deeper investigation into the effects of multimedia on assessments is needed (Kirschner et al., 2017).

Moving on to transferability, this research looked at effects of adding pictures to mathematics tests in students following the first year practical mathematics education of high school (Matte 1P). To properly frame this part of the discussion, two disclaimers about the choices made are presented, the first being the choice of mathematics students, and the second being the usage of a computer for the test. Thereafter, a more general discussion on relevant parts of what is known about transferability of the effects of multimedia is made.

As this master is in educational physics, studying physics students was of course considered. However, three main arguments supported the current choice of participants. First, the students selected for this experiment are critically under researched, in a field that is already sparsely developed, with primary school and university level being the most common sampling groups (Hu et al., 2021). Second, given the convenience sampling done to recruit participants, the chosen student group allowed for enough participants to sign up. Third, the interest in investigating decorative pictures seemed most relevant in this sampling group, as they are present on standardized tests, which is not the case for the standardized Norwegian physics exam.

The experiment of this study had students solve the test by reading test items off a computer screen and writing down their answers on a piece of paper. The paper was also available for scribbling and note-taking during the problem solving process. Although there are talks about fully digitizing the standardized written exams in Norway (UDIR, 2023), the current exams, as well as common practice at the school of the participants from this experiment, features tests given on paper. Some aspects of reading patterns might be different when reading off a monitor than when reading paper (Pardede, 2019), and the decrease of ecological validity in solving test items in a different way than students are used to are drawbacks of this. This choice was mostly based around practical limitations of the hardware and software used, as tracking eye movements while students solve tests on paper requires eye tracking glasses. The software also did not easily allow for clickable answer options viewed together with multimedia test items.

Onwards to concerns of transferability, the usage of multimedia in testing is multifaceted and could be affected by many underlying variables (Hu et al., 2021), including but not limited to: Problem difficulty and type, language and reading proficiency, the skill of interpreting multiple representations, types of multimedia used, and in the case of pictures, types of pictures used, both relative and absolute placement and size of different elements on the test items, constructed response formats versus closed response formats such as multiple choice, education level, expertise and age of participants, as well as possible differences resulting from common diagnoses like ADHD or dyslexia. Therefore, controlling for these variables, and investigating them in isolation would be a big step towards obtaining more knowledge in this research field.

Studies exploring some of these variables exist, such as the inclusion of representational pictures in realistic math problems (Dewolf et al., 2015) and in true or false physics problems (Ögren et al., 2016),the effects on performance of representational and decorative pictures on mathematics and science tests (Lindner, 2020), graphical representations in tests (Sass et al., 2017), and effects of representing necessary information visually instead of by text in biochemistry exams at university level (Arneson & Offerdahl, 2023).

Most of the studies analyzed by the meta-analysis done by Hu et al. (2021) on the multimedia effect, were on the topics of mathematics and science. Results show some homogeneity across topics, hinting at transferability between for example mathematics and physics, but the compared studies were few and differed in design, making it hard to say for certain. In addition to classifying studies based on topic, the meta analysis also classified compared studies based on problem difficulty. Below are some considerations on the transferability of both problem difficulty and topic.

For this study the test was in general a difficult test for the students that participated, with the mean test score being P=0.411(from table 5.1). Problem complexity might be a moderating factor on the multimedia effect in problem solving (Hu et al., 2021), which can be explained through CLT, as complex problems tend to overload the cognitive capacity of students. When asked about the perceived difficulty of the test, as compared to tests they usually encounter in their mathematics education, multiple students remarked that it was difficult to compare the two, as these test items were perceived as different from those on usual tests.

There may be differences in the multimedia effect across topics and domains, mainly because what is considered a representational picture in for example chemistry can be vastly different to representational pictures in mathematics (Hu et al., 2021). For the pictures used in this test, the transfer-value over to science and physics should therefore be relatively high, as most of the representational pictures used were schematic or physical representations of the problem at hand. This is similar to the use of figures to describe physical systems in physics. One important thing to note however, is the evidence for students benefiting the most from representational pictures when they are present in both the learning environment and the test (Lindner et al., 2021). Physics students might be more well versed in the usage of representational pictures than the participants of this study, because of the prominence of making figures as a beginning stage of problem solving in physics.

The final topic of this section is on possible directions for future research, based on the current state of multimedia assessment as a research field. The first natural extension to the investigation done in this study, is to more clearly link patterns in performance and reading patterns together, in order to gain insights that can help shape more ecologically valid testing environments that uses multimedia (Kirschner et al., 2017). For example by using a less homogenous participant group, clear differences in both reading patterns and performance can be investigated. This would help understanding of how to design multimedia test items that properly separates test takers based on their skill, as explained by CTMA.

Another possibility is to design an experiment that addresses the innate weakness of fixations on random areas when thinking about how to solve problems causing false

positives in AOI analysis. A suggestion here is to use a triangulation of research methods or a stimulated recall interview after the eye tracking session, as in the case study by Schindler and Lilienthal (2019).

