Regarding figures in physics exercises: A futile endeavour

Med blicken på figurer i fysikuppgifter: En fruktlös lösningsmetod

Karl Palm

Handledare / Supervisors

Jana Holsanova
Marcus Nyström

KOGM20

2015-06-04

Masteruppsats i kognitionsvetenskap
Avdelningen för kognitionsvetenskap
Filosofiska institutionen
Lunds universitet

Master's thesis (2 years) in Cognitive Science
Lund University Cognitive Science
Department of Philosophy
Lund University

Regarding figures in physics exercises: A futile endeavour

Karl Palm
kalle.palm@gmail.com

A common method for aiding students in solving physics exercises, is to add an image. That images together with text are better for learning than text alone is called the multimedia effect*. It is achieved through integration of textual and pictorial information. Based on this, it can be hypothesized that solving a physics exercise should be easier with an image. Using actual exercises and their corresponding images from Swedish textbooks in physics, a within-subjects experiment was designed. It was then performed by seventy upper secondary school students.*

The study searched for multimedia effects with control images and used eye tracking to examine integration of information. Integration of information was operationalised as attentional shifts between image and text. The results show that no multimedia effect occurred and that attentional shifts in fact correlated negatively with the proportion of correct answers from students. These results question a long-held belief among teachers and educators about the beneficial effect of images in exercises.

1 Introduction

All Swedish textbooks in physics are rich in images, purporting to support students in solving the exercises within thereby learning the material. But do they? Theories of multimedia learning claim that this works through integration of information from text and images. This claim can be measured by means of eye tracking. The aim of the study is to conduct an eye tracking experiment in order to test whether integrative behaviour occurs and whether integration of text and images aid students in their solutions.

The question of whether Swedish physics textbooks and their images are optimal or not is an important one in light of the last years' decline in the results of Swedish students in international comparisons (OECD, 2012; TIMMS, 2007). The importance of science, technology and mathematics in our post-industrial society cannot be overstated, and to be successful in these fields, the didactics of physics and the textbooks of the subject must be as effective as possible.

The following introduction will describe the theoretical framework (multimedia learning) and the empirical method (eye tracking) of the study. This will be followed by some previous studies in the field and the predictions of this study based on that. The hypotheses is that by integrating images and text, students will be better at solving exercises. Thereafter the design of the experiment to study the hypothesis will be discussed in detail, followed by a method description, the results of the study (that the hypothesis were not the case) and finally an interpretation of this.

Multimedia learning

The central theory for this research is *multimedia learning* (MML). What is meant by MML is generally the display of both pictorial and verbal stimuli with the intent of someone learning something (Mayer, 2001). Learning from images might be as old as cave paintings (Tversky, 2001) and depending on your definitions, MML in physics and mathematics can be traced back as far as Euclids *Elementa*, which is rich in figures (Heath, 1956). Also, technical illustrations to aid the understanding of scientific text has been common since the renaissance and Ferguson (1977) claims that they were one of the most important causes of the technological advances between the 16th and 18th centuries. However, serious research in the subject of MML and physics began in the twentieth century (Ferguson, 1977).

Information in printed material comes in two forms: pictorial and verbal, with different advantages. Text is generally considered best for abstract reasoning (Schnotz, 2002), whereas images are better at conveying structures (e.g. Johnson-Laird, 1980; Hegarty, 2011) even at very short presentation times (Eitel, Scheiter, Schüler, Nyström & Holmkvist, 2013). Also, realistic images make people think about concrete objects, whereas schematic visualisations give affordances to abstract thinking and symbolic interpretations (Schwartz, 1995). In current Swedish upper secondary school textbooks, the images that accompany the exercises are in fact of the latter kind, so they should actually aid thinking, according to Schwartz (1995). Sadly, Lowe (1999) found that students generally discover that which is perceptually salient about an image, rather than that which is conceptually relevant.

However, combining the text and images does give rise to MML. Several things are known about MML, such that learning is better with pictures and text than with only text, and that images and relevant words/text should be spatially and temporally close (Mayer, 2009). The two main theories for explaining MML (Eitel et al., 2013) are the *Cognitive Theory of Multimedia Learning* (CTML) and the *Integrative model of Text and Picture Comprehension* (ITPC). The CTML model proposes that an MM message (words and images) is registered by our senses (eyes and ears), processed in our working memory and finally integrated with prior knowledge from our long-term memory. This is based on three assumptions about human cognition. These are that we have a dual channel input[1], a limited cognitive capacity and that processing of information is active. The first assumption is supported by many sources (Mayer, 2001), for example by Baddeley (1992) and that working memory is limited is also well established (Miller, 1956; Baddeley, 1992). The last assumption, that knowledge is constructivistically organised, means that sensory data must

[1] There are different views on whether these two channel should be divided by sensory modality (Baddeley, 1992) or mode of presentation (Paivio, 1991).

be made sense of by the brain and is thus never "pure" sensory information.

These three assumptions together are according to Mayer (2001) sufficient to explain why a multimedia message is superior to a single media message for learning, with respect to both retention and transfer. The argument is that since processing is limited (assumption two) but must be performed to acquire knowledge (assumption three) we can ease processing demands by using our two channels (assumption one).

The second major theory of MML, ITPC, claims that there is a flaw in the CTML (Schnotz, 2002). Where the CTML proposes that the information from each of the dual channels is eventually integrated with prior knowledge into a final mental model, Schnotz (2002) says that this is improbable. The reason for this, according to Schnotz, is that text and images use different sign systems. Where text is descriptive, images are depictive and thus they must also be represented differently in our mind. The final stages in ITPC are propositional representations for text and mental models for images. Although they are not combined, there is interaction due to analogical relations and structural correspondence and this interaction can aid learning and understanding. The view of MM information in ITPC could thus be seen as complementary rather than integrative (Schnotz, 2002).

These theories have been tested in many studies and it is well established that images in conjunction with text give better learning outcomes than just text (Mayer, 2001). This is called *the multimedia principle*. Among the many results of MML research there are a few that are of relevance for this study.

Firstly, several studies show that text is attended to more than images (e.g. Liu & Chuang, 2009; Hyönä, 2010). There are also several studies showing that learners attend to text before images (Hyönä, 2010). However, Ozcelik, Karakus, Kursun and Cagiltay (2009) found that participants first look at images and then at the text. However, it should be noted that their images were very large in comparison to the text and also placed on top of the screen.

Many of the studies on MML, including this one, use eye movements as a measure and accordingly this will be described before going into detail regarding the MML studies.

Eye movements

A way to study thinking is through eye tracking (ET), which means measuring the movements of the eye (Holmqvist et al., 2011). When the head and what we are looking at are in a fixed position, the eye has two main behaviours: fixations and saccades. A fixation is when the eye is still, watching something. Then light enters the eye and gets processed first though the lateral geniculate nucleus, then into the visual cortex and further into the dorsal and ventral streams (Gazzaniga, Ivry, & Mangun, 2009). From there on, the information also reaches other parts of the brain and is in this way processed into what we experience as vision. However, even though our field of vision is large, the area in which we see sharply is very small, so the eye must constantly move to new fixations. This is done through saccades, which are very fast movements of the eye. They are in fact the fastest muscle movements humans can perform (Holmqvist et al., 2011). During saccades, no information hitting the retina is processed in the brain and we are virtually blind in the process. Since none of the information during saccades is processed, we are not aware that we are in fact doing this. Also, there is a functional difference between short and long fixations. A short fixation only assures that what you see is maintained in working memory, unlike long fixations which are the only possibility for deeper processing (Ballard, Hayhoe, Pook, Rao, 1997). This, however, will not be visible in this study as only average fixation durations will be studied.

