THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Mathematical Reasoning

In physics and real-life context

HELENA JOHANSSON

CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Division of Mathematics
Department of Mathematical Sciences
Chalmers University of Technology
And
University of Gothenburg
Gothenburg, Sweden 2015

ISBN 978-91-628-9431-3 (Printed)
ISBN 978-91-628-9432-0 (Online)

The thesis is available online:
http://hdl.handle.net/2077/38715

Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
SE-412 96 Gothenburg, Sweden
Telephone +46 (0)31 772 1000

Author e-mail: m_helena.johansson@swipnet.se

This doctoral thesis has been prepared within the framework of the graduate school in educational science at the Centre for Educational and Teacher Research, University of Gothenburg.

Centre for Educational Science and Teacher Research, CUL
Graduate school in educational science
Doctoral thesis 45

In 2004 the University of Gothenburg established the Centre for Educational Science and Teacher Research (CUL). CUL aims to promote and support research and third-cycle studies linked to the teaching profession and the teacher training programme. The graduate school is an interfaculty initiative carried out jointly by the Faculties involved in the teacher training programme at the University of Gothenburg and in cooperation with municipalities, school governing bodies and university colleges. www.cul.gu.se

Printed in Gothenburg, Sweden 2015

To Alice and Egil

Abstract

This thesis is a compilation of four papers in which mathematical reasoning is examined in various contexts, in which mathematics is an integral part. It is known from previous studies that a focus on rote learning and procedural mathematical reasoning hamper students' learning of mathematics. The aims of this thesis are to explore how mathematical reasoning affects upper secondary students' possibilities to master the physics curricula, and how real-life contexts in mathematics affect students' mathematical reasoning. This is done by analysing the mathematical reasoning requirements in Swedish national physics tests; as well as by examining how mathematical reasoning affects students' success on the tests/tasks. Furthermore, the possible effect of the presence of real-life contexts in Swedish national mathematics tasks on students' success is explored; as well as if the effect differs when account is taken to mathematical reasoning requirements. The framework that is used for categorising mathematical reasoning, distinguishes between imitative and creative mathematical reasoning, where the latter, in particular, involves reasoning based on intrinsic properties.

Data consisted of ten Swedish national physics tests for upper secondary school, with additional student data for eight of the tests; and six Swedish national mathematics tests for upper secondary school, with additional student data. Both qualitative and quantitative methods were used in the analyses. The qualitative analysis consisted of structured comparisons between representative student solutions and the students' educational history. Furthermore, various descriptive statistics and significance tests were used. The main results are that a majority of the physics tasks require mathematical reasoning, and particularly that creative mathematical reasoning is required to fully master the physics curricula. Moreover, the ability to reason mathematically creatively seems to have a positive effect on students' success on physics tasks. The results indicate additionally, that there is an advantage of the presence of real-life context in mathematics tasks when creative mathematical reasoning is required. This advantage seems to be particularly notable for students with lower grades.

Keywords: Creative mathematical reasoning, Descriptive statistics, Differential item functioning, Figurative context, Imitative reasoning, Mathematical Reasoning Requirements, Mathematics tasks, National tests, Physics tasks, Real-life context, T-test, Upper secondary school.

Preface

The thesis is written, a work that started five and a half years ago is completed. Now comes the hardest part, to in a few lines summarise the time as a doctoral student. As so many before me have pointed out, it is impossible to know what will come. Five and a half year is a long time, and my thoughts in the beginning of the period have been overshadowed by all new impressions and experiences encountered during the way. Overall, it has been a great time. Naturally, the doctoral education has differed a lot from my previous educations, since, in fact, half of the time was dedicated to my own research, the result of which lies in front of you.

Even if the progress sometimes has felt to be very slow, and skepticism about one's own capability have sneaked in, I never doubted that someday the thesis would be finished. This is entirely thanks to the support from my two supervisors, *Professor Mats Andersson* and *PhD. Jesper Boesen.* I thank you truly for all the encouragement, all wise comments, all valuable discussions, for always giving me feed-back and for always taking the time for any questions I might have. I think I was lucky to get you as my supervisors from the beginning of my doctoral studies.

Furthermore, I would like to thank: *Kerstin Wiklander* and *Staffan Nilsson*, who introduced me to statistical methods; and most of all, *Johan Tykesson* for his invaluable support with all the statistical analyses, and for giving quick responses to all my questions. You probably did not realise that it would take that much time when you first agreed to help as a statistical advisor. *Alan Schoenfeld* and the other participants in his 'functions group' at Berkeley University, as well as *Andrea diSessa*, who all made my time as a visiting student researcher in 2013 such a valuable experience. This visit was supported by scholarships from The Royal Swedish Academy of Science, as well as from Stiftelsen Paul och Marie Berghaus donationsfond and Adlerbertska Stipendiefonden. *Evá Fülöp*, my colleague, room-mate and friend, with whom I have discussed ups and downs and how to combine doctoral studies with other engagements in life; and for fun times at conferences, writing schools etc. *Rimma, Anna Ida, Semir, Tuula, Hoda, Anna,* and all other fellow doctoral students and friends in the FLUM group, for sharing experiences and for participating in making these five and a half year a great time.

Finally, I wish to thank and express my love to: my parents, who taught me to think that anything is possible; my mother for having inspired me to always think positively and my father for having encouraged me to think from various perspectives by always saying "Good, but if you think/do it in this way?"; as well as for all practical support that I have needed. My beloved children *Alice* and *Egil*, you are the best! My love *Joacim*, who has shared the last year with me, and boosted my energy by giving his love and support, and for bringing other values into the life.

Helena Johansson, April 2015

Contents

1 Introduction

Mathematics is a subject with many fields, and the character of the various fields is in some instances very different. Sometimes it is pure interest (as in basic research) that pushes the development forward, as was the case for many years in Number Theory, and other times it is the applicability that is the driving force of the research, as e.g. in Financial Mathematics. Some of the purely theoretical discoveries can turn out to have practical applications, such as prime numbers have had for encryption that today is a part of the modern man's everyday life.

No matter which mathematical field that is considered, Hirsch (1996) discusses in his article that all ideas, even the most abstract ones, ultimately origin from real-life experience. According to Wigner (1960), elementary mathematics was formulated to describe real-world entities; but since the concepts in the axioms only allow a handful of theorems to be formulated, more advanced concepts are required. From his point of view of mathematics as "the science of skilful operations with concepts and rules invented for this purpose" (Wigner, 1960, p.2), more advanced concepts are defined in order to permit ingenious logical operations that are considered beautiful in a formal sense, and thus do not have an origin in the real world.

Although not all of mathematics relates to science, mathematics is an integral part of most sciences. As mentioned above, mathematical concepts can turn out to have applicability far beyond the context they were originally developed in. The usefulness of mathematics in science is according to Wigner (1960) bordering to the mysterious without rational explanation. Hirsch (1996), on the other hand, concludes that the strength of mathematics to describe various phenomena depends on the fact that it has evolved to fit the analysis process, e.g. to separate components and reduce all unnecessary information. This can be regarded as a decontextualisation of the situation. Nevertheless, at some instances theories need to be experimentally verified and the mathematics is then used in a context.

Formally defined concepts form the basis of mathematics, but before the concepts are defined it is not unusual that they have been experienced in various forms (Tall & Vinner, 1981). This relates to Hirsch's discussion about how all mathematical ideas have an origin in the world of experience. How people understand a concept depends on the individual's mental pictures and associated properties and processes, which arise from different types of experience. Tall and Vinner use *concept image* to describe the cognitive structure forming the understanding of a concept. Seeing mathematics in context could thus be regarded as helpful when concept images are formed. In school, students have to consider this duality of mathematics when learning both mathematics and science. Sometimes it is necessary to see situations in which mathematics is applied in order to understand the concepts. Other times it is necessary to weed out some parts of the context presented in a task and reduce the task to pure mathematic in order to solve it.

Physics is one of the science subjects in school in which mathematics is a natural part. The relation between the school subjects mathematics and physics are reflected both in mathematics education research and in physics education research. Some of the discussions focus on how physics can influence the learning of mathematics, referred to below as *physics in mathematics*. Other discussions focus on the learning of physics and are concerned with various aspects of its relation to mathematics, and this is referred to as *mathematics in physics*. Since physics describe and formalise real-life phenomena, physics tasks that have to be solved by using some kind of mathematics, can be viewed as special cases of mathematics tasks with real life context.

Boaler (1994) discusses how bringing everyday contexts to mathematics tasks was motivated by that it would help bridging the abstract world of mathematics and students' real world outside the classroom. She separates the different arguments given for introducing contexts in the learning in three categories; 1. Learning is thought to be more accessible if students are given familiar metaphors. 2. Students are thought to become more motivated to learn if they are provided with examples enriching the curriculum. 3. Transfer of mathematical learning is thought to benefit from linking real world problems to school mathematics. The motivating argument, 2., is also noticed by Cooper and Dunne (2000) when they account for how school mathematics was related to the real world. They also saw another reason, the beneficial aspect, the mathematics should be relevant to what the students were supposed to need in their upcoming career and life. Motivation as a reason for tasks with realistic and everyday contexts is also used by Howson (2005) when he discusses school mathematics with meaning. The third category of the arguments discussed by Boaler (1994) could be related to Basson's (2002) paper, in which he discusses that it is valuable to relate mathematical concepts to relevant context from the students' real world in order to accomplish a more general understanding of the concepts.

The aim of this thesis is to contribute to research concerning implications of mastering mathematical reasoning. This is done by studying formal mathematical reasoning requirements in national physics tests, and how the ability to reason mathematically creatively influences students' success on physics tasks.

Furthermore, it is studied how real-life context in mathematics tasks, with different mathematical reasoning requirements, affects students' success on the tasks. The tasks that are used in the analyses come from national tests in both physics and mathematics. The dissertation consists of four papers. The first paper describes a qualitative analysis of the formal mathematical reasoning requirements in national physics tests. The second paper is a more quantitative analysis of students' success, taken required mathematical reasoning into account, on the different national physics tests. The third study uses conditional probability to examine how the ability to reason mathematically may influence success on physics tasks with various kinds of mathematical reasoning requirements. Finally, the fourth study explores how real-life context in mathematics tasks, that require different kinds of mathematical reasoning, affects students' success on the tasks. Various descriptive statistics and significance testing are used in the analyses, and grades and gender were also taken into account.

2 Background

2.1 Physics in Mathematics

Blum and Niss (1991) noticed already in the 80's that the relation between the two subjects mathematics and physics had become weakened in the mathematics education. The main reason for this diminished relation depends, according to Blum and Niss, on that new areas have developed, in which mathematics is important, and that these areas can provide examples suitable for mathematical instructions instead of examples from physics. They agree on the necessity of the opening of mathematics instruction to other applicational areas, but at the same time they stress that it is of great value to keep a close contact between mathematics and physics in school. Examples from physics provide good representative cases for validating mathematical models. They discuss how a separation between the two subjects can lead to unnatural distances between the mathematical models and the real situation intended to model. The weakened relation between the school subjects mathematics and physics is also observed by Michelsen (1998), who describes how the separation of the two subjects has evolved in the Danish school. In a paper by Doorman and Gravemeijer (2009), the authors discuss the advantage of learning mathematical concepts through mathematical model building and how examples from physics allow for a better understanding of the concepts. Hanna (2000), and Hanna and Jahnke (2002) propose that it is advantageous to use arguments from physics in mathematical proofs to make them more explanatory. They refer to Polya (1954) and Winter (1978) and continue discussing the benefits of integrating physics in mathematics education while learning and dealing with mathematical proofs. The importance of using physics to facilitate students' learning of various mathematical concepts is also discussed by Marongelle (2004), who concludes that using events from physics can help students to understand different mathematical representations.

2.2 Mathematics in Physics

Tasar (2010) discusses how a closer relation between the school subjects, mathematics and physics, can contribute to the understanding of physics concepts and can help ensure that students already understand the mathematical concepts needed in physics. Similar suggestion are made by Planinic, Milin-Sipus, Katic, Susac and Ivanjek (2012), who in their study of high school students' success on parallel tasks in mathematics and in physics concluded that students' knowledge is very compartmentalised and that stronger links between the mathematics and physics education should be established. According to Basson (2002), a closer relation might also decrease the amount of time physics teachers spend on redoing the mathematics students need in physics. The redo is likely a consequence of that "physics teachers claim that their students do not have the pre-requisite calculus knowledge to help them master physics" (Cui, 2006, p.2). Michelsen (2005) discusses how interdisciplinary modelling activities can help students to understand how to use mathematics in physics and to see the links between the two subjects. A weaker relation between the subjects, mathematics and physics, in school is also observed more recently in Sweden in the TIMSS Advanced 2008 report. A comparison between syllabuses for physics from different years revealed that the importance of mathematics in physics was more prominent ten years ago, than it is nowadays (Swedish National Agency for Education, 2009a).

Redish and Gupta (2009) emphasise the need to understand how mathematics is used in physics and to understand the cognitive components of expertise in order to teach mathematics for physics more effectively to students. Basson (2002) mentions how difficulties in learning physics not only stem from the complexity of the subject but also from insufficient mathematical knowledge. Bing (2008) discusses the importance of learning the language of mathematics when studying physics. Nguyen and Meltzer (2003) analysed students' knowledge of vectors and conclude that there is a gap between students' intuitive knowledge and how to apply their knowledge in a formal way, which can be an obstacle when learning physics.