Integrating motivational factors into CTML is a current research goal for multimedia learning theories (Mayer & Mayer, 2014).Including measures of these factors is done in some research (Lindner, 2020; Hu et al., 2021), and should be considered as an option when designing future studies.

7. Educational and personal experience and development

In this short section, I would like to make some personal remarks about my development and experience throughout working on this master thesis. I will begin by describing my experience with using eye tracking as a research method. This was a method that was unfamiliar to me at the start of the school year, and the learning process of getting to know the method has been uniquely challenging and fun. In the second section, some suggestions and reflections around how I could have improved this study and experiment knowing what I know today are presented. Finally, the last section contains a note about how I view the work on this master thesis in relation to my future profession of being a physics, mathematics and science teacher.

7.1 Experience of using eye tracking as a method in educational research

Using eye tracking as a method in educational research is not yet common practice for many researchers, despite its exponential rise in general. As such, I have devoted much time and effort to learning about the method. A large part of writing this thesis has been in an effort to learn and explain about different aspects of eye tracking, as I believe it is a method that can have a real, positive impact on research findings if implemented in a thoughtful and planned out way. This writing was also done in order to make the case for the method's unique benefits in answering my research questions. Using and learning about eye tracking as a method has been a novel, fun and challenging experience that I recommend for others to also experience. Hopefully this thesis can serve as an introductory reference or inspiration for others curious about or wishing to make use of eye tracking in their work.

7.2 Experimental improvements

As I have developed more knowledge about eye tracking methodology and skill in using both the hardware and software, my understanding of how to design a good eye tracking study has increased throughout the semester. This knowledge was not complete when planning for this study began, and so is not reflected in all my choices. In addition, hindsight is a great teacher. Therefore I would like to write something about what I would do differently if I had known what I do now.

The first improvement that I would want to make is to keep the item difficulty more homogenous. With the limited number of participants and test items, because of data collection time constraint, having similar item difficulty across test items would have made analysis easier and possibly more interesting. A dialogue with the students' teachers or a pre-test of potential test items could help filter out overly difficult or easy test items.

In a similar vein, a deeper understanding of other underlying variables and knowledge about previous research would have made it easier to design an experiment that isolates one of these variables and uses appropriate participation criteria and sample sizes. This

would make analysis more targeted and results more insightful. For example, smaller sample sizes would allow for a more detailed gaze analysis. On the other side of the coin, bigger sample sizes would make it easier to ascertain more statistically significant effect sizes between the chosen experimental conditions.

Finally, although I had found some examples of previous research similar to my own while designing this experiment, my overview and understanding of the research field has expanded considerably in the months working on this thesis, and examples of interesting analysis and data collection that could have been good to include in this study includes motivational factors and perceived difficulty per task and a closer analysis of the order of fixations between AOIs in any given time interval.

7.3 Personal and professional development and learning

Throughout the semester working on this master thesis, I have gained deep insights into the themes of this thesis, those being multimedia in learning and testing, eye tracking as a research method and the process of designing a study and writing a research paper. It has given me competence in learning and researching new, unfamiliar topics through scientific literature, as well as key insights and a new awareness about test design and more broadly, the usage of assessments in didactical settings. In general, I would like to think that being able to design valid assessment tasks that suitably measures learning outcomes is a crucial skill when working as a teacher, and I am thankful for the opportunity to have learned so much about these topics throughout the work of this master thesis. Both my choice of research topic and method were quite unique compared to my peers. As mentioned above, I view the research topic as highly relevant to my professional journey as a teacher, that will begin with the conclusion of my education, which is this master thesis.

8. Conclusion

Multimedia learning is a large and widely explored research area (Mayer & Mayer, 2014). Much is also known about assessment in all its forms. The combination of the two however, multimedia assessment, lacks this breadth and depth of empirical evidence and theoretical underpinnings. Some more attention has been given to this research area in the last few years (Kirchsner et al., 2017; Hu et al., 2021), but there is still much we do not know.

Because of this gap in the research literature, this study aimed to investigate how students read test items with representational and decorative pictures, and compared it with test items that had no pictures. To do so, an experiment consisting of students solving a mathematics test was conducted. Specifically, the two research questions of the study were:

RQ1: *"How does adding pictures to mathematics test items affect students' performance?"*

RQ2: *"How does adding pictures to mathematics test items affect students' reading patterns?"*

To answer the question about reading patterns, eye tracking was used to record what part of the test items the students were looking at. Eye movement analysis revealed that the participants mostly ignored looking at decorative pictures, and their presence did not cause much of a change in fixations on the task text or multiple choices of the test items when compared to test items with no picture. However a negative small to medium effect size (Cohen's d=-0.379, $p=0.403$ at $p<0.05$) on test scores was found, although this result was not significant. On the basis of this and previous inconclusive research on effects of decorative pictures (Hu et al., 2021), being precautious in their usage on tests specifically is recommended.