As we can only take in information during fixations and also must change them constantly, where we are looking can be regarded as what we are thinking about (Rayner, 1978; 1998). This idea about the connection of location of gaze and focus of thought is called the *mind-eye hypothesis* (Just & Carpenter, 1980) and is ubiquitous in modern eye tracking research (Rayner, 1998). However, ET cannot measure whether a participant understands something, only where he or she is looking (Hyönä, 2010). ET is nevertheless a good measure for learning experiments, since it gives an online measure of what is happening with the advantage of not introducing extra tasks for the participants (Hyönä, 2010). Hyönä's solution to accessing more complex aspects such as understanding is to complement ET with offline measures.

An important effect of prior knowledge in ET is that school children who have less of this, seem to have less use of images (Hannus & Hyönä, 1999). Hannus and Hyönä's explanation for this is that integration of information is a cognitively complicated process that these children have less access to. More knowledge does also improve ability to find task-relevant information (Canham & Hegarty, 2010), perhaps leading to ease in using images. However, when spatial information is conveyed in the text, high ability students do not attend to connected images as they manage to visualise a mental model without the image (Hegarty, Carpenter & Just, 1991). Ozcelik, Arslan-Ali and Cagiltay (2010) have showed that with increasing prior knowledge, participants have shorter fixation durations on the task-relevant information. Furthermore, in their review Van Gog and Scheiter (2010) found that visual attention to task-relevant details is strongly correlated with knowledge about the task, but also that these details are fixated faster.

Returning to fixations, there is a risk that participants were staring blankly at the screen or outside when performing mental arithmetic, while having long fixations but not actually thinking about what they are looking at (Epelboim & Suppes, 2001; Knoblich, Ohlsson & Raney, 2001). This might become a problem as not only are fixations longer when we try to understand something, but we also spend more time in general looking at aspects of problems that are harder to understand (Just & Carpenter, 1980).

Studies of multimedia learning with eye tracking

To start with, Mayer (2010) in his review points out that all studies that show an effect on learning by treatment also show an effect on eye movement behaviour. However, the problem with the research being made in the field of MML with ET so far is twofold and concerns ecological validity and the transferability from instructional reading to exercise solving. These two problems will be discussed in detail below. Regarding the first problem, studies with MML mostly rely on experiments where what the participants should learn is a complex system such as a heart (Scheiter & Eitel, 2015), a meteorologic system (Canham & Hegarty, 2010), a pump (Mayer, 2001), or a pulley system (Hegarty & Just, 1993; Eitel et al., 2013). In complex systems there is generally a structure that is essential to understanding the system. These kinds of structures are well known

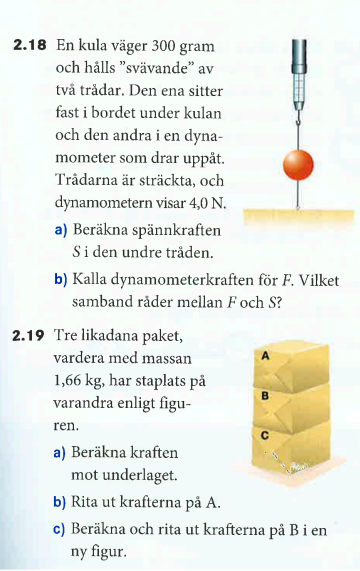

Figure 1: Example of design from physics textbook *Heureka* regarding exercises and their images.

to be much easier to understand via an image (Hegarty, 2011). There is a trade-off relation here between ecological validity and clear experimental results, where most experiments so far have been in the safe realm of high internal validity.

There are, however, some examples of naturalistic eye tracking MML studies (which all happen to be about science) such as Williamson, Hegarty, Deslongchamps, Williamson III, and Shultz (2013), who investigated students' gaze at two chemical models and which of them elicited the most correct answers. Also there is Yang, Chang, Chien, Chien and Tseng (2013) who observed students looking at Power Point presentations in geology. Another example is the study by Hannus and Hyönä (1999) about actually looking at textbooks. The effect sizes of their study are small, but still Hyönä (2010) claimed that not much ET research had been done on textbooks, making their study relevant nevertheless. They made a number of points such as images only aiding the understanding of the text that was related to the image (also Levie and Lentz, 1982) and that children must decide which images on the page to attend to. This relates to the second problem of MML and ET as these are examples of points very relevant to reading text but not to exercises. Also, the placement of text in relation to images is vital for achieving an MM effect, described by the *spatial contiguity principle* (Mayer, 2001). This principle states that closer proximity between text and images increases the MM effect and has been exhibited in several studies (e.g. Holsanova, 2014). However, the text of an exercise is often very short and thus all of it is more or less related to the image. The exercises and their corresponding images in Swedish physics textbooks are designed in a manner that makes it easy to understand which image is connected to which text. This can be done with placement, as can be seen in figure 1, or with thin lines between exercises.

Returning to transferability from instructional reading to exercise solving, the current research in the field mainly revolves around learning in the context of attending to a multimodal media. Exercises are a qualitatively different matter. When reading a text you are trying to grasp the information in front of you in an evident learning situation. It is possible to make a pre-test and post-test evaluation and compare performance/knowledge for different experimental conditions. This can be contrasted to an examination, where no learning is expected to occur, making a pre- and post-test rather meaningless (excluding formative assessment as detailed by Hattie and Timperley (2007)). Exercises however, are somewhere between a learning situation and an examination. That makes it problematic to equate learning from reading instructional text with solving exercises. It is also hard to investigate learning exercises by making pre- and post-tests to the exercises, because the exercises themselves are tests leading to a recursion of evaluation, since the pre- and post-test are also exercises and might be part of the learning.

Very few studies have been made in the field of ET and exercise solving (Nyström & Ögren, 2012) and they found no effect on student performance by adding pictures. This is perplexing since there is tacit knowledge among both teachers and writers of textbooks that images are of great aid. The experiment above was performed with university students, but there have also been experiments with upper secondary school students, but this study did not reveal any clear results concerning the effect of adding images to a problem either (Nyström, 2014).

The current study is inspired by an experiment by Nyström (2014) but with an attempt to circumvent two problems in that experiment. The first problem was the choice of exercises and images. These were chosen by the teacher of the students, without sufficient methodological or theoretical deliberations. One of the five exercises had ceiling effects and another had floor effects. The second problem was that the control condition was no image at all. This makes it hard to compare two conditions where the first has two *areas of interest* (AOI) (exercise and answer alternatives) and the other has three (exercise, image and answer alternatives). Furthermore, the results were hard to evaluate, since no one obviously looked at the field where there was no image.

There is, however, some research on ET and problem solving in general, which might be applicable to exercise solving. Epelboim and Suppes (2001) found that experts utilise images differently than novices. It is open for discussion whether upper secondary school students should be called experts in this context or not. Some research in ET and problem solving found that affecting attention also affects thinking (Grant & Spivey, 2003).

Hypotheses

Two major predictions can be made based on the theories above. If the results of previous MML experiments can be transferred to exercise solving, one can expect that a positive correlation will be found between image condition (meaningful and control) and the proportion of correct answers for each exercise. Furthermore, in accordance with Johnson and Mayer (2012) and Mason, Tornatora and Pluchino (2013), a positive correlation should also be expected between integration of text and image (operationalised as number of attentional shifts, explained below) and again the proportion of correct answers. This is, naturally, only relevant for the images in the experimental condition.