In a survey, Tuminaro (2002) analysed a large body of research, and categorised studies concerning students' use of mathematics in physics according to the researchers approach to the area. The four categories are (i) the observational approach; (ii) the modelling approach; (iii) the mathematical knowledge structure approach, and (iv) the general knowledge approach. The observational approach focuses on what students do when applying mathematics to physics problems and how they reason mathematically. Often there are no attempts to give any instructional implications. The modelling approach intends to describe the differences between experts and novices regarding their problem solving skills as well as to develop computer programs that can model the performance of the novices and the experts. Using results from these programs, one hopes to understand the learning process. Research placed in the mathematics knowledge structure approach aims to explain the use of mathematics from cognitive structures of novices and experts. General knowledge structure approach includes research oriented towards an understanding of concepts in general (not only mathematical), using various kinds of cognitive structures. Tuminaro sees a hierarchical structure in the four approaches and compares this structure with a trend in cognitive psychology, towards a refined understanding of cognition. According to Tuminaro, the four approaches towards an understanding of how students use mathematics in physics do not reach the fully sophisticated level, as the trend in cognitive psychology. Tuminaro therefore suggests that there still is a need for research about how the structure of students' knowledge coordinates when they draw conclusions about physics from mathematics.

Mulhall and Gunstone (2012) describe two major types of physics teacher, the *conceptual* and the *traditional*. Mulhall and Gunstone conclude that a typical teacher in the conceptual group presumes that students can solve numerical problems in physics without a deeper understanding of the underlying physics. A typical opinion among teachers in the traditional group is that physics is based on mathematics and that a student develops an understanding of the physics e.g. by working with numerical problems. Doorman and Gravemeijer (2009) notice (with reference to Clement 1985 and Dall'Alba et al. 1993) that most of the attention in both physics and mathematics is on the manipulations of formulas instead of focusing on the conceptual understanding of the formulas.

2.3 Service Subject

In some studies concerning the relation between mathematics and physics, the concept *service subject* emerges. Below follows a very brief review of found definitions/descriptions. Howson (1988) describes *mathematics as a service subject* when mathematics is needed as a complement in other major subjects the students are studying e.g. physics. He stresses that this does not "imply some inferior form of mathematics or mathematics limited to particular fields" (Howson. p. 1). Blum and Niss (1991) observe that focusing on mathematics as a service subject and on co-operation between mathematics and other subjects has been treated separately in the education. They discuss different kinds of mathematical modelling and conclude that for physics situations, mathematics is primarily used to describe and explain the physics phenomena. This use is different from how mathematics is used in models for e.g. economic cases, in which norms are established by value judgements. Niss (1994) discusses different aspects of mathematics, one of which is mathematics as an applied science. In this form, mathematics can serve as a service subject and provide help to understand phenomena in e.g. physics.

2.4 Learning Physics

When discussing learning of physics, there is, of course, a large body of additional literature that is relevant to consider depending on what questions one is studying. A lot of research about teaching and learning physics has been conducted by what Redish (2003) refers to as the Physics Education Research (PER) community. When studying how individuals learn physics, certain cognitive principles have to be considered (Redish, 2003). This approach is discussed by e.g. diSessa (2004), who emphasises the micro levels but from a knowledge-in-pieces perspective. This perspective is not restricted to the learning of physics, but is also applicable in mathematics. According to this micro-perspective, there are many different levels at which a concept can be understood, and contextuality has to be considered. Thus, in order to understand a student's learning, his or her understanding of a particular concept has to be studied in a variety of different contexts (diSessa, 2004).

2.5 Context

Context is one of those concepts that are used with an unequivocal meaning in the education literature. It can for example be used in order to describe the overall situation; or for describing various aspects of the learning situation in the classroom. Other times context is used to denote a subject or various areas of a subject. Furthermore, context can be used to denote students' expectations. Different sources were used in the search for literature discussing context in mathematics or science education in one way or another. The Mathematics Education Database, provided by Zentralblatt, was one of the major sources. Search terms as "context", "tasks", "mathematics" and "physics" were used. The abstracts of the publications in the result list were scanned for indications of possible descriptions of different use of the term "context". These publications were further surveyed for descriptions or definitions of context and references gave ideas of other relevant publications. The search terminated when no new references were found, i.e. references treating context in a different way than already found.

Below follows a description of the various use of context that were found in the literature review. They have been grouped with respect to some similarities, reflected in the headlines.

2.5.1 Cultural context

The cultural context can be considered as describing the overall situation and sets some of the boundaries for the situation the person facing the task is in.

Some of these boundaries could be if knowledge is supposed to be used in problem solving in a *school context* or in an *out of school context* (Verschaffel, De Corte & Lasure, 1994). An example of how an out of school context affects children's/students' learning of mathematics and their ability to use their knowledge to solve mathematical tasks is discussed by Nunes, Schliemann and Carraher (1993). They studied how children to street vendors in Brazil solved informal tasks, i.e. mathematical tasks as "I'd like ten coconuts. How much is that?" posed during interviews; and compared this to the methods the children used to solve tasks in formal tests, i.e. pen-and-paper tests in a school-like setting.

2.5.2 Context as the settings

Another boundary can be the *subject knowledge* assumed to be used when working with the task. If it is mathematical knowledge or physics knowledge that the students will need in order to deal with the problem, or knowledge from several areas together. This use of context can be seen in the paper of Hanna, de Bruyn, Sidoli and Lomas (2004), which concerns how students succeed on constructing mathematical proofs in the context of physics, i.e. students should create mathematical proofs based upon physical considerations.

Factors influencing the learning situation are the organisation of the learning and test situations, as well as the social structure of the situations. Factors like these are referred to as context in some of the literature. For instance, Shimizu, Kaur, Huang and Clarke (2010) use *instructional context* for how individual teachers organise their respective instruction. Other times it is not just the instructional context that is referred to, but the overall *learning classroom settings*; or the *social context*, which describes the learning situation e.g. whole class discussion, assessment or group work (Shimizu et al., 2010). Bell (1993) uses contexts to describe in which situations the mathematics is applied, where situations mostly refers to practical situations students in some sense can relate to. The discussion of learning is more related to teaching method, if students are working individually or if there are whole class discussions, c.f. instructional context.

In his dissertation about problem-solving, Wyndhamn (1993) discusses a study of how different learning settings influence students' success on tasks. He presents the same task for two different groups of students, one group involved in a mathematics lesson and the other group taking social-sciences. The task was to answer how much it would cost to send a letter that weighs 120 grams within Sweden. Students were provided the same table that the Swedish post office provides (see Table 1). Wyndhamn found that when the task was presented in the mathematics lesson, more students tried to calculate the answer instead of just reading of the table. This behaviour resulted in that the students in the social-sciences class were more successful to solve the task.

Table 1. Illustration of the postage table provided to the students.

Maximum weight grams	**Postage SEK**
20	2.10
100	4.00
250	7.50
500	11.50
1000	14.50

Planinic et al. (2012) studied how students understand line graph slope in mathematics and in physics (kinematics). They found that students did better on the parallel tasks when presented as mathematical, and that the added context in physics made the tasks more complex. Because of a lack of conceptual understanding of the relevant physics, students did not know which mathematical knowledge to use even though they sometimes possessed it (Planinic et al., 2012). These results indicate that students' success on tasks are dependent on the context of the settings. These result indicates that students' success on tasks are dependent on the context of the settings.

In a school context, subject areas can be further divided into *different courses* for respective subjects. Studies have shown that students' expectations in e.g. a task solving situation influence their choice of solution methods. This is most often referred to as the *psychological context* in the literature. For instance Bassok and Holyoak (1989) discuss different aspects of transfer and differentiate between psychological context and context as the physical components of the situations. Also Bing (2008) discusses how context can refer to students' expectations and in that way influence which resources that are activated when trying to solve a physics task. Bassok and Holyoak (1989) conclude in their study that contexts in physics tasks direct students' choice of solutions, while mathematics is regarded as more content free and knowledge can be used in new areas that require some novel solutions. Their study shows that students who had learned the general structure of arithmetic progressions, more spontaneously recognised that the same equations could be used to come up with solutions to physics tasks concerning specific areas, like velocity and distance.

In a test situation it is reasonable to assume that the course the test aims to assess influences what expectations students bring to the test situation. Further, the position of a particular task in a test likely influences the psychological context. It is common in many tests in the Swedish education system that “easier” tasks are placed in the beginning of the test and more demanding tasks come in the end of the test.

Several scholars use context when defining or describing intended *areas of a course*. For instance, in a study by Doorman and Gravemeijer (2009), context is used to define the area the physics are supposed to describe; which in their case is a weather forecast. A similar use of context is adopted by Engström (2011) when she for instance discusses how students are able to start a discussion in a physics context and then change to an environmental context, but still use their physics knowledge in the discussion of sustainable development.

2.5.3 Context in tasks

In this thesis, the focus is on how context influences success in solving mathematical tasks. Thus, how context is used by scholars discussing tasks in relation to mathematics education is of primarily interest. Shimizu et al. (2000) discuss how tasks have a central place in the mathematics classroom instructions. Individual teachers’ choice of various tasks thus influences students’ understanding of the mathematics and/or the physics that are taught. *Figurative context* is used by some scholars to describe how a task is posed, see e.g. Palm, (2002). Lobato, Rhodehamel and Hohensee (2012) use the single word *context* to describe the situation posed in the task; for example, a hose is used to fill a pool and the amount of water is graphed with respect to time. The task is to find and interpret the slope in the graph. Bing (2008) uses *context* for how physics tasks are presented. He discusses e.g. how the conception of Newton’s Second Law can be shown by consistent responses from a wide variety of contexts. Marongelle (2004) refers to Kulm (1984) and concludes that when context is used in the literature about mathematical problem solving, it often refers to the non-mathematical meanings that are present in the problem situation. This is similar to what Verschaffel et al. (1994) call *problem context*, a task embedded in some kind of described reality.

Verschaffel et al. (1994), as well as Boaler (1994), and Cooper and Dunne (2000), conclude that the context in tasks in school mostly are artificial and that it sometimes may be negative for students to use their common-sense knowledge as one usually does in real-life problems. This, although the intention of using real-life tasks **is** that students should practice applying formal mathematics in realistic situations. Contexts in questions that are posed in such way that students have to ignore what would have happened in real-life in order to provide the correct answer are called *pseudo-real contexts* by Boaler (1994). She further proposes a somewhat different approach to how context may be used, from what was common at that time. Instead of using context to present to students various specific real-life situation in which certain mathematics can be used, context can be valuable for giving students a real-life situation they have to reflect upon. Instead of trying to remember certain procedures for certain situations, it is valuable to discuss and think about the mathematics that is involved. In this way a mathematical understanding is developed that is easier to transfer to the real world. Boaler concludes from her study that context describing real-life situations only is valuable if the described real-life variables have to be taken into account to solve the posed question.

Palm (2002) discusses how different answers to tasks describing real-life events could be considered correct/reasonable, depending on how the students interpret the purpose of a particular task/question. For example, the decision of how many buses that are required in order to go on a school-excursion if each bus have 40 seats and there are 540 students and teachers in total at the school. Palm argues for how three possible answers (13, 13.5 and 14) can be regarded as correct solutions depending on how the purpose is interpreted. Is for instance the purpose to use the solution as information to decide how many buses that are required or is it the actual number of buses that one wants to order from the bus company that should be given as an answer. Further, should the student account for if more than one child can sit in one seat; and other real life considerations? Palm modified this "bus-task" to a more authentic variant by including an order slip to the bus company, which the students should fill out as an answer. The result showed that then 97 % of the students reflected about and discarded the "half bus" answer, compared to 84 % of the students who solved the original task.

Bergqvist and Lind (2005) investigate whether a change of context or numbers on some mathematical tasks affects how students succeed on the same tasks. In their study, tasks are categorised as either *intra-mathematical* or as *having contexts*. They use the term *intra-mathematical differences* to describe when the numbers in two tasks differ but the formulations are identical. If instead the mathematical content is the same but the real-life situations described in the tasks differ, then tasks are said to have different contexts. Their conclusions are that when two corresponding intra-mathematical tasks differ in numbers, the difficulties of the tasks are mostly not affected. Difference in difficulties could be noticed when the numbers were rational and calculators were not allowed. Difference in context generates likely no difference in difficulties if the formulation of the tasks are similar. If instead one of the tasks is formulated in a more standardised way and the other requires more interpretation of the situation, then differences in difficulties are more likely to occur.

A "similar" study as Bergqvist and Lind's, but on physics tasks, is carried out by Cohen and Kanim (2005), as they investigate if an unfamiliar context in a task contributes to the difficulties physics students have with interpreting linear proportionalities and placing the constant on the right side of the equal sign, called reversal error. They give an example of a typical task:

> Denise is on a journey where she visits planets A and B. Planet A has a radius that is three times as large as the radius of planet B. She finds that she weighs five times as much on planet B as she does on planet A. Write an algebraic expression relating the radius (R) of planet A to the radius of planet B, and a second expression relating the weight (W) of Denise on planet A to her weight on planet B. (p.1).

They noticed that 30% of the students put the constants wrong, i.e. that they wrote $R_B = 3R_A$ instead of $3R_B = R_A$ and/or $W_A = 5W_B$ instead of $5W_A = W_B$. In their study Cohen and Kanim wanted to find out if it was easier for the students to place the constant correct if they knew from the context that one of the variables ought to be larger than the other. For instance to write an expression for "*There are 20 times as many students as professors at this college.*" They found that context with clues did not help the students, on the contrary, a handful students made the error more often in these contexts. Through interviews and through studies about the sentence structure, their conclusion was that the reversal error was of a more syntactic nature, i.e. that the students wanted to stick to the algorithm they had learned for how to translate English into mathematics, even though they intuitively knew which variable should be the larger one (Cohen and Kanim, 2005).

Context free, as opposite to context, is used by some scholars to indicate that mathematical concepts are learned without relating to situations in real life or to other school subjects. For instance, Basson (2002) discusses how mathematical concepts, like function, have been treated in a context free way in South Africa, and how this can hinder students to realise that rules learned in the mathematics class are the same as learned/used in the physics class. This use of context free is similar to the way intra-mathematical is used to describe tasks that do not have a figurative context. The term "context-free" appears also in the paper of Blum and Niss (1991), when they describe how some teachers doubt that connections to other school subjects or other real-life applications belong to mathematics instruction. These teachers' view is that the power of the subject is based on the context-free universality of mathematics, and that involving applications in the instructions can distort the clarity, the beauty and this context-free universality of mathematics. This use of the term context free is thus not the same as how intra-mathematical is used in this thesis.