In regards to the effects of adding representational pictures, performance measures showed only small to no effect sizes, these being non-significant. Eye tracking analysis, on the other hand, revealed that only three of the six representational pictures were viewed notably more than the decorative pictures. This could indicate that the students struggle to utilize the pictures to their full potential. As many of the test items were quite difficult to solve for the students (Mean test score P=0.411), it is possible that they became overloaded, resulting in difficulty integrating the picture with the text. For future research, a deeper investigation into the relation between performance and multimedia would be helpful in order to help shape design principles of assessments, as explained by the Cognitive theory of Multimedia Assessment (Kirschner et al., 2017).

References

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis. *Psychological Science*, *15*(4), 225-231.

Angell, C., Bungum, B., Henriksen, E.K., Kolstø, S.D., Persson, J., Renstrøm, R. (2019) *Fysikkdidaktikk* (2nd ed.). Cappelen Damm Akademisk

Arneson, J.B., Offerdahl, E.G. Assessing the Load: Effects of Visual Representation and Task Features on Exam Performance in Undergraduate Molecular Life Sciences. *Res Sci Educ* 53, 319–335 (2023). https://doi.org/10.1007/s11165-022-10057-7

Bergstrom, J. R., & Schall, A. (2014). *Eye tracking in user experience design.* Morgan Kaufmann.

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, *155*, 49-62.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, *81*(1), 1-8.

Cheng, I., Goebel, R., Basu, A., & Safont, L. V. (2010). *Multimedia in education: Adaptive learning and testing*. World Scientific.

Dewolf, T., Van Dooren, W., Hermens, F., & Verschaffel, L. (2015). Do students attend to representational illustrations of non-standard mathematical word problems, and, if so, how helpful are they?. *Instructional Science*, *43*, 147-171.

Ehrhart, T., & Lindner, M. A. (2023). Computer-based multimedia testing: Effects of static and animated representational pictures and text modality. *Contemporary Educational Psychology*, 102151.

Fiedler, S., Schulte-Mecklenbeck, M., Renkewitz, F., & Orquin, J. L. (2019). Increasing reproducibility of eye-tracking studies. In M. Schulte-Mecklenbeck, A. Kühlberger, & J. G. Johnson (Eds.) *A handbook of process tracing methods* (pp. 65–75): Routledge.

Gaspelin, N., & Luck, S. J. (2018). The role of inhibition in avoiding distraction by salient stimuli. *Trends in cognitive sciences*, *22*(1), 79-92.

Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior research methods*, *48*, 1694-1712.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford, UK: Oxford University Press.

Holmqvist, K., Nyström, M., & Mulvey, F. (2012, March). Eye tracker data quality: What it is and how to measure it. In *Proceedings of the symposium on eye tracking research and applications* (pp. 45-52).

Holmqvist, K., Örbom, S. L., Hooge, I. T., Niehorster, D. C., Alexander, R. G., Andersson, R., ... & Hessels, R. S. (2023). Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior research methods*, *55*(1), 364-416.

Hu, L., Chen, G., Li, P., & Huang, J. (2021). Multimedia effect in problem solving: A meta-analysis. *Educational Psychology Review*, 1-31.

Jarodzka, H., Janssen, N., Kirschner, P. A., & Erkens, G. (2015). Avoiding split attention in computer-based testing: Is neglecting additional information facilitative?. *British journal of educational technology*, *46*(4), 803-817.

Jarodzka, H., Holmqvist, K., & Gruber, H. (2017). Eye tracking in Educational Science: Theoretical frameworks and research agendas. *Journal of Eye Movement research,* 10(1):3,1-18 DOI 10.16910/jemr.10.1.3

Johnson-Laird, P. N. (2005). Mental models and thought. *The Cambridge handbook of thinking and reasoning*, 185-208.

Jovanovska, J. (2018). Designing effective multiple-choice questions for assessing learning outcomes. *Infotheca*, *18*(1), 25-42.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Ivančić Valenko, S., Cviljušac, V., Zlatić, S., & Modrić, D. (2019). The impact of physical parameters on the perception of the moving elements in peripheral part of the screen. *Tehnički vjesnik*, *26*(5), 1444-1450.

Kirschner, P. A., Park, B., Malone, S., & Jarodzka, H. (2017). Toward a cognitive theory of multimedia assessment (CTMMA). *Learning, design, and technology: An international compendium of theory, research, practice, and policy*, 1-23.

Lenzner, A., Schnotz, W., & Müller, A. (2013). The role of decorative pictures in learning. *Instructional Science*, *41*, 811-831.

Lindner, M. A., Eitel, A., Thoma, G. B., Dalehefte, I. M., Ihme, J. M., & Köller, O. (2014). Tracking the decision-making process in multiple-choice assessment: Evidence from eye movements. *Applied Cognitive Psychology*, *28*(5), 738-752.

Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017). Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. *Learning and instruction*, *47*, 91-102.

Lindner, M. A. (2020). Representational and decorative pictures in science and mathematics tests: Do they make a difference?. *Learning and Instruction*, *68*, 101345.

Lindner, M. A., Eitel, A., Barenthien, J., & Köller, O. (2021). An integrative study on learning and testing with multimedia: Effects on students' performance and metacognition. *Learning and Instruction, 71*, 101100.