Some minor predictions can also be made based on the similarity of this study to other multimedia learning and/or eye tracking studies (detailed in the previous section). The average fixation duration should be close to 250 ms (see Rayner, 1998). The proportion of fixations of relevant areas should be higher for students with better grades/more prior knowledge (see Van

Gog & Scheiter, 2010). Also, the time spent on images should have a negative correlation with grades (see Hannus & Hyönä, 1999). Furthermore, the average fixation duration should be longer for images than for text (see Underwood, Jebbett, & Roberts, 2004 and Schwonke, Berthold, & Renkl, 2009).

Finally, the type of image (detailed in the section below) may affect the experiment. There are no predictions from theory about this, but it may be a confounding factor and will be analysed for completeness.

Design of stimuli

The stimuli will be adressed in two parts. The first is an argument about the choice of exercises to use and the pilot for selecting these. The second is a reasoning about the design of control images.

To ensure ecological validity, the exercises in the stimuli were chosen from the five currently available textbooks in physics for Swedish upper secondary school according to the current curriculum, Gy11 (Skolverket, 2011)[2]. All exercises with images were examined according to the following two criteria. Firstly, that the exercise could be solved without the image. That is, the image would not contain more information than the text, but only be of a different modality (image versus text). Secondly, that they should be solvable without pen and paper. For many exercises, a slight alteration of the text was made, such as adding that a circuit was in parallel or a number that was necessary to solve the exercise that was only included in the image. This was information that was easily recognised in the exercise image, but made the exercise impossible without it. A few already had multiple choice answers, but for the majority answers were added. Questions with arithmetic had their numbers simplified so they could be solved with mental arithmetic. Numbers were chosen so that they gave no clue about what operation was to be performed.[3] Conceptual questions were given false alternatives in accordance with common misconceptions in physics.

From exercises chosen according to the paragraph above, 66 were selected for a stimulus pilot study. The stimulus pilot study was made in four different versions. Two had the even-numbered exercises with control images and the other two had control images for the odd-numbered exercises. The two pairs were varied among them with some exercises using alternate wordings or difficulty. Out of the 87 exercises in the stimulus pilot, 21 were in two versions of different difficulty or different wording. The main purpose with the stimulus pilot was to find exercises with an appropriate solution rate. This was done to avoid ceiling or floor effects. All questions had five alternatives, so one could expect the proportion of correct answers to be in the interval 0.07 to 0.41 with 95% confidence if simply guessing.

The stimulus pilot was performed on 25 upper secondary school students (females, $N = 17$) from southern Sweden. The grades[4] in basic physics (*Fysik 1*) of the students ($M = 16.6$, $SD = 3.5$) were quite a lot higher than the the average grade of this course ($M = 13.5$, $SD = 4.5$) in Sweden (Henrik Sundström, *Skolverket* (The Swedish National Agency for Education), personal communication, March 17, 2015). Thus, it was more important that the exercises were sufficiently easy, than sufficiently hard. The result can be seen in table 1. Almost all questions were answered (96%). It took the students about an hour to answer all questions, so it can be estimated that each exercise took on average one minute to solve. Seven of the exercises elicited many questions from the students and/or contained ambiguities and were removed. The remaining exercises were finally examined by a PhD student in physics for errors.

The stimulus pilot was performed without eye tracking and with pen and paper instead of on-screen. One could argue that this would affect the data. However, in a review Mayer (1997) finds that there are no differences in learning between different media, as long as the information is the same.

All exercises from the pilot that had a percentage of correct answers between 36% to 80% were included in the experiment. The lower limit was determined so that the number of exercises would be 30. This number was chosen to avoid fatigue and was based on the time it took to solve the problems in the stimulus pilot. It is possible to argue that ceiling effects could be expected from some of the easier exercises. However, since the pilot students had such high grades compared to the mean in Sweden, this can be ignored. For exercises with different versions, the version closest to 50% correct answers was included. No analysis was done on whether images affected solution proportions, as this could lead to selection bias.

Even though the images in the study were all from Swedish textbooks, there were still some differences between them. In their text *Diagrams in the comprehension of scientific texts* Hegarty et al. (1991) define a schema for classifying figures (or as they call them diagrams) in scientific texts. They differentiate between *iconic diagrams*, *schematic diagrams* and *charts and graphs*. The first kind of image is one that refers to something concrete and where structural properties of the referent can be mapped onto the image. Spatial relations in the figure should be isomorphic to the referent's spatial relations. In the study there were 17 figures of this kind and an example can be seen in figure 2. Hegarty et al. (1991) found in their survey of textbooks in physics, biology and psychology that this was the most common type of image. A schematic diagram is defined as one that refers to abstract concepts and that relies on conventions for depicting both components and structure. Examples of this could be a Venn diagram, a flowchart or a circuit diagram. There was a total of 9 schematic figures in the study. Lastly, the diagram type *charts and graphs* are images that present quantitative information such as a map, a coordinate system or most graphs. This last type had only 4 examples in the experiment.

There are many considerations to be made with regards to

Table 1: Difficulty level of exercises in the stimulus pilot. This is operationalised as distribution of exercises according to what proportion of students answered correctly. Thus 18 of the exercises had a proportion of correct answers between 0 and 20%.

Range (%)	Exercises (#)
0-20	18
21-40	30
41-60	19
61-80	12
81-100	1

[2] They are: *Ergo* published by Liber, *Impuls* by Gleerups, *Heureka!* by Natur och kultur and *Orbit* by Zenith läromedel.

[3] If a math exercise has numbers not carefully chosen, students may solve them without actually reading the exercise. Such an example would be using the numbers 7 and 21 together in a text, giving strong affordances to answer with 3.

[4] The Swedish grading system ranges from 0 to 20 and is explained under the section *measures*.

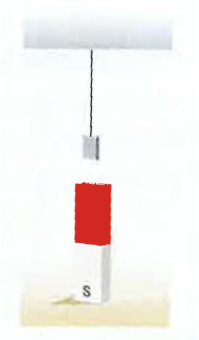

Figure 2: A figure used in the study and an example of a iconic diagram/figure according to the schema of Hegarty et al. (1991).

the control condition. It is problematic for an eye tracking study to have the control without any image, as mentioned previously. The control image must give almost the same stimuli to the participant as the experimental image, except that it does not contain the exercise-relevant information. Since the control image cannot simultaneously give literally the same stimuli and not contain the same information, some sacrifices must be made.

For the image there are several aspects to consider: relation to task, saliency, abstraction level, affordances/anchoring, distraction and brightness. In turn, how much the image is related to the task might influence how well the task is solved. For example, if the task is about a circuit diagram, a photo of a battery might give some inspiration, compared to a photo of an apple. Saliency will also affect the participants. An image of a person or a set of eyes strongly influences people's behaviour (Ernest-Jones, Nettle & Bateson, 2011) and attracts their gaze (Holmqvist et al., 2011), so the control image should not depict people, unless the experimental one also does. Next, as mentioned previously, the level of realism or abstraction might influence how the participants solve the exercise. It is thus unsuitable to contrast a circuit diagram with a photograph in general.

Another aspect of the images is the two phenomena of *affordances* (Norman, 2002) and *anchoring* (Tversky & Kahneman, 1974). Affordances are the manners in which the properties of an object influence decisions, and anchoring refers to the propensity of people, after having been presented with a (meaningless) number, to make estimates based on that number. Both these phenomena could influence the answer of the subject, so that if the control image contains three distinct objects and the the question is numerical, the answers might gravitate towards the number three. A possibility for the control image is to take the original and make some alterations so that it no longer aids the solution. For a circuit diagram exercise about some resistances in series, the control image might show the same diagram, but with the batteries in parallel. This, however, may attract a lot of fixations due to confusion and this distraction might also make participants solve the exercise worse than without an image and is therefore not suitable. Finally, the brightness of the images should optimally be relatively equal for purposes of the eye tracking quality (Holmqvist et al., 2011). To sum up, since even fundamental features like shape and curvature affect attention (Wolfe, & Horowitz, 2004), the control images should ideally contain identical graphical elements compared to the experimental ones, but arranged in a non-meaningful manner.