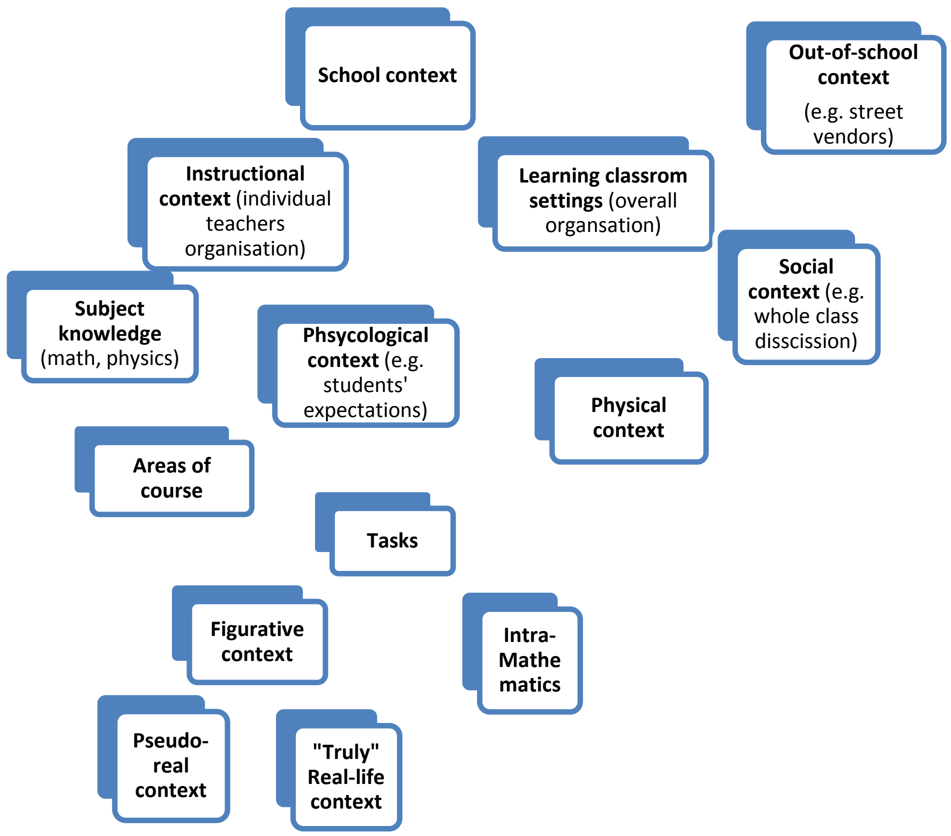


Figure 1. Brief structure of the different use of context

2.6 Gender Differences in Mathematics

In Boaler (1994) she reviewed a small-scaled piece of research and her findings suggest that contexts in some instances affect boys and girls differently. It is more likely that boys perform better than girls if the context in a task includes real world variables that should not be used nor taken into account to reach a solution. Boaler further discusses how this ignoring of the real world may be a reason to girls' disinterest in mathematics. According to the TIMSS Advanced 2008 report, there are significant differences in how boys and girls succeeded on the mathematics test in many of the participating countries. In all except one country, the difference is in boys' favour (Swedish National Agency for Education, 2009a). The analysis of PISA 2012 shows similar results; there is a significant difference in boys' favour regarding success on the test in mathematics when accounting for all 65 participating countries. The differences vary though within countries; in 23 of the countries no gender gap is observed and in five of the countries girls outperform boys (OECD, 2014). As mentioned in for example Ramstedt (1996) or Sumpter (2012, 2015), questions regarding gender differences can be viewed from many different perspectives, e.g. psychological, biological, sociological, historical etc., depending on the intention of the studies.

3 The Swedish Upper Secondary School

3.1 The National Curriculum

The upper secondary school in Sweden is governed by the state through the curriculum, the programme objectives and the syllabuses. In the curriculum are laid down the fundamental values that are to permeate the school's activities as well as the goals and guidelines that are to be applied. The syllabuses, on the other hand, detail the aims and objectives of each specific course. They also indicate what knowledge and skills students must have acquired on completion of the various courses. During the last decades there has been a gradual change toward a stronger focus on process goals and on students' competency to argument for their solutions and to make conclusions, and these goals are present in the curriculum from 1994 (Swedish National Agency for Education, 2006). The shifts are influenced by and similar to international reforms that aim at enriching both mathematics and physics. Content goals are complemented with process goals as those in the NCTM Standards (National Council of Teachers of Mathematics, 2000), and in the NGSS (Next Generation Science Standards Lead States, 2013) where it e.g. is explicated that "emphasis is on assessing students' use of mathematical thinking and not on memorisation and rote application of problem-solving techniques" when high school students use mathematics in physics (NGSS, 2013, HS-PS1-7, Matter and its Interactions). In the framework for PISA 2009 it is emphasised to focus on the mastery of processes and the understanding of concepts (OECD, 2009), and in the TIMSS framework the thinking process is explicated as one of the two dimensions to be assessed (Garden et al. 2006). For a more comprehensive discussion about the reforms and their backgrounds see e.g. Boesen et al. (2014, pp. 73-74). A central part of the reforms concerns reasoning and its central role in problem solving and in the individual's development of conceptual understanding

In the curriculum it is stated that the school should aim to ensure that students acquire good knowledge in the various courses that together constitute their study programme and that they can use this knowledge as a tool, for example, to "formulate and test assumptions" and to "solve practical problems and work tasks". One aspect of knowledge the curriculum focus on, is that school should take advantage of knowledge and experience students bring from "out-of-school" reality. It is the responsibility of the school to ensure that students, after they have finished school, can formulate, analyse and solve mathematical problems of importance for vocational and everyday life (Swedish National Agency for Education, 2006, p. 10 - 12).

Upper secondary school in Sweden is divided into *different national programmes*; *different specially designed programmes* and *programmes provided at independent schools*. A special designed programme could be considered similar to a national programme, and programmes at an independent school could be approved as one of the national programme. Two of the national programmes, the *Natural Science Programme* (NV) and the *Technology Programme* (TE), are oriented towards science and mathematics and include higher courses in mathematics and courses in physics. About 12% of all students in the upper secondary school in Sweden attend the Natural Science Programme or the Technology Programme (Swedish National Agency for Education, 2014).

According to the programme objectives (Swedish National Agency for Education, 2001), NV aims at developing the ability to use mathematics in the natural science and in other areas. It is also stated in the programme objective for NV that in order to develop concepts, students need an understanding of the inter-relationships within and between subjects. The importance of information technology (IT) in for example mathematics and science is outlined in the programme objective for TE. Therefore, one responsibility for TE is to give the students opportunity to attain familiarity with using computers as a tool and to use IT for learning and communication. The different courses in each programme are chosen to fulfil the aims in the different programme objectives. Courses in a school subject are labelled with capital letters, starting with A for the first course and B for the succeeding course and so on. For all students in NV, Mathematics A to D and Physics A are compulsory courses. For students in TE, Mathematics A to C and Physics A are compulsory. In each of the programmes students can choose between different branches. NV has three branches and TE has five. For the branch *Natural Science* for NV (NVNA), Physics B is compulsory and for the branch *Mathematics and Computer Science* (NVMD), Mathematics E is compulsory. Both Physics B and Mathematics E must be offered as optional courses to all students in NV regardless their choice of branch. None of the branches for TE includes requirements of more courses in mathematics or physics, but Physics B and Mathematics D to E must be provided the students as optional (Swedish National Agency for Education, 2001).

3.2 Syllabuses

Mathematics is one of the core subjects in Swedish upper secondary school, together with e.g. English, religion and social science, and Mathematics A is compulsory for all students. This importance of mathematics is expressed in the syllabuses for mathematics –a core subject– as e.g. “The school in its teaching of mathematics should aim to ensure that pupils: develop confidence in their own ability to … use mathematics in different situations, …, develop their ability with the help of mathematics to solve … problems of importance in their chosen study orientation” (Swedish national Agency for Education, 2001, p.112). In addition to core subjects there are programme-specific subjects, as for example physics for NV and TE. According to the syllabus in physics, some of the aims are to: “develop [students’] ability to quantitatively and qualitatively describe, analyse and interpret the phenomena and processes of physics in everyday reality, nature, society and vocational life”, …,” develop [students’] ability with the help of modern technical aids to compile and analyse data, as well as simulate the phenomena and processes of physics” (Swedish National Agency for Education, 2000).

Explicitly, mathematics is important when making quantitative descriptions and implicitly, when analysing data, although the analysing part is mentioned in relation to technical aids. In the syllabuses for the various courses Physics A and Physics B, mathematics is mentioned more explicitly. In Physics A, the students should “be able to make simple calculations using physical models”. In Physics B there is more than one aim that includes mathematics. The student should “be able to handle physical problems mathematically”. They should also “be able to make calculations in nuclear physics using the concepts of atomic masses and binding energy”. Physics B has Physics A as a prerequisite and the students should attain a deeper understanding for some of the physical concepts when studying Physics B. It is also explicated that there are higher demands on the mathematical processing in Physics B (Swedish National Agency for Education, 2000). Besides the aims are also the requirements for the grades in each course stated in the different syllabuses. The final grades students are awarded in a course depend on the achieved level of proficiency (Swedish National Agency for Education, 2000). The grades vary between *Not Pass* (IG), *Pass* (G), *Pass with distinction* (VG) and *Pass with special distinction* (MVG).

3.3 National Tests

The descriptions in the syllabuses of the goals and the different grade levels are quite brief and the intention is that the syllabuses and curriculum should be processed, interpreted and refined locally at each school. By reflecting on how knowledge is viewed in the policy documents, the national tests have several aims and two of them are to concretise the governmental goals and grade criteria, and to support equal assessment and fair grading (Ministry of Education, 2007). The tasks in the tests should contain, among other things, a realistic and/or motivating context. The character and the design of the tasks in tests stress what is covered in the taught curriculum. The tests also influence the teachers' interpretation of the syllabuses, which by extension stress what students focus on (Ministry of Education and Research, 2001; Swedish National Agency for Education, 2003).

The national tests in mathematics and physics are developed by the Department of Applied Educational Science at Umeå University, which has had this commission since shortly after a new national curriculum was implemented in 1994. National tests in mathematics are compulsory for upper secondary students, while the national tests in physics are not.

3.3.1 Mathematics

National tests in mathematics are given twice a year. Following the school year, there is one occasion in December and a second one in May. The occasions are decided by the Swedish Agency of Education. This accounts for the courses Mathematics A-D. Mathematics E on the other hand, is not compulsory and is provided once a year (in spring) by the National test bank in mathematics. Most of the tests are classified as secret for a period of ten years. There are a few tests that become open for public after they have been given. This occurs if there are some major changes in the interpretations of the syllabuses or changes in the assessment goals. The tests are distributed to the school and should be kept secured until the day for the test.

The manual for classifying tasks that are to be included in the national mathematics tests, uses context to describe areas that the tasks are part of (Umeå University, n.d.). A task is said to have no context if it only considers theoretical part of a subject area e.g. solving equations, calculate integrals or simplifying algebraic expressions. If instead the task is to make some geometrical calculations in which no physical objects are included but abstract graphical representations have to be considered, then the context is referred to as abstract. The rest of the different categories for context are divided into five real-life areas; economy-trade-society, industries-crafts, nature-technology, school-home-spare time and health-social care (Umeå University, n.d.). This division of context differs a bit from the ones discussed in Section 2.5. What is here called abstract context would been categorised as intra-mathematical or context-free, i.e. the same category as e.g. "solving an equation".

The national mathematics tests starts with seven to eight tasks that are to be solved without using any equipment other than a pencil. For the following eight to nine tasks a calculator is allowed. One of the tasks in the tests, often the last one, is an *aspect-task* that is assessed according to different aspects, e.g. choice of method, accomplishment, mathematical reasoning and use of concepts. This task should be easy to start with, but it should also include a challenge to more proficient students. During the whole test, students have access to a *formula sheet*, containing some mathematical formulas the students not have to remember. This is handed out together with the test.

3.3.2 Physics

Most of the material provided by the test bank in physics is not open to public, only to upper secondary teachers in physics, who have received a password. In total there are 847 tasks to choose from and 16 complete tests for each of the Physics A and Physics B courses, all classified. The first tests are from 1998 and the latest is from spring 2011. Besides the classified examples, there are five tests for each course that are open for students (or anyone interested) to practice on. These give the students get an idea of what the tests look like and what is required when taking a test. (Department of Applied Educational Science, 2011).

The provided tests comprise two parts; the first one consists of tasks for which a short answer is enough as a solution and the second part consists of tasks that require more analysing answers. For the last ten years, the final task in the tests is an aspect-task. Different aspects assessed are e.g. the use of concepts and models, the use of physics reasoning, and the accounting for the answer. The first three years, 1998-2000, it was an experimental part included in the tests; this part is not included in the analysis in this thesis. A part of the assessment support is that scoring rubrics are provided to the teachers with each test. The guidance in these rubrics has changed some over the years. In the more recent rubrics are e.g. more examples of acceptable answers outlined. Furthermore, the criteria for the highest grade were not explicated in the scoring rubrics for the earliest tests.

As opposed to national tests in mathematics, the teachers are not obligated to use the tests from the National test bank. However a majority of all registered teachers uses the provided physics tests as a final exam in the end of the physics courses (Swedish National Agency for Education, 2005). It is important to stress that National tests are not high-stake test, neither in mathematics nor in physics. The final grade in a course is not solitarily dependent on the achievement on the national test. In fact, teachers are not allowed to grade a course only on a single test, they have to account for all the various aspects the student has shown his/her knowledge during the entire course. After a test from the Test Bank is used, the teachers are intended to report back students' results on the test to the Test Bank.