Lindner, M. A., Schult, J., & Mayer, R. E. (2022). A multimedia effect for multiple-choice and constructed-response test items. *Journal of Educational Psychology*, *114*(1), 72.

Mayer, R. E., & Moreno, R. (1998). A cognitive theory of multimedia learning: Implications for design principles. *Journal of educational psychology*, *91*(2), 358-368.

Mayer, R. E. (2002). Multimedia learning. In *Psychology of learning and motivation* (Vol. 41, pp. 85-139). Academic Press.

Mayer, R., & Mayer, R. E. (Eds.). (2014). *The Cambridge handbook of multimedia learning*. (2nd edition) Cambridge university press.

McDonald, R. P. (2013). *Test theory: A unified treatment*. psychology press.

Meng-Jung T., Huei-Tse H., Meng-Lung L., Wan-Yi L., Fang-Ying Y. (2011). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education*

Moon, J. A., Lindner, M. A., Arslan, B., & Keehner, M. (2022). Investigating the Split-Attention Effect in Computer-Based Assessment: Spatial Integration and Interactive Signaling Approaches. *Educational Measurement: Issues and Practice*, *41*(2), 90-117.

NSD. (2020, 14. February). Personvernerklæring. In *NSD. https://www.nsd.no/om-nsd-norsk-senter-for-forskningsdata/personvernerklaering/*

Ögren, M., Nyström, M., & Jarodzka, H. (2017). There's more to the multimedia effect than meets the eye: is seeing pictures believing?. *Instructional Science*, *45*, 263-287.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, *38*(1), 1-4.

Pardede, P. (2019). Print vs Digital Reading Comprehension in EFL. *Journal of English Teaching*, *5*(2), 77-90.

Płużyczka, M. (2018). The first hundred years: A history of eye tracking as a research method. *Applied Linguistics Papers*, (25/4), 101-116.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.

Rayner, K., Smith, T. J., Malcolm, G. L., & Henderson, J. M. (2009). Eye movements and visual encoding during scene perception. *Psychological science*, *20*(1), 6-10.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). Psychology of reading (2nd ed.). New York, NY: Psychology Press.

Rayner, K., Reingold, E. M. (2015). Evidence for direct cognitive control of fixations during reading.
https://doi.org/10.1016/j.cobeha.2014.10.008

Reichle, E. D., & Reingold, E. M. (2013). Neurophysiological constraints on the eye-mind link. *Frontiers in Human Neuroscience*, *7*, 361.

Rodrigues, P. & Rosa, P. J. (2016). Chapter 1 Eye-Tracking as a Research Methodology in Educational Context: A Spanning Framework.

Rudolph, M. (2017). Cognitive theory of multimedia learning. *Journal of Online Higher Education*, *1*(2), 1-10.

Sadek, I., Sidibé, D., & Meriaudeau, F. (2015, March). Automatic discrimination of color retinal images using the bag of words approach. In *Medical Imaging 2015: Computer-Aided Diagnosis* (Vol. 9414, pp. 398-405). SPIE.

Sass, S., Schütte, K., & Lindner, M. A. (2017). Test-takers' eye movements: Effects of integration aids and types of graphical representations. *Computers & Education*, *109*, 85-97.

Schindler, M., & Lilienthal, A. J. (2019). Domain-specific interpretation of eye tracking data: towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics*, *101*, 123-139.

Smith, E. R. (1998). Mental representation. *The handbook of social psychology*, *1*, 391.

Strobel, B., Sass, S., Lindner, M. A., & Köller, O. (2016). Do graph readers prefer the graph type most suited to a given task? Insights from eye tracking. *Journal of Eye Movement Research*, *9*(4), 1-15.

Strobel, B., Lindner, M. A., Saß, S., & Köller, O. (2018). Task-irrelevant data impair processing of graph reading tasks: An eye tracking study. *Learning and Instruction*, *55*, 139-147.

Strobel, B., Grund, S., & Lindner, M. A. (2019). Do seductive details do their damage in the context of graph comprehension? Insights from eye movements. *Applied Cognitive Psychology*, *33*(1), 95-108.

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., & Reiss, K. M. (2020). Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics*, *104*, 147–200.

Sweller, J., van Merrienboer, J.J.G. & Paas, F.G.W.C. (1998) Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10, 251–296.

Sweller, J. (2010) Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educ Psychol Rev* 22, 123–138.

Sweller, J., Ayres, P., Kalyuga, S., Sweller, J., Ayres, P., & Kalyuga, S. (2011). The redundancy effect. *Cognitive load theory*, 141-154.

Sweller, J., van Merriënboer, J.J.G. & Paas, F. (2019) Cognitive Architecture and Instructional Design: 20 Years Later. *Educ Psychol Rev* 31, 261–292.