As mentioned above, there are numerous factors to consider for the control images and many of these are in contradiction to each other or at least incompatible. Since the participants were given a very clear task, it is reasonable to assume that *top-down* factors were more important than *bottom-up* factors. The former are knowledge-driven aspects of attention, such as when we are searching for something, whereas the latter is sensory-driven attention (Corbetta & Shulman, 2002). By this reasoning it was deemed more important that the control images were similar in terms of theme and level of abstraction and detail, rather than being similar in terms of brightness, number of lines, et cetera. Some cases can be seen in figure 3. As an example, the lower pair of images in figure 3 was from an exercise telling the student about two ballons repelling each other asking them about the possible electrical charge of them[5]. Even if the image of the ballons was not necessary to solve the exercise, it remains helpful if the participant has forgotten which means which, of the two words *repel* and *attract*.

Design of experiment

To investigate the hypotheses above, a within-subject experiment was designed. The experiment consisted of a number of physics exercises with the independent variable being whether the exercise was accompanied by an image or not. This was randomised for each exercise. Thus, each participant solved problems with and without images. They chose their answer in multiple choice format and were eye tracked in the process.

The within-subject design was chosen because the effect was estimated to be small and so that the number of participants would not be a problem. Also the difference between students was assumed to be large. If a between-subject design had been used, some students would only have seen control images and this would have led to them eventually ignoring the images (Scheiter & Eitel, 2015). This was also one of the reasons that the order of conditions was randomised.

The allocation of conditions for each exercise was performed as a randomisation for this first half of the participants. The other half then got a balanced version with inverse condi-

[5] They must both be of the same charge.

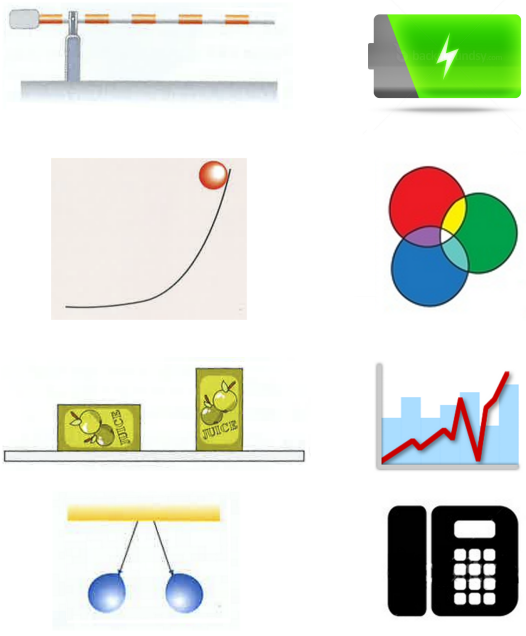

Figure 3: Four examples of original image (left) and their corresponding control image (right).

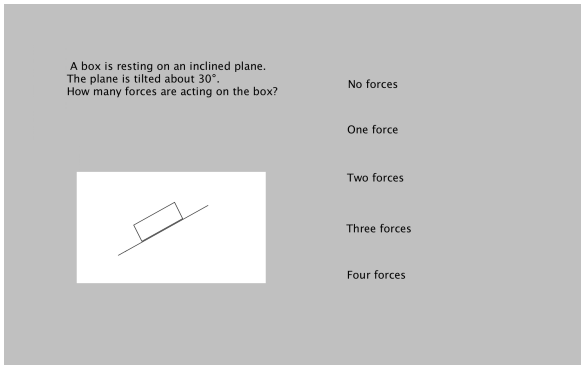

Figure 4: Example screen from pilot study.

tions. This ensured that each exercise received an equal number of participants in each condition. The order in with each exercise was presented was also randomised. A counterbalance could have been made, but was neither deemed necessary due to no expected order effects nor practical due to the large number of exercises.

Since the only eye tracking data that was interesting was the one performed during the process of solving the problems, the experiment was designed to be self-paced. Otherwise, fixations and attentional shifting would have been collected also when the students were finished with each problem and waiting for the timer to end. This could also potentially cause disruptions of the experiment due to restlessness of the participants. A self-paced experiment also gives the participants the possibility to investigate the image thoroughly. Eitel et al. (2013) found that short glances at an image only conveyed the general structure of the image and that more time was required to understand details.

However, Schmidt-Weigand, Kohnert, and Glowalla (2010) claim that the effect of multimedia learning mostly exists due to time constraints and that the slow reading speed is one of the reasons for an MM effect. This is supported by a meta-analysis by Ginns (2005), in which he found that MM effects are diminished or disappear when the MM content is presented self-paced rather than system-paced.

A common method when asking multiple choice questions is to ask the participants to judge their certainty and remove those who are guessing, i.e. rate their certainty to zero (Eitel et al., 2013). However, students are generally bad at estimating their confidence and there are also gender differences in these estimations (e.g. Lundeberg, Fox, & Punccohai, 1994). Two further advantages of not asking for confidence is lessening the cognitive demands and shortening the experiment. The problem of guessing was instead dealt with by using a high lower threshold on how many answered correctly in the stimulus pilot.

To test the experiment, a pilot was performed with five participants and two exercises. The exercises and associated images were not from physics textbooks, but constructed by the author. These were presented with two condition which were randomised so that everyone received one exercise with a meaningful image and one without. An example can be seen in figure 4.

The participants answered by mouse. The experimental images evoked almost twice as many attentional shifts, although not at a level of statistical significance. A 2-sample test for equality of proportions with continuity correction gave $X^2 = 1.87$, $df = 1$, $p = 0.17$. A power analysis in the form of a two-sample comparison of proportions gave that for regular alpha and 0.9 in power, 820 persons would have been needed in each group. However, according to the same statistical tests, the control group spent significantly more time on the images, but to prove this with power according to the one mentioned above, 116 participants would have been needed. A power level of 0.9 is in any case rare to achieve in a behavioural experiment.

These results should not be overinterpreted, as the participants in the pilot were neither physics students nor in the correct age range. Besides, only one person managed to solve one of the two exercises. The pilot merely emphasised that exercises need to be chosen very deliberately and that appropriate participants are necessary.

Measures

To start with, the students were asked about their gender[6] and grade in basic physics. In Sweden grades are given as A, B, C, D or E as passing grades and F as failing. These are translated into the numbers 20 (A), 17.5 (B), 15 (C), 12.5 (D), 10 (E) and 0 (F) when calculating the grade average for the diploma (Skolverket, 2011). These numbers will be used to measure grades. From the experiment, the proportion of correct answers were registered. This will hereafter be referred to as *performance*, as it is a value of how well the participants performed with regards to the exercises.

There are many possible measures to take from the eye-tracker. In a review by Mayer (2010), investigating learning with graphics, all researchers used the time looking at relevant areas as their main measure. That will thus be used in this study as well. Gaze duration on an image can be interpreted as how much the image was processed. Rayner (1998) reports that longer fixation durations are also a sign of more detailed processing (also Ozcelik et al., 2009). Regarding measures, Rayner (1998) furthermore claims that number of fixations, fixation duration, duration time and scan paths are all generally relevant when studying learning. From several ET/MML studies it is known that fixations on images are generally longer than on text (Underwood, Jebbett, & Roberts, 2004; Schwonke, Berthold, & Renkl, 2009).