This thorough description of the National Test tests' purpose and their influence on the mathematics and physics education hopefully clarifies and motivates the choice to use these tests as an indicator of what are formally required mathematically from upper secondary students while studying mathematics and physics.

4 Conceptual framework

4.1 Mathematical Problem Solving

The conceptual framework used in this thesis is related to the various phases of problem solving (Lithner, 2008). Problem solving is used in various contexts with different meanings. Solving mathematical problems can include everything between finding answers to already familiar tasks and trying to proof new theorems. In this thesis *problem* implies when an individual does not have easy access to a solution algorithm (Schoenfeld, 1985). The term *task* on the other hand comprises most work students are involved in during class and while doing homework (Lithner, 2008), which in this thesis narrows down to the work students do while taking a test. Different advantages of working with mathematical problem solving in school are that students' ability to reason mathematically improves, their problem solving skills develop and they become more prepared for life outside school, compared to not working with problem solving (Lesh & Zawojewski, 2007; Schoenfeld, 1985; Wyndham et al., 2000). Learning mathematics through problem solving can also help students to develop their mathematical thinking and their skills in reason mathematically in other areas than pure mathematics, for example physics, (Blum & Niss, 1991).

4.2 Mathematical Reasoning

The impact of mathematical reasoning *on mathematical learning* has been discussed and studied from multiple perspectives. Schoenfeld (1992), for example, points out that a focus on rote mechanical skills leads to bad performance in problem solving. Lesh and Zawojeskij (2007) discuss how emphasising on low-level skills does not give the students the abilities needed for mathematical modelling or problem solving, neither to draw upon interdisciplinary knowledge. Lithner (2008) refers to his studies of how rote thinking is a main factor behind learning difficulties in mathematics. The definition of mathematical reasoning and the conceptual framework that is used for the analyses in this thesis are developed by Lithner (2008) through his empirical studies of how students are engaging in various kinds of mathematical activities. As a result, *reasoning* was defined as ”the line of thought adopted to produce assertions and reach conclusions in task solving” (p. 257).

Just as problem solving, mathematical reasoning is a term that is used with different meanings in various contexts (Yackel & Hanna, 2003). For some scholars, mathematical reasoning is used as a synonym for a strict mathematical proof (e.g.

Duval, 2002; Harel, 2006); others talk about pre-axiomatic reasoning e.g. Leng (2010). The NCTM (2000) distinguishes between mathematical reasoning and mathematical proofs when setting the standards for school mathematics. Ball and Bass (2003) equate mathematical reasoning with a mathematical ability every student need in order to understand mathematics. In this thesis, to be considered as *mathematical reasoning* the justifications for the different reasoning sequences should be anchored in mathematical properties and mathematical reasoning is used as an extension of a strict mathematical proof (Lithner, 2008). When reasoning, one starts with an *object*, a fundamental entity that can be a function; an expression; a diagram etc. To this object, a transformation is done and another object is acquired. A series of transformations performed to an object is called a procedure (Lithner, 2008). The mathematical properties of an object are of different relevance in different situations. This leads to a distinction between *surface* properties and *intrinsic* properties, where the former ones have little relevance in the actual context and the latter ones are central and have to be regarded. How the student makes and motivates the choices in the reasoning sequences is dependent on what resources he/she has access to. Schoenfeld (1985) defines the term resources as the tools; e.g. mathematical knowledge; the student has access to when solving a task. The justification for a choice does not have to be mathematical correct, but it has to be a plausible argument. This means that there is some logic to why a guess would be more reasonable, form a mathematical point of view, than another guess (Polya, 1954). Depending on whether this reasoning is superficial or intrinsic, the framework distinguishes between *imitative reasoning* and *creative mathematical founded reasoning* (Lithner, 2008)

One example, described in Bergqvist, Lithner and Sumpter (2008), of when only surface properties are considered, is a student who tries to solve a max-min problem: "Find the largest and the smallest values of the function $y = 7 + 3x - x^2$ on the interval [-1, 5]". This task can be solved with a straightforward solution procedure: One first uses that the function is differentiable on the whole interval to find all possible extreme points in the interval, (i.e. solve $f'(x) = 0$). If there are extreme points, the values at these points are calculated and compared with the values at the endpoints. In the situation described, the student does not remember the whole procedure, but reacts on the words largest and smallest and starts differentiating the function and solves $f'(x) = 0$. This calculation only gives one value and the answer demands two. Instead of considering intrinsic mathematical properties, the student seeks a method that will provide two values and instead solves the second degree equation $7 + 3x - x^2 = 0$. Two points are now obtained and the function values at these points are accepted as the solution by the student. Although the student gives these values as an answer, it is with some hesitation because the method used did not involve any differentiation, something remembered by the student to be related to a max-min problem.

4.3 Creative Mathematical Founded Reasoning

Creativity is another term that is used in various contexts and without an unequivocal definition, just as problem solving and mathematical reasoning are. There are though mainly two different use of the term: one where creativity is seen as a thinking process which is divergent and overcomes fixation; and another one, where creativity is used when the result is a product that is ascribed great importance to a group of people (Haylock, 1997). Regardless of context, there are two main components that can be crystallised when discussing creativity; these are the usefulness and novelty (Franken, 2002; Niu & Sternberg, 2006).

When creativity is discussed in a mathematical context, it has often been an ability ascribed to experts (Silver, 1997). A quantitative study by Kim (2005) shows a nominal correlation between students' creativity and their scores on IQ-tests, a result supporting the view of not ascribing creativity only to experts or "genius". In a study by Schoenfeld (1985), where he compares novices' problem solving abilities with experts', he concludes that professional mathematicians succeed because of their different way from students of tackling a mathematical problem. These are abilities that can be developed and improved by the students (Schoenfeld, 1985). Silver (1997) makes a similar conclusion in his paper when he discusses the value for educators in mathematics of changing their view of creativity from professional mathematicians' skills, to a mathematical activity every student can improve in school. Sriraman (2009) makes a definition of mathematical creativity "as the process that results in unusual and insightful solutions to a given problem, irrespective of the level of complexity" (p.15).

In the framework used in this thesis, the creativity perspective from Haylock (1997) and Silver (1997) is adopted. That means that creativity is seen as a thinking process that is novel, flexible and fluent. The flexibility indicates that the students have overcome fixation behaviours at some level. The two types of fixation that are intended are content universe fixation, which limits the range of elements that are seen as useful; and algorithmic fixation, which concerns the repeated use of an algorithm once successful (Haylock, 1997).

Creative mathematical reasoning[1] *(CR)* fulfils all of the following criteria:

i. " *Novelty*. A new reasoning sequence is created or a forgotten one is recreated.
ii. *Plausibility*. There are arguments supporting the strategy choice and/or strategy implementation motivating why the conclusions are true or plausible.
iii. *Mathematical foundation*. The arguments made during the reasoning process are anchored in the intrinsic mathematical properties of the components involved in the reasoning." (Lithner, 2008, p.266).

4.4 Imitative Reasoning

The other kind of reasoning used is *imitative reasoning* (IR). The difference between imitative and creative mathematical reasoning, is that there is no flexibility in the thinking process. There are no new reasoning sequences created and the arguments for the chosen solution method (i.e. the reasoning), could be anchored in surface mathematical properties. The reasoner just uses a solution procedure that seems to fit that kind of task. Imitative reasoning is distinguished into *memorised reasoning* (MR) and *algorithmic reasoning* (AR). When it is enough just to recall an answer to be able to solve a task, this is regarded as MR, for example the proof of a theorem.

"MR fulfils the following conditions:

i. The strategy choice is founded on recalling a complete answer.
ii. The strategy implementation consists only of writing it down." (Lithner, 2008, p. 258)

If some kind of calculations is required to solve the task, there is often no use in remembering an answer. Instead it is more suitable to recall an algorithm. Algorithm is here used in a wide sense and refers to all the procedures and rules that are needed to reach the conclusion to a specific type of tasks, not only the calculations.

"AR fulfils the following conditions:

i. The strategy choice is to recall a solution algorithm. The predicted argumentation may be of different kind, but there is no need to create a new solution.

[1] Originally called *creative mathematical founded reasoning*

ii. The remaining parts of the strategy implementation are trivial for the reasoned, only a careless mistake can lead to failure." (Lithner, 2008, p.259).

AR is subdivided into three different categories, depending on how the proper algorithm is argued for. The categories are *familiar algorithmic reasoning* (FAR), *delimiting algorithmic reasoning* and *guided algorithmic reasoning* (GAR). In this thesis/licentiate, only the categories FAR and GAR are used during the various analyses. FAR fulfils:

i. "The reason for the strategy choice is that the task is seen as being of a familiar type that can be solved by a corresponding known algorithm.
ii. The algorithm is implemented." (Lithner, 2008, p. 262).

If the reasoner does not recall any algorithm or is not able to delimit any from the known ones, there can be a need for guidance from an external source to perform the reasoning. The guidance can either be text-guided, e.g. when following an example in the text book that look similar on the surface, or person-guided when for instance the teacher tells every step in the reasoning sequence that has to be made to fulfil the reasoning, without discussing any intrinsic-based mathematical arguments for the choices. GAR fulfils

i. "The strategy choice concerns identifying surface similarities between the task and an example, definition, theorem, rule or some other situation in a text source.
ii. The algorithm is implemented without verificative argumentation." (Lithner, 2008, p.263).

4.5 Local and Global Creative Mathematical Reasoning

Lithner (2008) introduces a refinement of the category CR into *local CR* (LCR) and *global CR* (GCR) that captures some significant differences between tasks categorised as CR. This sub-division has been further elaborated by other scholars e.g. Boesen, Lithner and Palm (2010), and Palm, Boesen and Lithner (2011). In LCR, the reasoning is mainly MR or AR but contains a minor step that requires CR. If instead there is a need for CR in several steps, it is called GCR, even when some parts contain AR and/or MR.

4.6 Non-mathematical Reasoning

The analytical framework in this thesis introduces an additional category called *non-mathematical reasoning* (NMR). This consists of those tasks that can be solved by using just a knowledge of physics. Physics knowledge here refers to relations and facts that are discussed in the syllabuses and textbooks of the physics courses but not in the mathematics courses, for example, the fact that angle of incidence equals angle of reflection. In the same way, the concept of mathematics refers to school mathematics that is introduced in mathematics courses for students at upper secondary school or the mathematics assumed to already be known according to the syllabuses.

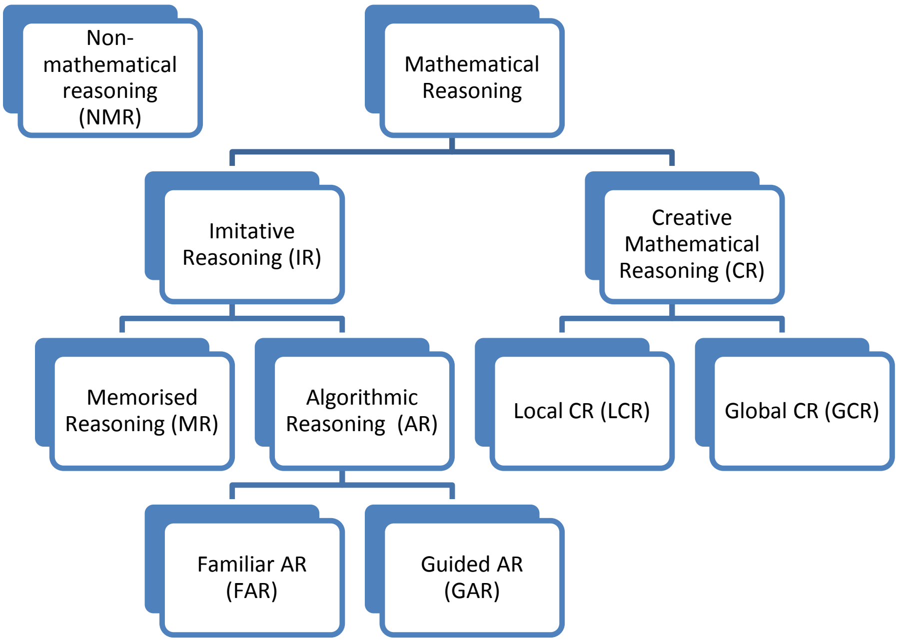


Figure 2. Overview of the conceptual framework.

5 Research Related to the Framework

5.1 Rote Learning, Procedural Knowledge and Imitative Reasoning

Although IR refers to the kind of knowledge that is learned by heart/rote and that studies have shown that rote learning contributes to learning difficulties, this thesis does not imply that students should not learn algorithms. Sfard (1991) discusses how operational and structural knowledge are complementary, and that one of them cannot exist without the other. A three stage hierarchical model is presented and it is stated that before a concept is fully understood, the student has to learn processes/operations that are related to the concept (Sfard). That is, it is necessary to learn some algorithms in order to achieve a deeper mathematical understanding, but it is not enough. In the paper of Gray and Tall (1994), the dichotomy between procedures and concepts is discussed and they introduce a new word *procept*, referring to both the concept and the process that are represented by the same symbol. Although there is an agreement that procedural knowledge is important, it is not enough when students learn mathematics (Baroody, Feil & Johnson, 2007; Gray & Tall, 1994; Sfard, 1991; Star, 2007).

Further there is an argumentation about whether deep procedural knowledge can exist without involvement of conceptual knowledge (Baroody, Feil & Johnson 2007; Star, 2005, 2007). To be successful in mathematics it is necessary for the students to do *proceptual* thinking, which includes the use of procedures. But as Grey and Tall (1994) stress, the proceptual thinking is also flexible i.e. it includes the capacity to view the symbols as a procedure or a mental object depending on the situation. The definitions of the various subcategories of imitative reasoning accounted for above include no such thing as flexibility. On the contrary, the reasoning could be very fixed.