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006) An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research & Method in Education* 29:2, 185-208

Tobii. (2014). *User's Manual Tobii X2-60 Eye tracker version 1.0.3.* Tobii Technology AB

Utdanningsdirektoratet. (2023, 31. january) *Endringer i eksamen etter nye læreplaner.* UDIR. https://www.udir.no/eksamen-og-prover/eksamen/slik-endrer-vi-eksamen/

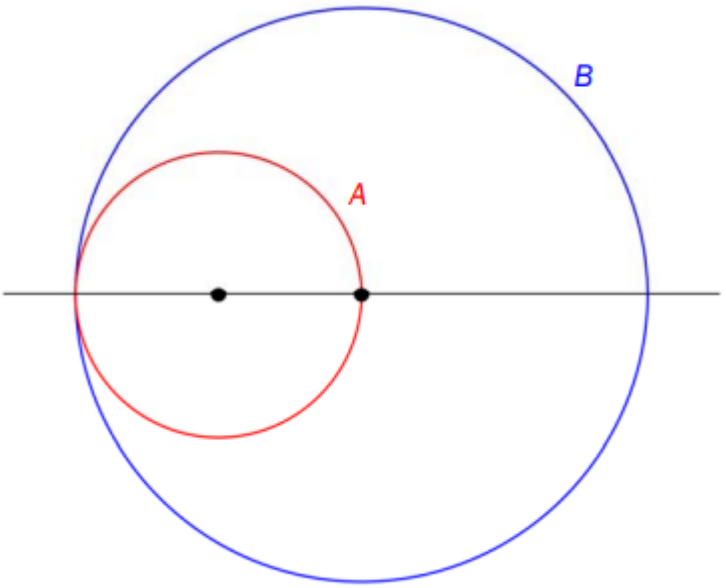
Underwood, G., & Everatt, J. (1992). The role of eye movements in reading: Some limitations of the eye-mind assumption. In *Advances in psychology* (Vol. 88, pp. 111-169). North-Holland.

Wiliam, D. (2011). *Embedded formative assessment*. Solution tree press.

Appendix A. Test items

Each of the below pages are exactly as the test items were presented on the computer screen during the experiment. Ordered in the following way:

- Test item 1 with a representational picture
- Test item 1 with a decorative picture
- Test item 1 without a picture
- Test item 2 with a representational picture
- Test item 2 with a decorative picture
- Test item 2 without a picture
- Test item 3 with a representational picture
- Test item 3 with a decorative picture
- Test item 3 without a picture
- Test item 4 with a representational picture
- Test item 4 with a decorative picture
- Test item 4 without a picture
- Test item 5 with a representational picture
- Test item 5 with a decorative picture
- Test item 5 without a picture
- Test item 6 with a representational picture
- Test item 6 with a decorative picture
- Test item 6 without a picture



Diameteren til sirkel A er lik radiusen til sirkel B. Hva kan vi si om arealet til sirkel B sammenliknet med arealet til sirkel A?

a) Det er dobbelt så stort

b) Det er større, men vi kan ikke bestemme nøyaktig hvor mye større

c) Det er tre ganger så stort

d) Det er halvparten så stort

e) Det er fire ganger så stort

Diameteren til sirkel A er lik radiusen til sirkel B. Hva kan vi si om arealet til sirkel B sammenliknet med arealet til sirkel A?

a) Det er dobbelt så stort

b) Det er større, men vi kan ikke bestemme nøyaktig hvor mye større

c) Det er tre ganger så stort

d) Det er halvparten så stort

e) Det er fire ganger så stort

Diameteren til sirkel A er lik radiusen til sirkel B. Hva kan vi si om arealet til sirkel B sammenliknet med arealet til sirkel A?

a) Det er dobbelt så stort

b) Det er større, men vi kan ikke bestemme nøyaktig hvor mye større

c) Det er tre ganger så stort

d) Det er halvparten så stort

e) Det er fire ganger så stort

En kopp med 220ml cappuccino koster 22 kroner. Hvor mye koster cappuccinoen per liter?

a) 44 kroner

b) 66 kroner

c) 100 kroner

d) 50 kroner

e) 220 kroner

En kopp med 220ml cappuccino koster 22 kroner. Hvor mye koster cappuccinoen per liter?

a) 44 kroner

b) 66 kroner

c) 100 kroner

d) 50 kroner

e) 220 kroner

En kopp med 220ml cappuccino koster 22 kroner. Hvor mye koster cappuccinoen per liter?