Another measure to capture processing of images is *integrative saccades*. Holsanova, Holmberg and Holmqvist (2009, p. 1222) define an integrative saccade as "transitions between semantically related pieces of verbal and pictorial information, indicating the process of readers' construction of referential connections between text and illustration". Integrative saccades may be a sign that you are integrating the information (Holsanova, 2014; Johnson & Mayer, 2012; Mason et al., 2013) or that you are having trouble combining the information (Holsanova, 2009). Furthermore, Schwonke et al., (2009) found that this behaviour increases when participants are performing worse. This experiment could test how this actually is correlated with performance.

This phenomenon of alternating between text and image has several names, but is often referred to as *attentional splitting* (de Koning, Tabbers, Rikers, & Paas, 2010; Schmidt-Weigand et al., 2010). However, Hyönä (2010) argues that this terminology is unfortunate as it implies dual processing, while our attention is in fact serial and suggests the term *attentional shifting*, which will be used throughout this text. The reason for not using integrative saccades is that this term spec-

[6] A third option: "Other/I don't want to answer" was added, but selected by no one.

ifies the semantic content that was saccaded between, and the exact point in the content where saccades occur will not be measured. Another, less common, name for this phenomenon is *inter-zone scanning* (Yang et al., 2013).

Concerning the ET data quality, Rayner (1998) in his review reports a mean fixation duration of about 250 ms, which should also be achieved here. Holmqvist, Nyström, and Mulvey (2012) suggest a set of measures to report ET data quality and here a subset of these measures were used: mean and standard deviation of the calibration. More measures were not deemed necessary due to the size of the AOIs and the lack of detailed analysis, such as where singular fixations fell.

Ethical considerations.

The students who took part were all over 18 and thus all of age to consent to participate in the experiment. After the experiment, consent forms were given to all participants and they were informed about all details and had the possibility to ask questions. No participant was estimated to have been harmed. Since they were all studying physics at upper secondary school level, they might even have benefitted from the experiment. This is because they were solving physics exercises and might have learnt something from solving them or at least being submersed in physics for a while.

2 Method

The experiment consisted of 30 physics exercises with two different conditions: an image from the actual textbook exercise or a control image. Apart from the image, each exercise contained a text presenting the problem and five answer alternatives.

Participants and design

In the final experiment 70 subjects (12 females) took part. They were all third year students (ages 18 to 19) of Swedish upper secondary school with normal or corrected-to-normal vision. Furthermore, they were all enrolled in the natural science program (N = 17) or technological program (N = 53) and were thus studying physics or had been studying physics. Two schools in southern Sweden were chosen with a non-probabilistic sampling method of convenience.

For each task, students were given the exercise, an image (control or treatment) and five answer alternatives. After deciding on an answer, they pressed the space button and again were prompted with the answers and only the answers. They then chose what they believed to be correct. Eye tracking was performed during the first of these two phases for each task. An example of a screen shot can be seen in figure 5. It should also be pointed out that no stimuli were placed near the corners, since calibration normally is worse there (Holmqvist et al., 2011).

Since the experiment was performed in a laboratory with a group setup of 25 computers and eye trackers, certain aspects had to be considered. Because the experiment was self-paced, students might get bored when finished and disturb their peers. To avoid this, a silent movie was added at the end of the experiment to entertain them while waiting. Post-test boredom and disturbance did not become an issue.

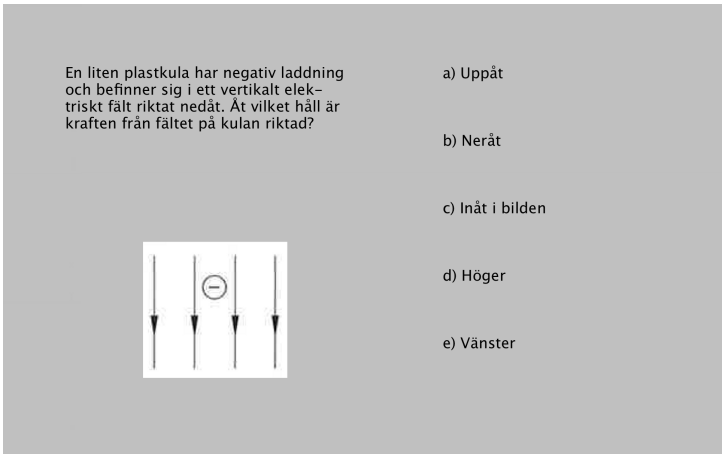

Figure 5: Example of exercise.

Materials and apparatus

The experiment was performed in the so-called *Digital classroom* of the Humanities lab in Lund University, which can be seen in figure 6. The eye-trackers used were SMI RED-m, set to a sampling rate of 120 Hz. The eyes were measured as a binocular average. The presentation of stimuli was made with SMI Experiment Center. Detection of fixations was made with the default settings of BeGaze 3.5. The screen size was 17" with a resolution of 1680×1050 pixels.

As previously mentioned, the stimuli were physics exercises, all containing a text, an image and five answer alternatives. The texts and images were from Swedish textbooks in basic physics for upper secondary school. The answer alternatives were in a few cases from the textbook, but mostly constructed for this experiment.

Procedure

The experiment was performed in five sessions with groups of 20, 11, 12, 10 and 17 participants. The information before the session to the students was, in addition to ethical concerns and consent, a false face validity of the experiment as to be about physics and mental arithmetic. This was also stated as the reason why they were not allowed to use pen and paper. It is important in ET research to give a false face validity that ideally does not direct gaze as instruction is known to affect eye behaviour (Yarbus, 1967). The participants were debriefed after the experiment, and in post-test group interviews only a handful of students identified that the experiment had in fact something to do with images. Finally, the true purpose was revealed and consent once again sought.

Figure 6: The digital classroom of the humanities lab in Lund.

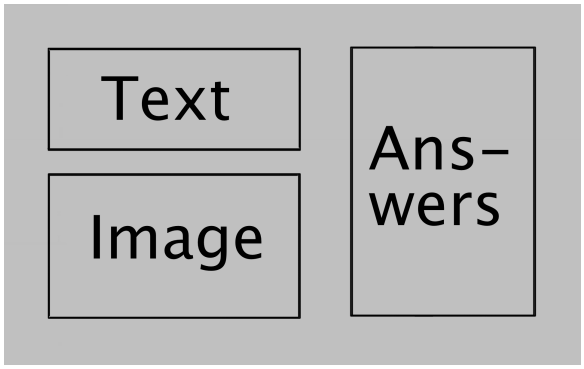

Figure 7: The three areas of interest of each task in the experiment.

The calibration of the eye tracker was performed as a five-point calibration with a four-point validation. Participants were recalibrated if the deviation on either eye was more than 1°. If after three attempts, the calibration still did not reach this threshold, the best calibration was chosen and they still participated. For the sizes of the areas of interest used in the experiment, this level of quality was sufficient.

Data analysis

Some exclusions were made from the data. Data errors occurred for several questions for one participant, resulting in multiple problems and the eventual exclusion of that participant. One participant took a bathroom break in the middle of the experiment and the question that was open during the break was excluded. Another participant left in the middle of the experiment and the last question open was excluded. The remaining data for that participant was still included. Finally, one participant wanted to discuss two questions with the experiment leader during the experiment. These talks were quite brief, but sufficiently disrupting to grant exclusion of these two questions.

BeGaze delivered data in several formats which were merged with some python scripts. The resulting file was imported in R where the statistical analysis was made. For investigations of correlations, a linear model was fit and the value of the slope was calculated. As a measure of certainty, a 95% confidence interval was added as well as p-values. For differences between two groups, t-tests were used and for multiple groups an ANOVA with post hoc-tests was performed. The alpha level used was 0.05. Also, generalized linear mixed model (GLMM) of the binomial family were used. Fitting was performed with the glmer function from the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) and confidence intervals calculated for these with the MCMCglmm package (Hadfield, 2010).