5.2 Physics Reasoning

Since mathematical reasoning in physics tasks is one of the focuses in this thesis, it seems natural to include a brief review of how scholars discuss reasoning in physics; and descriptions of the most commonly used concepts. diSessa (1993) uses the term *p-primes* to describe people’s sense of physical mechanism. P-primes are described as small knowledge structures that in some cases are self-explanatory i.e. things happen because that is the way they are. The p-primes originate from the students’ experiences of the real world. Through learning, appropriate p-primes are activated in relevant situations and new ones can be generated. The function of a p-prime as self-explanatory, also change during learning, as it must be consistent with the physics laws. P-primes are neither wrong nor right in themselves, in some circumstances they are correct and in others not. In this respect, p-primes can be used both when studying reasoning about unproblematic situations and problematic situations. This is different from using the concept misconception, then only wrongly understood situations can be analysed (diSessa, 1993).

Bliss (2008) accounts for studies, conducted by Joan Bliss, Jon Ogborn and others from a period of twenty years, about how students use *common sense reasoning* to explain/describe physical phenomena. Common sense reasoning refers to when students use experiences from everyday life in their reasoning and is explained as “It is the type of reasoning we use to make sense of what is happening around us, or what may have happened, or what will happen” (Bliss, 2008, p.126). One of the results of the studies in Bliss (2008) was that concrete physical schemes are developed through the interaction with real world experience. These schemes are combined to mental models and used when one is trying to understand or predict different physical events i.e. reasoning about physics.

Another concept used for reasoning about physics situations is *qualitative reasoning* or *qualitative physics* (Forbus, 1981, 2004). This concept is mainly used for an area of artificial intelligence (AI) that is modelling the world, from a scientific perspective, using the intuitive notions of human mental models instead of mathematical models. The origin of qualitative reasoning is peoples intuitively reasoning about the physical world, i.e. their common sense reasoning (Klenk, Forbus, Tomai, Kim & Kyckelhahn, 2005). Qualitative reasoning seems also being used by physicists when first trying to understand a problem and later when interpreting quantitative results (Forbus, 2004).

Wittmann (2002) introduces the concept *pattern of association* when discussing reasoning in physics. This refers to the linked set of reasoning resources brought by a student to some specific situation. Some of the resources can be described by diSessa’s (1993) concept p-primes. How these resources are organised when explaining a physics situation is what distinguishes novices from experts, not the existence of the resources (diSessa).

6 Aims and Research Questions

As accounted for in the previous sections, there is a lot of educational research on the relation between the school subjects mathematics and physics that support the necessity of different mathematical competencies when learning physics. Some of the research accounted for shows that how students reason mathematically affect their learning of mathematics. If students only look for superficial properties when they are solving a mathematical task, i.e. using imitative reasoning, it is likely that they end up not understanding the underlying mathematical concepts. Focusing on surface properties is a kind of rote-mechanical procedure that contributes to poor performance in mathematical problem solving. At the same time skills in mathematical problem solving is considered to have a positive effect on the ability to reason mathematically in other areas than mathematics.

Physics is a science developed to describe and model our real world and mathematics is essential in order to formulate the models. If students focus on imitative reasoning when solving physics tasks, one can assume from the previous discussion that they will be given less opportunities to understand the underlying mathematical concepts that occurs in the models. This possible lack of understanding of the mathematics presumably affects the understanding of the physics and then also students' learning of physics. It seems that the mathematical reasoning that is required by students when they are solving physics tasks is not as well studied as the reasoning students use in physics classes.

Physics can furthermore be considered as a real-life context in mathematics tasks in a school setting. As accounted for in Section 2.5, there is a lot of research on how context in mathematics influences students' success and choices of solutions to various tasks. It seems that the relation to various types of mathematical reasoning has not been specifically studied

This thesis concerns the relation between mathematical reasoning and various contexts where mathematics is required. The overall aims concern:

A. How mathematical reasoning affect students' possibilities to master the physics curricula.

B. How real-life contexts in mathematics affect students' mathematical reasoning.

In order to address aim A, it is studied what kind of mathematical reasoning is required to master the physics curricula, as this is concretised through national tests. The first three papers in the present thesis treat this aim by the following specified research questions (RQ), outlined paper-wise:

Paper I, RQ

I.1. Is mathematical reasoning required of upper secondary students to solve national physics tests from the Swedish national test bank?

I.2. If mathematical reasoning is required, what is the distribution of physics tasks requiring CR compared to tasks that are solvable by IR?

Paper II, RQ

II.1. Is it possible for a student to get one of the higher grades, VG and MVG, without using CR?

II.2. If it is possible, how common is it?

Paper III, RQ

III.1. Does the success on a physics task that requires CR affect the probability to succeed on any other task in the same test?

III.2. Does the success on a physics task solvable by IR affect the probability to succeed on any other task in the same test?

Aim B is explored by answering how the presence of context in national mathematics tasks influences students' success in solving tasks requiring different mathematical reasoning. This is analysed in Paper IV through the following RQ

Paper IV, RQ

IV.1. Does the presence of figurative context influence the solution rates on mathematics tasks?

IV.2. Does the presence of figurative context influence the solution rates on mathematics tasks when mathematical reasoning requirements are taken into account?

IV.3. Are there significant differences in students' solving rate on CR-C tasks and on CR-M tasks, and solving rate on IR-C tasks and on IR-M tasks?

IV.4. Does the presence of figurative context have different influences on students' success depending on their grades?

IV.5. Does the presence of figurative context have different influences on students' success depending on their grades if required mathematical reasoning is taken into account?

IV.6. Does the presence of figurative context have different influences on students' success depending on their grades if gender is taken into account?

IV.7. Does the presence of figurative context influence girls and boys significantly differently when account is taken for that they have the same mathematical ability?

7 Methods

For the analysis in Paper I, the following ten national physics tests were used: the December 1998, May 2002, December 2004, May 2005 and December 2008 tests for the Physics A course and the May 2002, May 2003, May 2005, February 2006 and April 2010 tests for the Physics B course. The first tests that were chosen were the non-classified tests, i.e. publicly available, cf. Section 3.3.2, so that examples could be discussed in the thesis. To have five tests from each course, the remaining tests were randomly selected among the classified tests. In order to examine various aspects of students' success in relation to mathematical reasoning, Paper II and III, students' results on the previously categorised physics tests were required. Student data were used by permission from Department of Applied Educational Science at Umeå University, and data possible to get access to were for the May 2002, December 2004 and May 2005 tests for the Physics A course and the May 2002, May 2003, May 2005, February 2006 and April 2010 tests for the Physics B course. The number of students for each test varies from 996 to 3666.

The data used for the analysis in Paper IV come from six Swedish national mathematics tests for three consecutive mathematics courses for upper secondary school; Mathematics B, C and D from December 2003; and Mathematics B, C and D from May 2004. Each task has previously been categorised with respect to mathematical reasoning requirements, i.e. IR or CR, (Palm et al., 2011). In addition to the tests, various information about the students' that have taken the tests are available. The student data contain information about students' score on each task, their total test score, their grade on the test and their course grade, as well as their school, their gender, if Swedish is mother-tongue or not and their attained programme. The number of students varies for the six different tests, from 829 to 3481.

7.1 Categorisation of Mathematical Reasoning Requirements

To categorise physics tasks according to reasoning requirements, solutions to respective task are needed. Required reasoning refers to what kind of reasoning that is sufficient to solve a task, and the used analysis procedure together with the chosen framework gives the possibility to determine this. In order for a task to be categorised as requiring algorithmic reasoning or memorised reasoning, the student should be able to recognise the type of task. This in turn depends on the education history of the solver. The solutions used in the analysing procedure were constructed by the researcher. The fact that these solutions are plausible students' solutions are based on the experience as a physics teacher and the access to the *solution manuals*. Solution manuals are provide by the National test bank, and there is one manual for each test. The manuals comprise suggestions of various acceptable solution to each of the tasks in a test, and directions of how different solutions should be scored. These manuals are a part of the assessment support, and contribute to equal assessment as well as to help interpret the goals in the curricula. Some of the solutions in the manuals are authentic student solutions. In fact, several of the tasks in the tests have been tested on real students, and typical solutions have been selected and included in the manual.

As no students were present in this study, there were no actual learning history to consider. According to studies of how the education in physics and mathematics are organised, a major part of the learning activities seems to be controlled by the textbooks in respective subject (Engström, 2011; Swedish National Agency for Education, 2003 & 2009a; Swedish Schools Inspectorate, 2010; Ministry of Education and Research, 2001). A description of the references' respective findings can be found in Paper I. The learning history of an average student is in this thesis therefore reduced to the content in the textbooks. There are of course other things that play a role in individual students previous experience e.g. tasks discussed during classes and/or physics situations met outside school. Since the learning history is so complex, to equate students' learning history with what is included in the text books is a necessary reduction to be able to perform the study. In view of the references above it is a reasonable reduction. Both textbooks in mathematics and physics were considered in the analysis. Since students are allowed to use a physics handbook during a physics test, the access to formulas and definitions in this handbook also has to be taken into account when analysing the reasoning requirements in the tasks. The textbooks and the handbook were chosen among the books commonly used in the physics courses in upper secondary school. There are about three to four commonly used books in mathematics and physics, respectively. This leads to about 16 different combinations of text books an upper secondary student could have. Even if not all students have the same combination of books chosen for this analysis, the assumption that the books represent the learning history of an average student is reasonable. The chosen mathematics books are "Matematik 3000 Kurs A och B" (Björk & Brolin, 2001) and "Matematik 3000 Kurs C och D" (Björk & Brolin, 2006). The chosen physics books are "Ergo Fysik A" (Pålsgård, Kvist & Nilsson, 2005a) and "Ergo Fysik B" (Pålsgård, Kvist & Nilsson, 2005b), and the chosen physics handbook is ”Tabeller och formler för NV- och TE- programmen" (Ekbom et al., 2004). The procedure for analysing the tasks is given by the chosen framework and an analysis sheet was used to structure the procedure. The steps in the procedure are outlined below and are used previously in e.g. Palm et al. (2011).

I. Analysis of the assessment task – Answers and solutions
   a) Identification of the answers (for MR) or algorithms (for AR)
   b) Identification of the mathematical subject area
   c) Identification of the real life event

II. Analysis of the assessment task – Task variables
   1. Assignment
   2. Explicit information about the situation
   3. Representation
   4. Other key features

III. Analysis of the textbooks and handbook – Answers and solutions
   a) In exercises and examples
   b) In the theory text

IV. Argumentation for the requirement of reasoning

Table 2. The different sources used in the steps in the procedure.

Step I	Tasks in national physics tests	Solution manuals	Physics text books	
Step II	Tasks in national physics tests			
Step III	Tasks in national physics tests	Mathematics text books	Physics text books	Physics handbook

Below follows a thoroughly description of the steps in the procedure

I. Analysis of the assessment task – Answers and solutions: The first step in the procedure consisted in constructing a plausible student solution. The solution was then looked at from a mathematical perspective and categorised according to relevant mathematical subject areas that were required for the solution, e.g. asking if the solution included working with formulas, algebra, diagrams, solving equations, etc. Tasks with solutions not including any mathematical object were identified and categorised as NMR tasks (cf. Section 4.6). Mathematical objects refer to entities to which mathematics is applied. The first step also includes the identification of ’real-life’ events in the task formulation. This identification is relevant because a described situation in the task could give a clue to a known algorithm by means of which one can solve the task (see the Weightlifter (a) example below).

II. Analysis of the assessment task – Task variables: The next step in the procedure was to analyse the solution according to different task variables. The first variable was the explicit formulation of the assignment. The second variable was *what* information about the mathematical objects was given explicitly in the task compared to what information the students need to obtain from the handbook or that they have to assume in order to reach a solution. The third task variable concerns *how* the information was given in the task, e.g. numerically or graphically or whether it was interwoven in the text or explicitly given afterwards. The task could also include keywords, symbols, figures, diagrams, or other important hints the student can use to identify the task type and which algorithm to use. These features were gathered into the fourth task variable.

III. Analysis of the textbooks and handbook – Answers and solutions: The third step in the analysis process focused on the textbooks and the handbook. Formulas used in the solution algorithm were looked for in the handbook, and the available definitions were compared to the constructed solution to the task. The textbooks were thoroughly looked through for similar examples or exercises that were solved by a similar algorithm. The theory parts in the text-books were also examined in order to see whether they contained any clues as to solve the task.

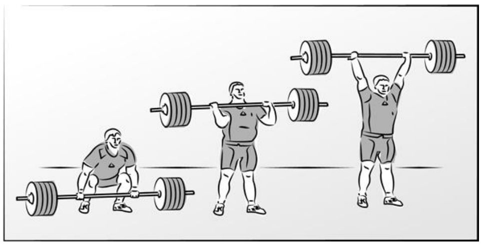
IV. Argumentation for the requirement of reasoning: In the final step, the researcher produced an argument, based on steps I to III, for the categorisation of the reasoning requirement for every task. In order to be categorised as FAR, there must have been at least three tasks considered as similar in the textbooks. It could then be assumed that the students will remember the algorithm, which might not be the case if there are fewer occasions. Three similar tasks was found to be an appropriate number in the study by Boesen et al. (2010). If the task was similar to a formula or definition given in the handbook, it was assumed that the student could use this as a guidance to solve the task. Thus only one similar and previously encountered example or exercise was required for tasks categorised as requiring GAR. To be categorised as requiring MR, tasks with the same answer or solution should have been encountered at least three times in the textbooks. It was then assumed that the student could simply write the same answer for the task. If none of the above reasoning types were sufficient for solving the task and there was a need to consider some intrinsic mathematical property, the task was categorised as requiring some kind of CR.

7.1.1 Examples

The examples below are chosen to represent and illustrate the different types of analysis and the categorisation of the tasks in the national physics tests. All of the tasks are chosen from public tests. Normally, subtasks are treated separately since the task variables and the analysis of the textbooks can be different. The outline of all tasks in a test begins in the same way; first the number of the task in the test is given and after that, enclosed in brackets, the task's number in the National test bank for physics. On the next line the maximum scores for the task are given. The scores are divided into two different categories, G-scores and VG-scores. The maximum scores for each category are separated with a slash, for example 2/0 means that a student can get a maximum of two G-scores and no VG-scores on that particular task. In the same way, 1/1 means that the maximum is one G-score and one VG-score. If the task consists of subtasks: a, b, etc.; the total scores for the subtasks are separated with commas.