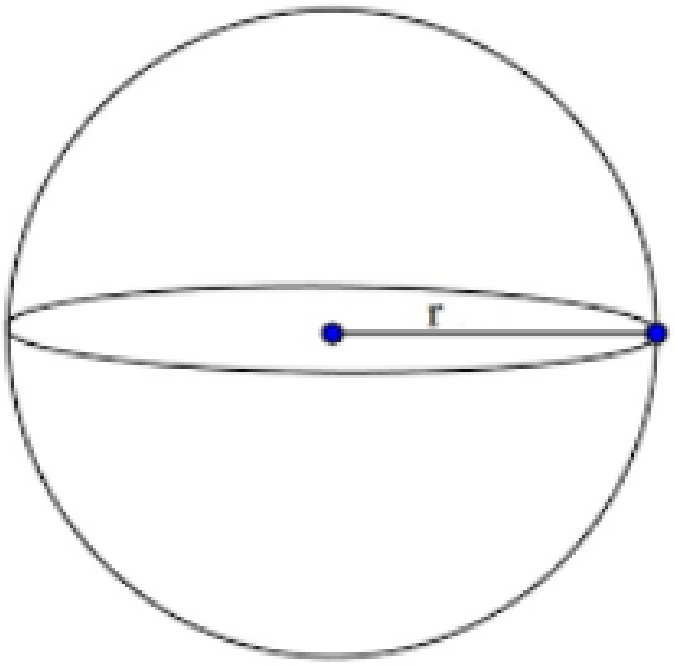
a) 44 kroner

b) 66 kroner

c) 100 kroner

d) 50 kroner

e) 220 kroner



Volumet V til ei kule kan regnes ut ved å bruke formelen $V = \frac{4}{3}\pi r^3$

Hvor mange ganger større blir volumet til kula, dersom radiusen blir tre ganger så stor?

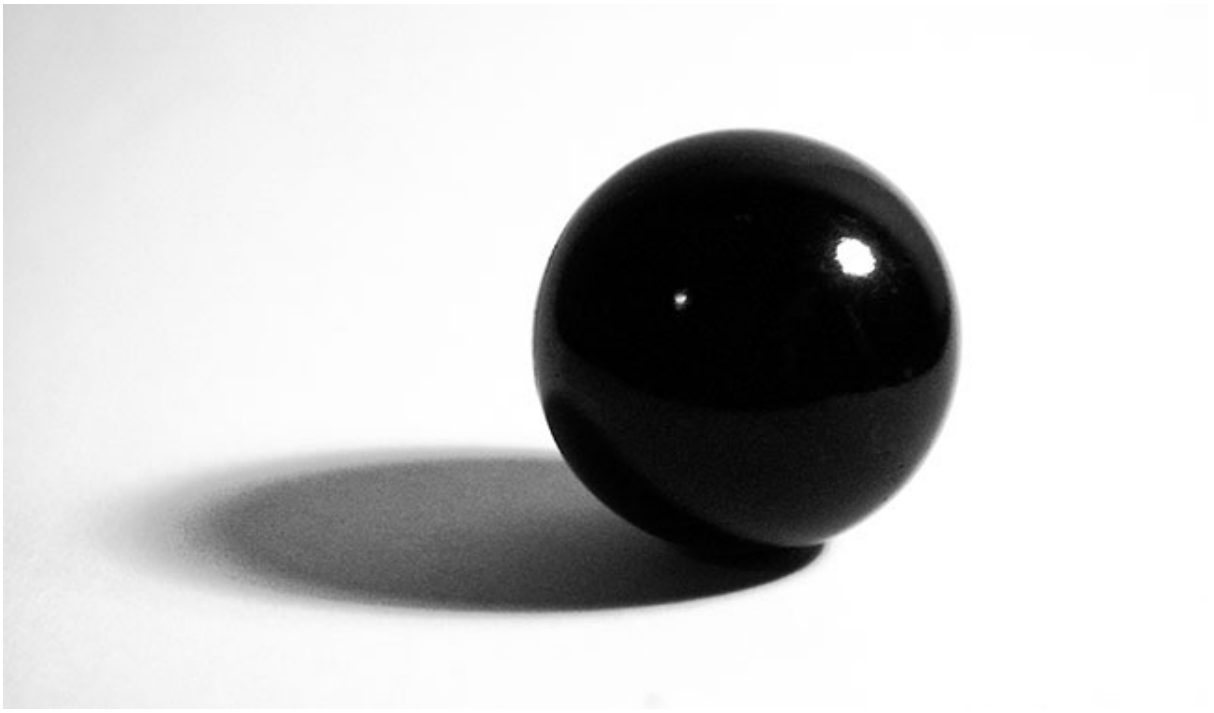
a) 3

b) 27

c) 9

d) 6

e) 30



Volumet V til ei kule kan regnes ut ved å bruke formelen $V = \frac{4}{3}\pi r^3$

Hvor mange ganger større blir volumet til kula, dersom radiusen blir tre ganger så stor?

a) 3

b) 27

c) 9

d) 6

e) 30

Volumet V til ei kule kan regnes ut ved å bruke formelen $V = \frac{4}{3}\pi r^3$

Hvor mange ganger større blir volumet til kula, dersom radiusen blir tre ganger så stor?