There was also some analysis done on AOI level, such as time spent on images. For this analysis, the AOIs were of the size according to figure 7.

3 Results

After some data analysis measures were calculated. First is some quality measures examining the reliability of the experiment. Thereafter comes the experimental results, detailing the two main findings. The first one is that giving a meaningful image to participants has no effect on proportion of correct answers and the second is that more attentional shifts in participants being correlated with a less degree of correct answers.

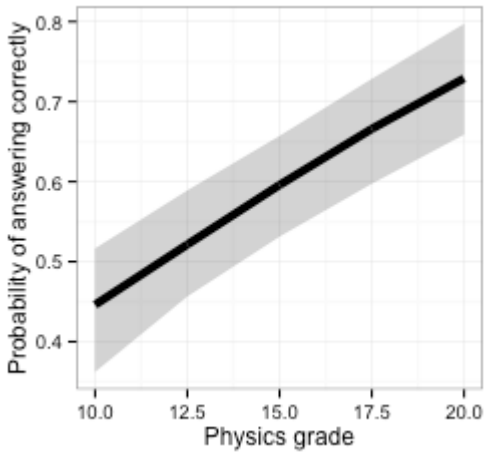

Figure 8: Plot of generalised linear model of probability of correct answer as a function of grade in basic physics for all participants (N = 69). The grey area denotes confidence interval of the line.

Quality measures

The mean calibration error of the participants was for both x- and y-axis 0.5° (SD = 0.2°). The mean level of tracking was 92% (SD = 11%, median 95%). The mean grade of the students in basic physics were 13.8 (SD = 3.1). For ET measures, the average fixation duration for all participants and all exercises was 241 ms (SD = 53 ms). Also, the average fixation duration for images (M = 262 ms, SD = 113 ms) compared to text (M = 228 ms, SD = 52 ms) was greater in the experimental condtition. According to a one-sample t-test, this difference was significant ($t(1014)$=74, $p < .001$). In contrast, the control group had shorter average fixation durations on images (M = 188 ms, SD = 137 ms) than those on text (M = 238 ms, SD = 58 ms). This difference was also significant ($t(1026)$=43, $p < .001$).

The performance of the students in the test compared to their grades in basic physics can be seen in figure 8. A binomial GLMM with random intercepts for each participant and each question suggests that the effect of physics grade on performance is highly significant (likelihood ratio test: L = 26.3, df = 1, $p < .001$). The coefficient on grade in the model was .13, meaning that each grade increase (2.5 points) is equal to increased odds of giving a correct answer by 39%[7].

Experimental measures

The general difference in performance between the experimental image and the control image can be seen in figure 9. The difference in proportions of correct answers for the two groups was 1.6% (95% CI [−4, 7]). A one sample t-test gave no significant difference between the conditions ($t(29) = 0.60$, p = .54). A binomial GLMM with random intercepts for each participant and each question, with condition and image type added to the model still gave no significant differences (likelihood ratio test for condition: L = 0.6, df = 1, $p > .4$ and image type: L = 0.06, df = 2, $p > .8$).

A function of proportion of correct answers in relation to attentional shifts can be seen in figure 10 (for experimental images only). This proportion significantly decreased with a percentage of 10 for each doubling of attentional shifts (95% CI [0.5, 19], p = 0.039). This can also be averaged over time,

[7] Calculated as $e^{0.13 \star 2.5} = 1.39$.

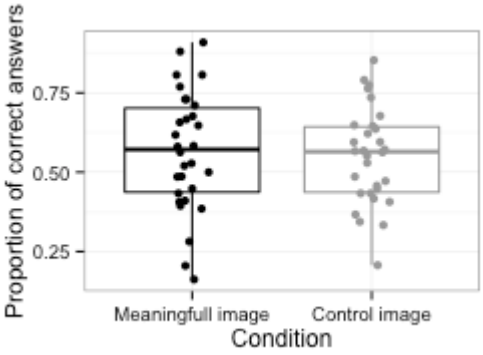

Figure 9: Boxplot of proportion of correct answers by condition for all exercises ($N = 30$). A slight jitter of the points was added in the x direction for visibility.

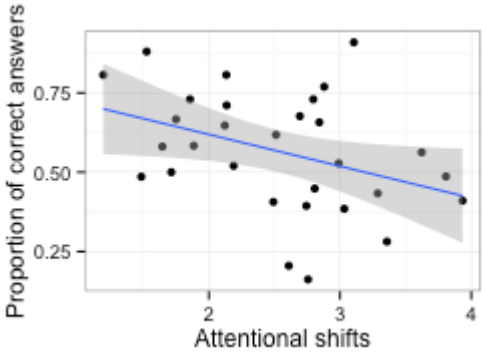

Figure 10: Scatterplot of proportion of correct answers as a function of number attentional shifts for all exercises with the experimental condition ($N = 30$). Attentional shifts were $\log_2$ transformed. A line has been fitted for the data, and the grey area denotes standard error of the line.

resulting in a decreased effect with less significance. Looking at the proportion of correct answers in relation to attentional shifts for the control condition, the effect is similar but less (8 percentage points) and not significant (95% CI [−0.4, 17], $p > .05$). For attentional shifts a one-way between subjects ANOVA [$F(2, 27) = 4.08$, $p = .03$] revealed a significant effect of image type. Post hoc comparisons using the Tukey HSD test indicated a significant difference only between iconic and schematic figures at $p = .02$. There was no significant difference between image kinds in average fixation duration on images.

There was also a correlation between grade and proportion of fixations on any of the relevant AOIs. For the experimental condition this was text, image and answers and for the control only text and answers. This proportion increased by 1.1 percentage points for each increase in grade (95% CI [0.6, 1.5], $p < .001$). The proportion of time spent on the image in the experimental condition however, decreased with increasing grade with 0.02 percentage points (95% CI [0.01, 0.03], $p < .001$).

A Poisson GLMM model was fit to explore the effect of total time spent on each exercise and the number of attentional shifts. The number of attentional shifts was consistently higher in the experimental condition compared to control condition ($L = 1249$, $df = 1$, $p < .001$), indicating that meaningful images were indeed integrated with the text. Furthermore, the number of attentional shifts increased as a function of total time in both conditions, but more rapidly when the image was relevant to the task. This interaction between condition and looking time was highly significant ($L = 39$, $df = 1$, $p < .001$; see figure 11).

Regarding speed versus accuracy for solving the exercises, there are two possible analyses: on exercise level or on participant level. For exercises each extra second spent solving it meant the proportion of correct answers changed by −0.6 percentage points for both experimental and control condition. For the experimental condition, this was not significant (95% CI [−1.3, 0.04], $p = .06$), but it was for control condition (95% CI [−0.011, 0,0011], $p = .02$). On the participant level, an extra second spent on the exercise in the experimental condition corresponded to a 0.4 percentage point increase in proportion of correct answers (95% CI [0.03, 0.7], $p = .03$). For the control condition, the increase was of similar size, 0.3, but not significant (95% CI [−0.09, 0.6], $p = .13$). This can be seen in figure 12.

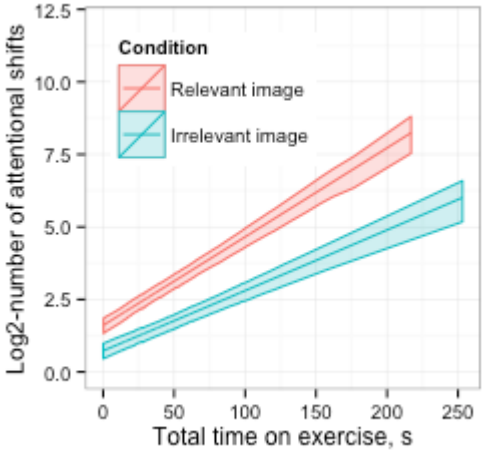

Figure 11: Plot of a generalised linear model of number of attentional shifts as a function of total time spent on the exercise. The grey area denotes credibility interval of the line.