Task no. 3 (1584)
2/0, 1/0

A weightlifter is lifting a barbell that weighs 219 kg. The barbell is lifted 2.1 m up from the floor in 5,0 s.



a) What is the average power the weightlifter develops on the barbell during the lift?

Short account for your answer:

b) What is the average power the weightlifter develops on the barbell when he holds it above the head during 3.0 s?

Short account for your answer:

Analysis of 3a

I. Analysis of the assessment task – Answers and solutions: A typical solution from an average student could be derived by the relation between power and the change of energy over a specific period of time. In this task, the change of energy is the same as the change of potential energy for the barbell. Multiply the mass of the barbell by the acceleration of gravity and the height of the lift and then divide by the time to get the power asked for. The mathematical subject area is identified as algebra, in this case working with formulas. The identification of the situation to lift a barbell can trigger the student to use a certain solution method and is, therefore, included in this analysis as an identified "real-life" situation.

II. Analysis of the assessment task – Task variables: The assignment is to calculate the average power during the lift. The mass of the barbell, the height of the lift, and the time for the lift are all considered as mathematical objects. In this example, all of the objects, cf. Section 4.2, are given explicitly in the assignment in numerical form. In the presentation of the task, there is also an illustrative figure of the lift.

III. Analysis of the textbooks and handbook – Answers and solutions: *Handbook:* Formulas for power, $P=\Delta W/\Delta t$, with the explanation "ΔW = the change in energy during time Δt"; for "work during lift", $W_l = mg \cdot h$, with the explanatory text, "A body with weight mg is lifted to a height h. The lifting work is…"; and for potential energy with the text "A body with mass m at a height h over the zero level has the potential energy $W_p = mg \cdot h$". *Mathematics book*[2]*:* Numerous examples and exercises of how to use formulas, e.g. on pages 28-30. *Physics book*[3]*:* Power is

[2] The mathematics text book in all examples is Björk & Brolin (2001)
[3] The physics text book in all examples is Pålsgård et al. (2005a)

presented as work divided by time, and in on example work is exemplified as lifting a barbell. An identical example is found on page 130. An example of calculating work during a lift in relation to change in potential energy is found on page 136. Exercises 5.05 and 5.10 are solved by a similar algorithm.

IV. Argumentation for the requirement of reasoning: The analysis of the textbooks shows that there are more than three tasks similar to the task being categorised with respect to the task variables, and these tasks can be solved with a similar algorithm. As mentioned in the method section, if the students have seen tasks solvable by a similar algorithm at least three times, it is assumed that they will remember the solution procedure. This task is then categorised as solvable using IR, in this case FAR.

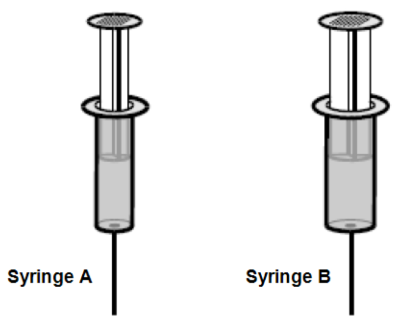
Analysis of 3b

I. Analysis of the assessment task – Answers and solutions: It is not necessary to use any mathematical argumentation in order to solve this task, and solution can be derived by physical reasoning alone. There is no lifting and, therefore, no work is done, and this means that no power is developed. This task is a typical example of an analysis resulting in the NMR categorisation.

Task no. 13 (1184)

0/2

A patient is going to get an injection. The medical staffs are reading in the instructions that they are supposed to use a syringe that gives the lowest pressure as possible in the body tissue. Which of the syringes A or B shall the staff choose if the same force, F, is applied and the injection needles have the same dimensions? *Argue for the answer*

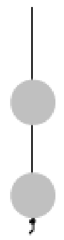


I. Analysis of the assessment task – Answers and solutions: To solve this task, the student can use the relation between pressure, force, and area (p=F/A). Neglecting the hydrostatic pressure from the injection fluid, if the force applied to the syringe is the same then it is the area of the bottom that affects the pressure. The larger the area, the lower the pressure. The staff should choose syringe B. The mathematical subject area is identified as algebra, such as to work with formulas and proportionality.
II. Analysis of the assessment task – Task variables: The assignment is to choose which syringe that gives the minimum pressure and to provide an argument for this choice. Only the force is given as a variable, and this is represented by a letter. Key words for the students can be *force* and *pressure*. The situation is illustrated by a figure in which it appears that syringe B has a greater diameter than syringe A.
III. Analysis of the textbooks and handbook – Answers and solutions: *Handbook:* The relation p=F/A is defined. *Mathematics book:* Proportionalities are discussed and exemplified but are not used for general comparisons. *Physics book:* One example about how different areas affect the pressure and one exercise that is solved in a similar way by using a general comparison between different areas and pressure.
IV. Argumentation for the requirement of reasoning: There is only one example and one exercise that can be considered similar with regard to the task variables and the solution algorithm. The formula is in the handbook, but there has to be some understanding of the intrinsic properties in order to be able to use the formula in the solution. This task is, therefore, considered to require some CR, in this case GCR, in order to be solved.

During the analysis process situations occurred where the analysis was not as straight forward as in the preceding examples. All these tasks were discussed in the reference group and below is one example of a borderline case that arose.

Task no. 12 (1214)
1/2

In order to determine the charge on two small light silver balls, the following experiment was conducted. The balls, which were alike, weighed 26mg each. The balls were threaded on a nylon thread and were charged in a way that gave them equal charges. The upper ball levitated freely a little distance above the other ball. There were no friction between the balls and the nylon thread. The distance between the centres of the balls was measured to 2.9 cm. What was the charge on each of the balls?



I. Analysis of the assessment task – Answers and solutions: To derive a solution, the forces acting on the upper ball must be considered. Because it is levitating freely, it is in equilibrium and, according to Newton's first law, the net force on the ball is zero. The forces acting on the ball are the downward gravitational force, F = mg, and the upwards electrostatic force from the ball below, $F = kQ_1Q_2/r^2$. Setting these expressions equal to each other and solving for Q_1 (and assuming that $Q_1 = Q_2$) will give the charges asked for. The mathematical subject area is identified as algebra, such as to work with formulas and to solve quadratic equations.
II. Analysis of the assessment task – Task variables: The assignment is to calculate the charges on the balls. The mass of the balls and the distance between their centres are mathematical objects given numerically and explicitly in the assignment. The information about the charges' equal magnitude is textual and is a part of the description of the situation. There is also a figure of the balls on the thread illustrating the experiment.
III. Analysis of the textbooks and handbook – Answers and solutions: *Handbook:* Formulas for power, $P=\Delta W/\Delta t$, with the explanation "ΔW = the change in energy during time Δt"; for "work during lift", $W_l = mg\cdot h$, with the explanatory text, "A body with weight mg is lifted to a height h. The lifting work is…"; and for potential energy with the text "A body with mass m at a height h over the zero level has the potential energy $W_p = mg\cdot h$". *Mathematics book:* Numerous examples and exercises of how to use formulas, e.g. on pages 28-30. *Physics book:* Power is presented as work divided by time, and in one example work is exemplified as lifting a barbell. An identical example is found on page 130. An example of calculating work during a lift in relation to change in potential energy is found on page 136. Exercises 5.05 and 5.10 are solved by a similar algorithm.
IV. Argumentation for the requirement of reasoning: The analysis of the textbooks shows that there are more than three tasks similar to the task being categorised with respect to the task variables, and these tasks can be solved by a

similar algorithm. As mentioned in the method section, if the students have seen tasks solvable by a similar algorithm at least three times, it is assumed that they will remember the solution procedure. This task is then categorised as solvable using IR, in this case FAR.

Tasks categorised as solvable by FAR, like 3a above, are hence forward called FAR-tasks and tasks solvable by GAR are called GAR-tasks. Altogether, these kinds of tasks are referred to as IR-tasks. Tasks categorised as solvable by only using physics, that is, when no mathematics were required, like 3b above, are henceforward called NMR-tasks. Tasks requiring GCR to be solved, like 13 above, will be called GCR-tasks and in the same way tasks requiring LCR, like 12 above, will be called LCR-tasks. Tasks requiring either LCR or GCR will be called CR-tasks.

7.2 Comparing Grades with Kinds of Tasks Solved

As mentioned in Section 3.2, the grades a student can receive on a test vary between IG–Not Pass, G–Pass, VG–Pass with distinction, and MVG–Pass with special distinction. To get the grade MVG, students need to fulfil certain quality aspects besides the particular score level. To decide if it is possible for a student to get one of the higher grades without using any kind of CR, each test was first analysed separately. First the score level for each grade was compared with the maximum scores that were possible to obtain, given that the student only has solved (partly or fully) IR- and/or NMR- tasks. The available student data did not give any information about which of the qualitative aspects required for MVG the students had fulfilled, but the data sheets include students' grades; thus MVG could be included in the analyses as one of the higher grades. After analysing if it was possible at all to receive the grades VG or MVG without solving any CR-tasks, students' actual results on the categorised tasks for those particular tests were summed up. The proportion of students who only got scores from IR- and NMR-tasks was then graphed with respect to the different grades.

7.3 Categorising tasks according to context

The tasks used for the analyses corresponding to RQ 7 to 13 come from six Swedish national mathematics tests for three consecutive mathematics courses for upper secondary school; Mathematics B, C and D from December 2003; and Mathematics B, C and D from May 2004. Each task has previously been categorised with respect to mathematical reasoning requirements, i.e. IR or CR, (Palm et al., 2011). In addition to the tests, various information about the students' that have taken the tests are available. The student data contain information about students' score on each task, their total test score, their grade on the test and their course grade, as well as their school, their gender, if Swedish is mother-tongue or not and their attained programme. The number of students varies for the six different tests, from 829 to 3481.

In order to say anything about whether contexts in tasks affect students' success, it is desirable to analyse the tasks from various perspectives and with different methods. Before the analyses of the success could start, the tasks had to be divided into different groups according to if a figurative context was present or if the tasks were intra-mathematical. The tasks were further grouped with respect to required mathematical reasoning, CR or IR. The categories, with respect to which the tasks will be analysed, are: *context task(s)*–tasks with a figurative context, *intraMath task(s)*–tasks without a figurative context, *CR-C task(s)*–context tasks requiring CR, *CR-M task(s)*–intraMath tasks requiring CR, *IR-C task(s)*–context tasks solvable by IR, and *IR-M task(s)*–intraMath-tasks solvable by IR. It is further noticed whether the figurative context in respective context task is a real context or a pseudo-real context (cf. Boaler, 1994).

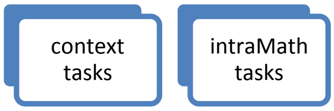


Figure 3. Tasks grouped according to the presence of figurative context.

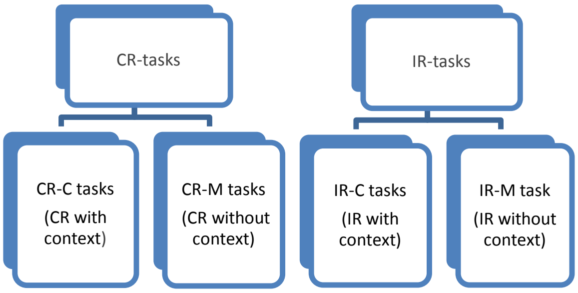


Figure 4. Overview of the subdividing of tasks according to mathematical reasoning and presence of figurative context.

All the tasks that are analysed are solved in a test situation in a school context. The test situations are assumed to be approximately the same for all students. It is further assumed that the setting as a test situation influences average students in similar ways, i.e. it is a test situation and the students' intentions are to manage as well as they can. Therefore, these aspects of the settings are considered to be fixed in this dissertation.

There are settings/factors that do vary, and thus could be important to consider in the other analyses. One of them is the mathematics course respective test assesses. It is assumed that students prepare themselves by studying the relevant areas of mathematics they know will be tested. Because of what is known about the testing system, it is further assumed that students expect that there will not be any tasks assessing any other areas of mathematics than which are specified beforehand. Another one of the factors is a task's position in the test. It is known that the position influences students' expectations regarding whether the task is assumed to be easy or more difficult. The character of the tasks vary depending on whether a calculator is allowed or not, and if the task is an aspect-task. These are further factors worth considering. During the categorisation of the tasks, notes are thus taken about "mathematics course", "test year", "task placement", "calculator" or "no calculator". At the same time it is also identified which mathematical area is involved in the task, e.g. if it is to solve a quadratic equation or maybe to estimate the probability of an event.

7.4 Quantitative Methods

7.4.1 Comparing ratios between conditional and unconditional probabilities

To decide whether there exists a dependence between success on a particular task R, the reference task, and the success on another task X, it was decided to compare the conditional probability to solve X with the unconditional probability to solve X. That is, the ratio

$$\frac{P(X=1|R=1)}{P(X=1)} \qquad (1)$$

was estimated, where X = 1 and R = 1 denote that the tasks have been fully solved, respectively. If this ratio is larger than 1, the probability to succeed on the task X is higher if students successfully have solved the task R than if they have not. The probabilities in (1) are estimated by computing the arithmetic means from the available student data for each test. To estimate P(X=1|R=1), the number of students who had solved both X and R were divided by the number of students who had solved R. The probability P(X=1) was estimated by calculating the number of students who had solved X by the total number of students who had taken the test.

7.4.2 Comparing solution rates

Each test was analysed separately. For every task, the number of students who had solved the task partly or completely was divided by the total number of students who had tried to solve the task. This resulted in a solution rate for each task. For example, consider a task with the maximum score 2. Assume that for this particular task there are 978 students who have got 1 or 2 scores, and there are 1257 students in total, who have tried to solve the task, i.e. students who have 0, 1 or 2 scores. Then the solution rate for that task is 978/1257=0.778.