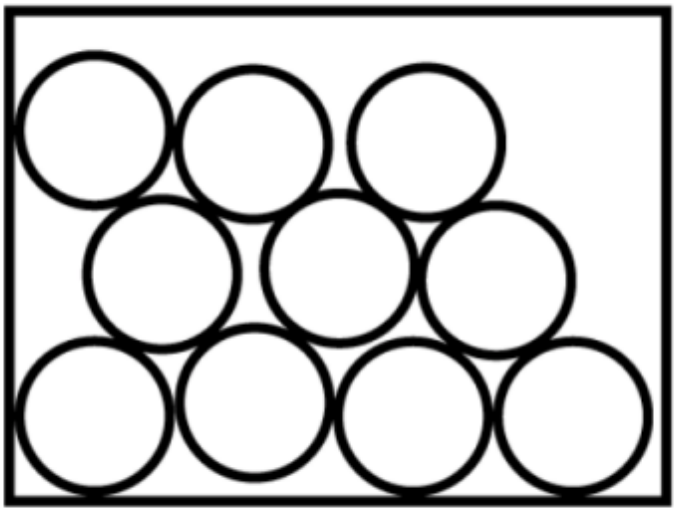
a) 3

b) 27

c) 9

d) 6

e) 30



I en boks er det til sammen 10 gule, røde og grønne kuler.

Hver gang vi trekker 7 kuler, er det minst 1 gul kule.
Hver gang vi trekker 7 kuler, er det minst 2 røde kuler.
Hver gang vi trekker 5 kuler, er det maksimalt 1 grønn kule.

Hvor mange kuler er det av hver farge?

a) 3 gule, 5 røde, 2 grønne

b) 1 gul, 2 røde, 7 grønne

c) 4 gule, 5 røde, 1 grønn

d) 4 gule, 1 rød, 5 grønne

e) 3 gule, 2 røde, 5 grønne

I en boks er det til sammen 10 gule, røde og grønne kuler.

Hver gang vi trekker 7 kuler, er det minst 1 gul kule.
Hver gang vi trekker 7 kuler, er det minst 2 røde kuler.
Hver gang vi trekker 5 kuler, er det maksimalt 1 grønn kule.

Hvor mange kuler er det av hver farge?

a) 3 gule, 5 røde, 2 grønne

b) 1 gul, 2 røde, 7 grønne

c) 4 gule, 5 røde, 1 grønn

d) 4 gule, 1 rød, 5 grønne

e) 3 gule, 2 røde, 5 grønne

I en boks er det til sammen 10 gule, røde og grønne kuler.

Hver gang vi trekker 7 kuler, er det minst 1 gul kule.
Hver gang vi trekker 7 kuler, er det minst 2 røde kuler.
Hver gang vi trekker 5 kuler, er det maksimalt 1 grønn kule.

Hvor mange kuler er det av hver farge?

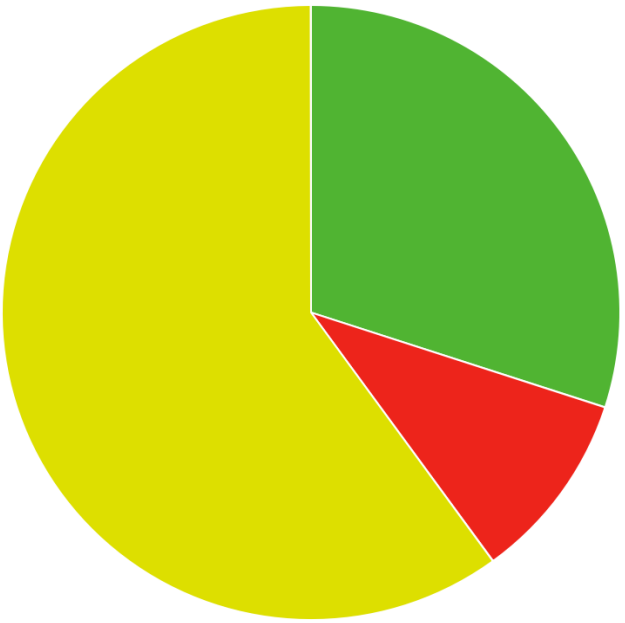
a) 3 gule, 5 røde, 2 grønne

b) 1 gul, 2 røde, 7 grønne

c) 4 gule, 5 røde, 1 grønn

d) 4 gule, 1 rød, 5 grønne

e) 3 gule, 2 røde, 5 grønne



I en boks ligger det gule, røde og grønne kuler.

3/5 av kulene er gule, og 1/10 av kulene er røde. Hvor mange prosent av kulene er grønne?

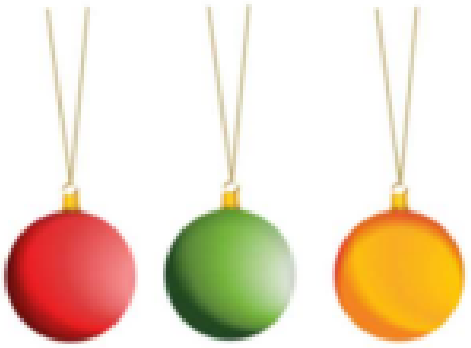
a) 3%

b) 10%

c) 60%

d) 30%

e) 50 %



I en boks ligger det gule, røde og grønne kuler.

3/5 av kulene er gule, og 1/10 av kulene er røde. Hvor mange prosent av kulene er grønne?

a) 3%

b) 10%

c) 60%

d) 30%

e) 50 %

I en boks ligger det gule, røde og grønne kuler.