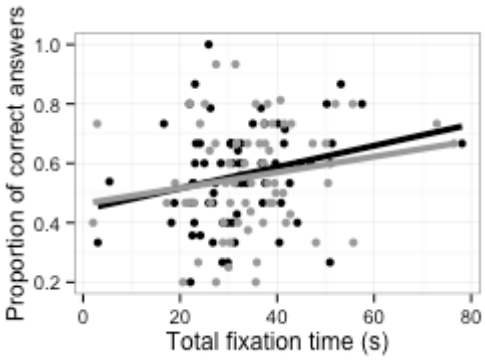

Figure 12: Scatterplot of proportion of correct answers for all participants as a function of average total fixation time for both conditions ($N = 69 \cdot 2$). The black dots and line represents the experimental condition and the grey ones represent the control.

4 Discussion

The multimedia effect is a well-researched phenomenon which implies that images in conjunction with text increase learning compared to only text. It has mainly been researched in the context of reading, but tacit knowledge among teachers is that it should work for exercise solving as well. The assumption that images aid the solving of exercises was operationalised as two hypotheses. First, that a positive correlation will be found between image condition and the proportion of correct answers for each exercise. Second, that a positive correlation should also be expected between integration of text and image and the proportion of correct answers. These were tested in an experiment with upper secondary school students who had taken basic physics. They answered multiple choice questions from textbooks in basic physics with either a meaningful image or a control image.

Before looking at the results, the validity and reliability of the experiment will be discussed. Calibration can be considered successful regarding deviation as described in the results section. The level of tracking was also good and the high standard deviation of tracking level was mainly due to one participant. The average fixation duration is close to what Rayner (1998) reports in his review, which can be seen as an indication that the ET data is typical. As expected from the study by Hannus and Hyönä (1999), students with less prior knowledge (operationalised as grade in basic physics) also used the images less. This is, however, somewhat in conflict with the results by Hegarty et al. (1991) claiming that high ability students (again operationalised as grade in basic physics) do not use images if they are informative but redundant in relation to the text. Perhaps this operationalisation does not hold or the extent of using an image might not be proportional to how much it is fixated. Finally, regarding prior knowledge, Vag Gog and Scheiter's (2010) results that better prior knowledge leads to more looking at task-relevant information was confirmed. Since the task-relevant information in this experiment was more or less all the information (in the experimental condition), it is an open question what students with less prior knowledge were thinking when looking at the grey space or outside the screen.

From the results it was also evident that the higher the grade, the better the students performed in the experiment. This is a good indication of validity, with respect to the experiment's resemblance to actual physics exercises in school. It is also an indicator that knowledge of physics is more important than chance for determining performance. The physics grades of the students were also close to the national average, which should make the results generalisable in that respect. As expected from previous studies (Underwood et al., 2004; Schwonke et al., 2009), average fixation duration was longer for images than text. The general explanation is that it is less cognitively demanding to identify words, compared to images. However, this was only in the experimental condition. That this was not the case in the control condition is presumably due to participants quickly disregarding the image as meaningless and then not spending any long fixations on it.

Regarding the main hypotheses, the general difference in performance between the experimental image and the control image was not significant. Mayer (2010) claims that few or no results from treatment are seen in MML studies using multiple choice and that this should be replaced by open-ended question. However, this limits the number of questions you can ask (for practical reasons) and there are also several counter-examples to this, such as Williamson et al. (2013), Tsai, Hou, Lai and Yang (2012) and Lindner, Eitel, Thoma, Dalehfte, Ihme, and Köller (2014). That images did not increase the proportion of correct answers is in line with Nyström and Ögren (2012), yet still puzzling in light of multimedia theories and tacit knowledge of teachers. Both Ginns (2010) and Schmidt-Weigand et al. (2010) claim the MM effects disappear when experiments are self-paced. If this is the case, and the explanation for lack of effect in this study, one can ask whether there is any point in multimedia learning at all. Solving problems is seldom done under the short time constraints, as in for example Eitel et al. (2013), either in school or in real life.

A question regarding the lack of multimedia effect is whether the students understood that there was no information to be gained from the control images. Since the average fixation duration was considerably lower on images in the control condition, it is reasonable to assume that the participants understood this and thought less about them. As a parenthesis, a few of the images did exhibit strong multimedia effects, but no analysis was made on individual exercise level. This could be a subject for a future study.

One reason for the lack of effect might be that despite using naturalistic stimuli, the study deviated from a normal classroom experience in certain aspects. Students normally read from a book and not a screen, but as Mayer (1997) concludes this should not have an effect. However, students do normally write down their answers by pen and, more importantly, are allowed to take notes. Perhaps there is something in the physical act of writing a solution or drawing your own image that interacts with the exercise and its image in a manner of embodied and situated cognition (e.g. Kirsh & Maglio, 1994). A future study should examine how students utilise their self-drawn figures and images and perhaps this is where the multimedia effect can be found.

Another explanation is that there might be a speed-accuracy trade-off. The results do show that this effect exists for the analysis level of exercises, but that only means that easy exercises are solved faster. However, analysis on participant level demonstrates that spending more time on an exercise did lead to increased solution rates, but only for the experimental condition. This could be interpreted as images being helpful in the context of longer thought processes. That is, when students take the time to really try to figure out a hard problem, an image might be helpful. However, this is slightly speculative and requires more research. That research should investigate the relationship between multimedia effects and variations of timed versus self-paced setups. If the multimedia effect is only an effect of limited time, then it is not very interesting to investigate. Either teachers and multimedia researchers are wrong or there might be some other, yet unknown effect of images.

The hypothesis based on the two experimental studies of attentional shifting and learning (Johnson & Mayer, 2012; Mason et al., 2013) was that a positive correlation should be found between attentional shifts and the proportion of correct answers for all exercises. However, the results of the current study show the opposite. The correlation was negative and there was no correlation at all for the control condition. Both in the experiment by Mason et al. (2013) and the one by Johnson and Mayer (2012), attentional shifting was related to learning performance and not to problem solving. Thus these conflicting results can be seen as an argument that the results of MML with regards to learning cannot be generalised to problem solv-

ing.

Interestingly, the exercises in the experimental condition elicited more attentional shifts per time, than in the control condition. Evidently, a relevant image does make students look back and forth to a greater extent. It is possible to argue that the number of attentional shifts should be averaged over time, but this is probably not a good idea. This would give the rate of attentional shifts and there are no predictions from the literature how the speed of shifting would affect performance. Whether humans oscillate quickly or slowly between information in different forms and the consequences of this behaviour for learning and/or problem solving is not within the scope of this thesis, but could be a suggestion for a new topic of investigation.

Another interpretation could be that increasing numbers of attentional shifts are in fact not a sign of integration of information, but rather, as the conjecture by Holsanova (2009) states, a sign of *trouble* integrating the pictorial and verbal information. If faced with a problem one cannot solve, it is reasonable that the gaze wanders back and forth between the information that is available. This was also supported by Schwonke et al. (2009) as stated in the introduction.

One consideration that might support the idea that high attentional shifting is in fact a sign of trouble with integrating information is the fact that attentional shifts were significantly higher for schematic figures compared to iconic ones. This might be because schematic figures require an interpretation what they mean before they can be utilised. Also, that average fixation duration on images had no significant differences based on image type, supports that the differences between these image types is in difficulty of integration with text, rather than difficulty merely to understand in general.