The tasks, separated by test, were then grouped according to the six categories listed above, and a mean solution rate for every category was calculated. If there are for example 7 intraMath tasks on a test, the solution rates for these tasks are summed and divided by 7. For every test there were now a mean solution rate for each of the six categories; context tasks, intraMath tasks, CR-C task, CR-M tasks, IR-C tasks and IR-M tasks. These different mean solution rates were then compared in order to see if the presence of figurative context could be a reason to any differences.

7.4.3 Paired sample T-test

For the quantitative analysis and significance testing of how figurative context might influence students' success on tasks, the results for individual students were required. For each student, individual solving rates were calculated with respect to the four different sub-categories; CR-C task, CR-M tasks, IR-C tasks and IR-M tasks. For example, to calculate the solving rate on CR-C tasks for a particular student, the student's scores on all CR-C tasks is summed and then divided by the total scores possible to obtain by solving all CR-C tasks. Thus, if the student has got 15 out of 18 of the scores for the CR-C tasks, the student's solving rate for CR-C tasks is 15/18=0.83.

The paired T-test was used for hypothesis testing of the difference between students' means of the solving rates for the pairs CR-C and CR-M tasks, and IR-C and IR-M tasks. The tested null hypothesis is: H_0*: the mean value of the differences between the pairs is zero*. In order to use a parametric test, such as the paired T-test, data have to be normally distributed. Since the t distribution tends to a normal distribution for large sample size, the normality condition could be neglected if the sample size is at least 30 (Sokal & Rohlf, 1987, p. 107). The sample sizes in the present study, and thus the differences (the data), fulfil the criteria, therefore the T-test can be used. At the same time, large sample size always tends to give significant differences, even though they are very small in practice. In order to decide if the significant differences are to be accounted for, Cohen's *d* is used as an index of the effect size. This number is defined as

$$d = \frac{\bar{x}_D}{s_D}, \tag{2}$$

where $\bar{x}_D$ is the difference of the group means and s_D is the standard deviation of the difference. The effect size is classified as small if $d = 0.2$, as medium if d 0.5, and as large if $d \geq 0.8$ (Sullivan & Feinn, 2012).

7.4.1 Descriptive statistics

To be able to say more about if/how figurative context influences students' success, it is desirable to account for students' ability. As a measure of the ability, students' course grades were used. The grades are, as outlined in 3.2 and 7.2: IG, G, VG and MVG. After grouping students with respect to their grades, each grade-group's total score on the Mathematics B 2004 test was summed and compared to the scores respective group had received on each of the categories intraMath tasks and context tasks. After this, the different grade-groups' total scores on CR tasks and IR task were summed and compared to respective group's scores on the CR-M and CR-C tasks, and on the IR-M and IR-C tasks, respectively.

For example, sum the scores on all CR-M tasks that students with the grade MVG have received. Divide this sum by the sum of the total scores the students with the grade MVG has received on all the CR tasks. Make the same calculations for the scores students with the grade MVG has received on all the CR-C tasks in the test. Then repeat this for the remaining grade-groups. By graphing the obtained proportions, a descriptive comparison of the influence of figurative context is obtained.

To also account for gender, students were grouped by gender and kept sub-grouped by grades. For every subgroup, the logarithmic differences between the odds for students' success on intraMath tasks and on context tasks on the Mathematics B 2004 test were calculated. Letting f_M denote an individual student's proportion of intraMath scores and $1 - f_M$ denote the proportion of the intraMath scores not received, the individual student's odds for intraMath-tasks is $f_M/(1 - f_M)$. The odds for context tasks is calculated in the same way, which gives the logarithmic differences as

$$\log\left(\frac{f_M}{1-f_M}\right) - \log\left(\frac{f_C}{1-f_C}\right). \quad (3)$$

The logarithmic differences were then graphed for boys and girls, respectively, and a local regression line (LOESS) was fitted to each graph.

7.4.2 Differential Item Functioning

To significantly test if the presence of figurative context might influences boys and girls differently despite that they have the same ability, the tasks are tested for differential item functioning (DIF). DIF exists if people with the same knowledge/ability, but belonging to different groups, have different probabilities to give the right answer to an item/task. Group belongings could be with respect to, for example, gender (as in the present study), ethnicity, culture or language. A widely used method for detecting DIF is the Mantel-Haenszel procedure (MH) (see e.g. Guilera, Gómez-Benito & Hidalgo, 2009). Holland and Thayer (1988) were the first ones to use MH to detect DIF. Ramstedt (1996) developed and used a modified version of MH to analyse if there were differences between how boys and girls succeeded on national physics tests depending on their gender.

Since one part of MH consists in calculating an *odds ratio* that is used as a measure of the effect size, the concept odds ratio will be explained before the description of MH.

7.4.3 Odds ratio

Odds ratio can be used to measure the dependency between different nominal variables. It is commonly used in various clinical research (Haynes, Sacket, Guyatt, & Tugwell, 2006) or in biological statistics (McDonald, 2009). Because odds ratio can be used for qualitative data and the results only show influence from one variable and remain undisturbed from others, this model is used in various social science research (Ribe, 1999). Keeping the groups fixed, o*dds* is defined as the probability p for an event to happen divided by the probability for the same event not to happen, O = p/(1-p). Odds ratio is then defined as the ratio between the different odds for the event with respect to different groups (see below).

Table 3. Example of a probability matrix

	Y happens	Y does not happen
Group 1	p_1	$1 - p_1$
Group 0	p_0	$1 - p_0$

Odds ratio: $\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{p_1(1-p_0)}{(1-p_1)p_0}$ (4)

If the odds ratio is equal to 1, the probability for the event to happen does not depend on the factor differentiating the groups. Calculations of the odds ratio can thus tell how the probability for success in one group differs from the probability for success in another group. Ribe (1999) describes an example where the odds ratio is used to see how the risk to be unemployed is affected by the country of birth. First the data is stratified so that other variables that also might affect unemployment are held constant. The two groups that are compared are people born in Iran and Sweden, respectively.

Table 4. Situation 1: woman 27 – 39 years, no upper secondary education, single.

Country of birth	**Probability to be unemployed**	**Odds to be unemployed**
Iran	0,78365	3,62214
Sweden	0,32274	0,47654

Odds ratio = 3,62214/0,47654 = 7,6

Table 5. Situation 2: Man 40 – 49 years, higher education, married.

Country of birth	**Probability to be unemployed**	**Odds to be unemployed**
Iran	0,24359	0,32203
Sweden	0,04065	0,04237

Odds ratio = 0,32203/0,04237 = 7,6

The conclusion in this example is that the country of birth affects the risk to be unemployed and that the probability to be unemployed is much higher if one is born in Iran than in Sweden.

7.4.4 The Mantel-Haenszel procedure

MH was originally developed for data analyses from retrospective studies in the clinical epidemiology area. The purpose was to test whether there were any relations between the occurrence of a disease and some factors. The disease could for instance be lung cancer and one factor could be cigarette smoking (Mantel & Haenszel, 1959). A retrospective study can be performed on already collected data and does not require as big sample size as a forward study (also called prospective study) does. In a retrospective study of a disease one looks for unusually high or low frequency of a factor among the diseased persons, while in a forward study it is the occurrence of the disease among persons possessing the factor that is looked at (Mantel & Haenszel). The calculations involved in MH are quite simple and this is probably a contributing factor to that the method is commonly used in various areas today, e.g. epidemiology (Rothman, Greenland & Lash, 2008), biology/biological statistics (McDonald, 2009) and social/educational sciences (Fidalgo & Madeira, 2008; Guilera et al., 2009; Holland and Thayer, 1988; Ramstedt, 1996).

To use MH, data should first be stratified into 2x2 contingency tables. In these tables the rows and the columns represent the two nominal variables that will be tested for dependence. Usually the variable that is placed in the rows is the one that is tested whether it explains/affects the outcome of the variable placed in the columns. The row variable is therefore sometimes called the explanatory variable and the column variable is called the response variable. The different contingency tables represent a third nominal variable that identifies the repeat. The two nominal variables could for example be: a disease and a factor; a plant and a habitat; group belonging and success on tasks. Examples of the repeat variable are different medical centers, different seasons, different teachers etc.

Table 6. Contingency table for repeat i.

Table *i*	Y = 1	Y = 0	Totals
X = 1	a_i	b_i	n_{i1}
X = 0	c_i	d_i	n_{i0}
Totals	m_{i1}	m_{i0}	n_i

In Table 6, X and Y represent the two nominal variables. Both variables are coded by the values 0 and 1 for the respective object included in the study. Belonging to the group of diseased persons might then be represented by X = 1 and not being diseased by X = 0. In the same way, the occurrence of a factor may be represented by Y = 1 and non-existence of the factor by Y = 0. The letters a_i, b_i, c_i and d_i denote the frequencies for respective occurrence and $n_i = a_i + b_i + c_i + d_i$. A diseased person possessing the factor will then be one of those contributing to the frequency a_i. The probability p for an event is estimated by the relative frequency $\hat{p}$. For example, the relative frequency for the event X = 1 and Y = 1 is $\hat{p} = a_i/n_i$.

The method consists in estimating the *common odds ratio*, α_{MH}, for the different contingency tables. The number α_{MH} is estimated as the sum of the weighted odds ratios for the individual contingency tables. From Table 6 follows that the odds for X = 1 and Y = 1 is estimated by a_i/b_i and the odds for X = 0 and

Y = 1 is estimated by c_i/d_i. This gives that the odds ratio for contingency table *i* is estimated by

$$\alpha_i = \frac{a_i/b_i}{c_i/d_i} = \frac{a_i d_i}{b_i c_i}. \tag{5}$$

The common odds ratio calculated in the MH-procedure is defined as

$$\alpha_{MH} = \frac{\sum_i a_i d_i/n_i}{\sum_i b_i c_i/n_i} = \frac{\sum_i w_i \alpha_i}{\sum_i w_i}, \tag{6}$$

where α_i is the odds ratio for table *i* and

$$w_i = \frac{b_i c_i}{n_i} \tag{7}$$

is the weight associated to α_i. The summations run over all contingency tables, i.e. $i = 1,\dots,k$, where *k* is the number of contingency tables. The assumed null hypothesis, H_0, is that there is no dependence between the variables X and Y, i.e. $\alpha_{MH} = 1$.

A main step in the procedure is the calculation of a MH test statistic, which tells whether α_{MH} differs sufficiently from 1 so that H_0 can be rejected. The most commonly used test statistic, χ^2_{MH}, is approximately chi-square distributed, and is compared to a chi-square distribution with one degree of freedom (Mantel & Haenszel, 1959; Ramstedt, 1996; Mannocci 2009; McDonald, 2009). The definition of χ^2_{MH} is

$$\chi^2{}_{MH} = \frac{(|\sum_i a_i - \sum_i E(a_i)| - ½)^2}{\sum_i Var(a_i)}, \tag{8}$$

where $E(a_i) = {n_{i1} m_{i1}}/{n_i}$ is the expected value for a_i under H_0 and

$$Var(a_i) = \frac{n_{i1} n_{i0} m_{i1} m_{i0}}{n_i^2 (n_i - 1)} \tag{9}$$

is the variance for a_i (Mantel & Haenszel, 1959). The value ½ that is subtracted in the numerator for each of the statistics is a continuity correction value (Mantel & Haenszel, 1959; McCullagh & Nelder, 1989).

7.4.5 Modified DIF method

The original method for detecting DIF on dichotomous tasks uses MH as it is described above. The method is thus based on 2x2 contingency tables (Table 7) were the rows indicate group belongings, usually called reference group (R) and focal group (F), and the columns indicate success or not on the task that is analysed. There is one table for every measurement *i* of the ability, which in the present study is students' course grades.

Table 7. Contingency table for repeat *i*. a_i, b_i, c_i and d_i represent the frequencies for right and wrong for the groups R and F. $n_{Ri} = a_i + b_i$, is the number of students in the reference group, and $n_{Fi} = c_i + d_i$, is the number of students in the focal group, $n_i = n_{Ri} + n_F$.

Group	Score on the task		Total
	1	0	
R	a_i	b_i	n_{Ri}
F	c_i	d_i	n_{Fi}
Total	m_{1i}	m_{0i}	n_i

Since the χ^2_{MH} test statistic is dependent on sample size, large sample size tends to always give significant differences, even though they are very small in practice, and small sample size can result in large differences though the result is not significant (also discussed in Section 7.4.3). In order to decide whether the detected DIF is practically significant, α_{MH}, cf. (6), is, as mentioned in Section 7.4.2, used as a measure of the effect size and is called the MH index of the DIF. If $\alpha_{MH} = 1$ there is no difference between the groups' success on the task, if $\alpha_{MH} > 1$ the task is in favour of the reference group, and if $\alpha_{MH} < 1$ the task is in the focal group's favour. Since α is an odds ratio, this means that when for example α_{MH}= 1.6, the odds for the reference group to succeed on the task is on average 60 % higher than the odds for the focal group; and if α_{MH} = 0.6, then the odds for the focal group to succeed is on average 67 % (1/0.6 = 1.67) higher than the odds for the reference group.