3/5 av kulene er gule, og 1/10 av kulene er røde. Hvor mange prosent av kulene er grønne?

a) 3%

b) 10%

c) 60%

d) 30%

e) 50 %

En bensintank har form som et rett, firkantet prisme. Tanken er 40 cm bred, 90cm lang og 30 cm høy. Hvor stort volum har tanken?

a) 108000 cm^3

b) 10800 cm^3

c) 180 cm^3

d) 1800 cm^3

e) 18 m^3

En bensintank har form som et rett, firkantet prisme. Tanken er 40 cm bred, 90cm lang og 30 cm høy. Hvor stort volum har tanken?

a) 108000 cm^3

b) 10800 cm^3

c) 180 cm^3

d) 1800 cm^3

e) 18 m^3

En bensintank har form som et rett, firkantet prisme. Tanken er 40 cm bred, 90cm lang og 30 cm høy. Hvor stort volum har tanken?

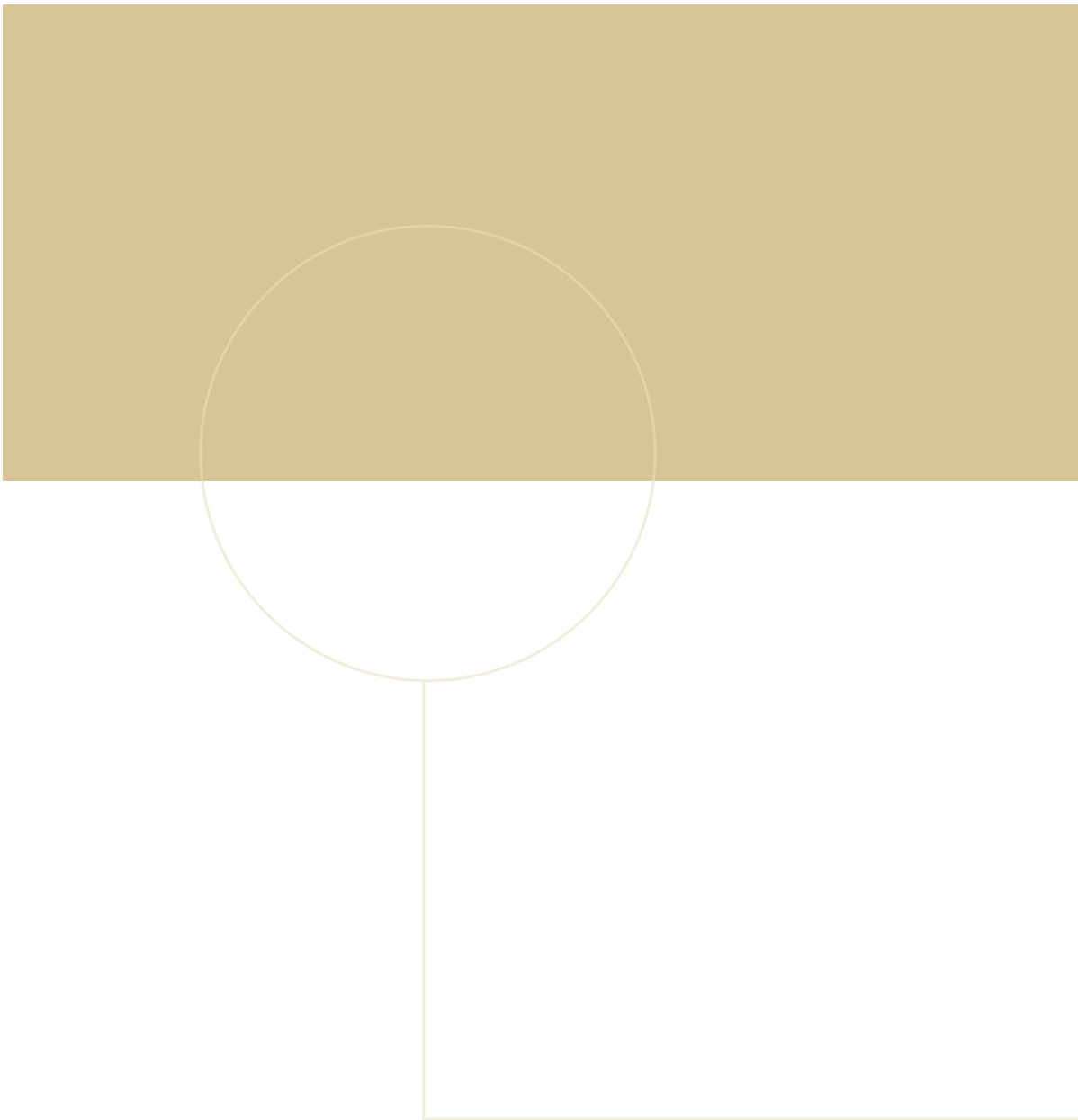
a) 108000 cm^3

b) 10800 cm^3

c) 180 cm^3

d) 1800 cm^3

e) 18 m^3



NTNU

Kunnskap for en bedre verden