The level of attentional shifting in the control condition is presumably affected by the choice of control images. These could have been improved by adopting them more to the experimental images in terms of bottom-up features, and not only top-down features as in the current study.

If images in physics exercises do not aid in solving them, what are the consequences of that? Moreover, the images used here were specifically designed by the authors of the textbooks to be helpful. There is an ongoing discussion about whether or not teachers should train students in how to use figures and other representations (Ainsworth, 2006). Schwonke et al. (2009) point out that knowledge about using images in one domain does not follow automatically from general knowledge in that domain. Hence Schwonke et al. (2009) are of the opinion that this should be taught. This sentiment is also shared by Hannus and Hyönä (1999), who found that when students are told to look at images, they do so to an larger extent and perform better. However, if images do not aid in exercise solving, there is no point in teaching students to look more at them. As it stands right now, there is yet a study to be published that shows that images actually aid in exercise solving.

5 Acknowledgements

First I would like to thank my supervisors Jana Holsanova and Marcus Nyström for insightful questions and continued support. I would also like to thank the entire cognitive science class of '15 (especially Trond Tjøstheim, Andrey Anikin, Ludvig Londos, Amanda Bjernestedt and Martin Lingonblad) for comments, discussion and proofreading. Also, I would like to thank Manuel Oliva and Rasmus Bååth for help with the digital classroom and statistics, respectively. Furthermore, I would like to thank Annika Wallin and Christian Balkenius from the cognitive science department for miscellaneous help. I also thank the Humanities Lab in Lund for its generous loan of equipment. Finally, I would like to thank Kerstin Wolf for support, comments, discussions, practical help with the experiment and much more.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183-198.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556-559.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*(04), 723-742.

Bates, D., Maechler, M., Bolker, B., & Walker, S., (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7*, http://CRAN.R-project.org/package=lme4

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction, 20*(2), 155-166.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201-215.

De Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2010). Attention guidance in learning from a complex animation: Seeing is understanding? *Learning and Instruction, 20*(2), 111-122.

Eitel, A., Scheiter, K., Schüler, A., Nyström, M., & Holmqvist, K. (2013). How a picture facilitates the process of learning from text: Evidence for scaffolding. *Learning and Instruction, 28*, 48-63.

Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vision Research, 41*(12), 1561-1574.

Ernest-Jones, M., Nettle, D., & Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior, 32*(3), 172-178.

Ferguson, E. S. (1977). The mind's eye: Nonverbal thought in technology. *Science, 197*(4306), 827-836.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2009). *Cognitive neuroscience: the biology of the mind.* New York: W. W. Norton & company.

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction, 15*(4), 313-331.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving guiding attention guides thought. *Psychological Science, 14*(5), 462-466.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software 33*(2), 1-22.

Hannus, M., & Hyönä, J. (1999). Utilization of illustrations during learning of science textbook passages among low- and high-ability children. *Contemporary Educational Psychology, 24*(2), 95-123.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-122.

Heath, T. L. (Ed.). (1956). *The thirteen books of Euclid's*

Elements. Courier Corporation.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: implications for design. *Topics in Cognitive Science, 3*(3), 446-474.

Hegarty, M., Carpenter, P. A., & Just, M. A. (1991). Diagrams in the comprehension of scientific text. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research: Vol. 2* (pp. 641-668). New York: Longman.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*(6), 717-742.

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 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). ACM.

Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology, 23*(9), 1215-1226.

Holsanova, J. (2014). Reception of multimodality: Applying eye tracking methodology in multimodal research. In Carey Jewitt (Ed.) *Routledge Handbook of Multimodal Analysis,* (pp. 285-296). London: Routledge

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*(2), 172-176.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science, 4*(1), 71-115.

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied, 18*(2), 178.

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

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science,* 18, 513-549.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition, 29*(7), 1000-1009.

Levie, W. H., & Lentz, R. (1982). Effects of text illustrations: A review of research. *ECTJ, 30*(4), 195-232.

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.

Liu, H. C., & Chuang, H. H. (2011). An examination of cognitive processing of multimedia information based on viewers' eye movements. *Interactive Learning Environments, 19*(5), 503-517.

Lowe, R. K. (1999). Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education 14*(2), 225-244.

Lundeberg, M. A., Fox, P. W., & Punccohai, J. (1994). Highly confident but wrong: gender differences and similarities in confidence judgments. *Journal of Educational Psychology, 86*(1), 114.

Mason, L., Tornatora, M. C., & Pluchino, P. (2013). Do fourth graders integrate text and picture in processing and learning from an illustrated science text? Evidence from eye-movement patterns. *Computers & Education, 60*(1), 95-109.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*(1), 1-19.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction, 20*(2), 167-171.

Norman, D. A. (2002). *The design of everyday things.* New York: Basic books.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.

Nyström, M. (2014, March 4). Eye-movements and learning: From vector calculus to dyslexia. Seminar presented at *Thinking in Time: Cognition, Communication and Learning*, Lund.

Nyström, M., & Ögren, M. (2012). How illustrations influence performance and eye movement behaviour when solving problems in vector calculus. In *LTHs 7:e Pedagogiska Inspirationskonferens.*

OECD (2012). *PISA 2009 Technical Report*. OECD publishing.

Ozcelik, E., Arslan-Ari, I., & Cagiltay, K. (2010). Why does signaling enhance multimedia learning? Evidence from eye movements. *Computers in Human Behavior, 26*(1), 110-117.

Ozcelik, E., Karakus, T., Kursun, E., & Cagiltay, K. (2009). An eye-tracking study of how color coding affects multimedia learning. *Computers & Education, 53*(2), 445-453.

Paivio, A. (1991). Dual coding theory: retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 45*(3), 255.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*(3), 618.

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

Scheiter, K., & Eitel, A. (2015). Signals foster multimedia learning by supporting integration of highlighted text and diagram elements. *Learning and Instruction, 36*, 11-26.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and Instruction, 20*(2), 100-110.

Schnotz, W. (2002). Commentary: Towards an integrated view of learning from text and visual displays. *Educational Psychology Review, 14*(1), 101-120.

Schwartz, D. L. (1995). Reasoning about the referent of a picture versus reasoning about the picture as the referent: An effect of visual realism. *Memory & Cognition, 23*(6), 709-722.

Schwonke, R., Berthold, K., & Renkl, A. (2009). How multiple external representations are used and how they can be made more useful. *Applied Cognitive Psychology, 23*(9), 1227-1243.

Skolverket (2011). *Gymnasieskola 2011*. Skolverket.

TIMSS (2007). *International science report: findings from IEA's trends.*

Tsai, M. J., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education, 58*(1), 375-385.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Tversky, B. (2001). Spatial schemas in depictions. In Gattis, M. (Ed.) *Spatial schemas and abstract thought* (pp. 79-111). London: MIT Press.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *Quarterly Journal of Experimental Psychology Section A, 57*(1), 165-182.

Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction, 20*(2), 95-99.

Williamson, V. M., Hegarty, M., Deslongchamps, G., Williamson III, K. C., & Shultz, M. J. (2013). Identifying student use of ball-and-stick images versus electrostatic potential map images via eye tracking. *Journal of Chemical Education, 90*(2), 159-164.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*(6), 495-501.

Yang, F. Y., Chang, C. Y., Chien, W. R., Chien, Y. T., & Tseng, Y. H. (2013). Tracking learners' visual attention during a multimedia presentation in a real classroom. *Computers & Education, 62*, 208-220.

Yarbus, A. L. (1967). Eye movements and vision (B. Haigh, trans.). New York: Plenum Press.