To decide whether a task should be classified as a DIF-task, Ramsted (1996) refers to the critical values for the effect size used by ETS (Educational Testing Service) (Longford, Holland & Thayer, 1993, p.175). The effect of DIF is divided into three different groups depending on the value of α_{MH} and χ^2_{MH}; these groups are:

A. negligible DIF when $0.65 < \alpha_{MH} < 1.54$, or ${\chi^2}_{MH} < 3.84$, i.e. the statistic is not significant at the 5% level
B. moderate DIF when ${\chi^2}_{MH} \geq 3.84$ and either a) $0.53 < \alpha_{MH} < 0.65$ or $1.54 < \alpha_{MH} < 1.89$ or b) ${\chi^2}_{MH} < 3.84$ for $\hat{\theta}_{MH} < 0.65$ or $\alpha_{MH} > 1.54$
C. large DIF when $\alpha_{MH} < 0.53$ or $\alpha_{MH} > 1.89$ and ${\chi^2}_{MH} \geq 3.84$ for $\alpha_{MH} < 0.65$ or $\alpha_{MH} > 1.54$

Since the sample sizes for this analysis in this thesis can be considered large, only values for α_{MH} that are significant at the 5 % level will be considered. This implies that if ${\chi^2}_{MH} \geq 3.84$, the tasks will considered a moderate DIF-task if: $0.53 < \alpha_{MH} < 0.65$, or $1.54 < \alpha_{MH} < 1.89$; and a large DIF-task if $\alpha_{MH} < 0.53$, or $\alpha_{MH} > 1.89$.

To be able to use MH for detecting DIF on polytomous tasks, Ramstedt (1996) introduced a modified version of MH in his study about differences in boys' and girls' success on national physics tests. The method could be considered as an approximate dichotomous method in which the polytomous tasks are dichotomised. Instead of letting the frequencies in the contingency tables in the original MH represent the number of boys and girls that have solved vs. not solved the task, the frequencies represent the number of "boy-scores" and "girl-scores" for the different cells. The analogues of the frequencies in Table 7 are calculated according to $a_i = p_{Ri} \cdot n_{Ri}$ and $b_i = (1-p_{Ri}) \cdot n_{Ri}$, where p_{Ri} is the proportion solved tasks (scores) and $1-p_{Ri}$ is the proportion non-solved tasks (non-scores) for group R. The frequencies c_i and d_i for group F are calculated in the same way, that is, $c_i = p_{Fi} \cdot n_{Fi}$ and $d_i = (1-p_{Fi}) \cdot n_{Fi}$, where p_{Fi} is the proportion solved tasks (scores) and $1-p_{Fi}$ is the proportion non-solved tasks (non-scores) for group F. As an example, Ramstedt shows how a polytomous task with the maximum score 3 is dichotomised. On such a task, the proportion "reference group-scores"

$$p_{Ri} = \frac{n_{Ri3} \cdot 3 + n_{Ri2} \cdot 2 + n_{Ri1} \cdot 1 + n_{Ri0} \cdot 0}{3 \cdot n_{Ri}}, \quad (10)$$

is calculated, where n_{Ri3} is the number of students in the reference group with 3 scores on the task, n_{Ri2} is the number of students with 2 scores and so on, with respect to each i. Just as in Table 7, the estimated number of correct solutions is $p_{Ri} \cdot n_{Ri}$ and the estimated number of 0-score solutions is $(1-p_{Ri}) \cdot n_{Ri}$ for the reference group. The generalised solution proportions p_{Ri} and p_{Fi} are thus defined as

$$p_{Ri} = \frac{\sum_j n_{Ri} \cdot i}{M \cdot n_{Ri}} \quad (11)$$

and

$$p_{Fi} = \frac{\sum_j n_{Fi} \cdot i}{M \cdot n_{Fi}}, \quad (12)$$

where the summations run over $j = 0, \ldots, M$, and M is the maximum score on the task.

Table 8. Contingency table for repeat i with the frequencies for right and wrong solutions (maximum and 0 scores), as well as the total frequencies for the groups R and F.

Group	Score on the task		Total
	M	0	
R	$a_i = p_{Ri} \cdot n_{Ri}$	$b_i = (1-p_{Ri}) \cdot n_{Ri}$	n_{Ri}
F	$c_i = p_{Fi} \cdot n_{Fi}$	$d_i = (1-p_{Fi}) \cdot n_{Fi}$	n_{Fi}
Total	$m_{1i} = p_{Ri} \cdot n_{Ri} + p_{Fi} \cdot n_{Fi}$	$m_{0i} = (1-p_{Ri}) \cdot n_{Ri} + (1-p_{Fi}) \cdot n_{Fi}$	n_i

By using the expressions in Table 8 for a_i, b_i, c_i and d_i and putting them into (6), the following expression for the MH DIF index is derived,

$$\alpha_{MH} = \frac{\sum_j pRj \cdot nRi \cdot (1-pFj) \cdot nFj/\mathrm{n}_j}{\sum_j (1-pRj) \cdot nRj \cdot pFj \cdot nFj/\mathrm{n}_j}. \quad (13)$$

The summations run over all contingency tables, i.e. $j = 1, \ldots, k$, where k is the number of contingency tables. The MH test statistic is derived according to equality (8), but with

$$E(a_i) = \frac{n_{Ri} \cdot (pRi \cdot nRi + pFi \cdot nFi)}{n_i} \quad (14)$$

as the expected value for a_i under H_0 and

$$Var(a_i) = \frac{nRi \cdot nFi \cdot (pRi \cdot nRi + pFi \cdot nFi) \cdot ((1-pRi) \cdot nRi + (1-pFi) \cdot nFi)}{n_i^2 (n_i - 1)} \quad (15)$$

as the variance for a_i. By these modifications, MH can be used for detecting DIF on tests with a mixture of dichotomous and polytomous tasks (Ramstedt, 1996). MATLAB is used for the calculations of DIF in the present thesis.

8 Methodology

By using both qualitative and quantitative approaches, different kinds of questions can be answered. Qualitative analyses aim to understand and explain the specific ones, whereas quantitative analyses intend to find generality and causality (Lund, 2012). Lund further discusses how qualitative methods are more suitable for generating hypotheses while quantitative methods are more appropriate for testing hypotheses. By combining these two approaches, the research benefits of the strength from both of them, and is not as sensitive of the weaknesses of respective approach as had been the case if they had been used in separate studies. It is then possible to obtain both the depth from the qualitative analysis, as well as the objectivity and generalisability from the quantitative analysis. To combine qualitative and quantitative methods in a single study, or in a coordinated cluster of individual studies, is a quite young approach and has been established as the formal discipline *Mixed methods* since around 2000 (Lund, 2012).

Mathematical reasoning is the common basis for all conducted studies in the four papers constituting this thesis. During the qualitative analysis of mathematical reasoning requirements in national physics tests, questions were generated about how different kinds of mathematical reasoning affect students' success on the tasks, as well as whether there are any dependence between successes on different kinds of tasks, with respect to mathematical reasoning requirements. Since physics describes and formalises real-life phenomena, physics tasks that have to be solved by using some kind of mathematics, can be viewed as special cases of mathematics tasks with real-life context. Thus questions about how context (not restricted to physics) in mathematics tasks might affect students' success arose. The thesis is permeated with an explorative approach, in which hypotheses are generated by the outcome from the previous analyses, cf. Lund (2012) discussed above. The decision to account for grades and gender in some of the analyses, was based on the fact that this information was available in the data.

The reasons for choosing Lithner's (2008) framework for categorising mathematical reasoning requirements in physics tests are: the different reasoning categories are well-defined and can be used as a concrete tool for categorising empirical data; the framework is anchored in empirical data; the framework has been used in previous studies, e.g. Bergqvist (2007), Palm et al. (2011) and Sumpter (2013); and a part of this thesis relies on some of the categorisations in Palm et al. (2011).

The procedure used in this thesis to categorise physics tasks, described in Section 7.1, was developed by Palm et al. during their categorisation of mathematics tasks. The resulting categorisation of tasks is only meaningful if it represents the reasoning actually used by students while solving the tasks; and this can be achieved with the well-documented criteria required for each category; together with a routine for agreement-discussions about the categorisation. Alternatively, higher reliability could be reached with a less complex phenomenon, e.g. by defining creative mathematical reasoning as solutions consisting of more than three steps. This, on the other hand, would give a very low validity of the meaning of creative mathematical reasoning.

The validity of the analysis of mathematical reasoning requirements is dependent both on the appropriateness of the procedure used for the categorisation and on the fact that the outcome from the categorisations are in accordance with students' actual reasoning. The appropriateness is argued for above, and the concordance with students' reasoning will be argued for below.

The construction of a typical solution in the first step of the procedure is one of the methodological considerations. Identification of the mathematical subject area and the task variables depends on this typical solution and the results from the identification affect the categorisation of the required mathematical reasoning. Hence, how this typical solution is constructed can affect the result, i.e. the distribution of the mathematical reasoning requirements could differ from the one presented in this study. The third step of the procedure consist in an analysis of the textbooks in mathematics and physics. As mentioned earlier, one textbook each for mathematics and physics have been chosen to represent an average upper secondary students' used literature. There are about four different textbooks for each of the courses and the choice of mathematics and physics books is often made locally at each school. The combination of the textbooks that students in one school use could differ from the combination used by students in other schools. Although the textbooks cover essentially the same subject areas, examples and exercises could vary between the books. This influences the number of similar tasks, as the one analysed, that can be found in the textbooks, which in turn affects the categorisation of the analysed task and eventually the presented distribution of the mathematical reasoning requirements. If the examples the teachers discuss during classes would have been included in the analysis, the number of similar tasks might be higher than when only textbooks are used as a representation of the learning history. The number of IR-tasks would then have been higher and then consequently, the number of CR-tasks would have been lower.

In the last step of the procedure, a task is argued for to be a FAR-task if similar tasks have been met at least three times before. The fact that three is an appropriate assumption is supported by a study by Boesen et al. (2010). They use three as a minimum to categorise a task as FAR and found that when students are put in front of FAR-tasks in national mathematics tests, the students try to recall appropriate algorithms to solve the test tasks. It is clear that another choice than three as the minimum number will affect the number of tasks categorised as FAR in general. It is also likely that the number of similar tasks different students need to have met to be able to remember a solution differ.

An argumentation for concordance between the theoretical established reasoning requirements and the reasoning an actual student would use is based on results from Boesen et al. (2010). In that study, real students' actual type of mathematical reasoning that were used to solve tasks on tests in mathematics, was compared with the prior theoretically established reasoning requirements for the same tasks (according to the same procedure as described in Section 7.1). It was shown that only 3 % of the tasks were solved with a less creative reasoning than what was judged to be required; and 4 % of the tasks were either solved with more creative reasoning or not solved at all. These results indicate that the establishment of reasoning requirements in this way provides meaningful results. The construction of a plausible student solution is one of the four steps in the analysis procedure. As the author's experience of physics students and physics tests can be considered similar to the experience of mathematics and mathematics tests of the scholars in Boesen et al. (2010), their result about the method's validity should be relevant in this thesis as well.

The categorisation of all physics tasks in the present thesis was made by the author. During the analysing process, both tasks where the categorisation was straightforward and tasks where the categorisation could be considered as border-line cases occurred. Typical examples of the different kinds of categorisation were continuously discussed in a reference group consisting of the author, a mathematics education researcher well familiar with the analysis procedure, and a mathematician. All categorisations considered as border-line cases were discussed in the group, thus no inter-reliability estimate was calculated.

The categorisation of mathematics tasks according to whether they consist of a figurative context or not is much less complex than categorising tasks according to mathematical reasoning requirements. The criterion for a mathematics task to be categorised as a context task is that a real-life event should be described in the task. Thus, both the validity and the reliability for this categorisation is high.

As described in 7.4.3, the paired samples t-test was chosen for testing if the mean difference of students' solving rates on two different categories of tasks was zero. For a t-test to give valid results some assumptions have to be fulfilled, these are: the dependent variable should be continuous, the independent variable should consist of related groups and the distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed. In this thesis the dependent variable is the students' solving rate, which is on a continuous scale; and the independent variable is the students taking the test, which contains both groups of tasks that the solving rates are calculated for, and these groups are thus related. Since the data consist of solution frequencies, which only can attain a finite number of values, data are not normally distributed, and thus neither is the difference. But at the same time, the sample sizes in the present study are large ($n > 829$), and from the Central limit theorem it then follows that the sample means are approximately normally distributed when the sample size is 30 or greater (Sokal & Rohlf, 1987, p.10). The conclusion is that results from the paired samples t-test are valid.

An alternative to the t-test could have been to use a non-parametric test, e.g. Wilcoxon signed rank test that does not presume normally distributed data. On the other hand there are no confidence interval obtained, that together with the effect size, can be used to determine the practical significance. Since large sample size always tends to give statistically significant p-values, the practical significant is desirable to consider in order to draw any conclusions. As described in Section 7.4.3, Cohen's *d* is used as the parametric index of the estimated effect size.

The use of the Mantel-Haenszel procedure in order to test tasks for DIF is a well-established method. Calculated values for both the χ^2_{MH} test-statistic and the α_{MH} index for effect size are used when the effect of DIF is estimated according to the scale developed by the ETS. The results from the present DIF analysis could thus be considered both valid and reliable. As described in Section 7.4.1, students' course grades were used as a measure of the ability. An alternative would have been to use their test scores, which would have provided a more fine grained scale for the ability. On the other hand, tests score is dependent on the observed task itself, which could lead to circle dependencies. In any case, the course grade is enough for this study, and since the course grade is based on both the national test result and other performances made during the course, possible circle dependencies can be reduced.

All the national tests involved in the studies as well as the student data are used by permission from the Department of Applied Educational Science at Umeå University. Furthermore, student data are anonymous, and therefore no ethical conflicts exits.

9 Summary of the Studies

9.1 Paper I

By analysing the mathematical reasoning required to solve tasks in national physics tests, the idea in Paper I is to capture the mathematical reasoning that is required to master or fully master the physics curricula for upper secondary school. It is explicated in the physics syllabuses that the use of mathematics is incorporated in the goals and that the national tests are the government's way of concretising the physics curricula. As outlined in Section 0, RQ I.1 and I.2, i.e. if mathematical reasoning is required to solve physics tasks and how the required reasoning is distributed, is studied in Paper I.

To answer these questions, 209 tasks from ten different physics tests from the National test bank in Physics were analysed. The first tests chosen were the unclassified tests, cf. Section 3.3.2, so that examples could be discussed. In order to have five tests from each course, the remaining tests were randomly selected among the classified tests. The analysis consisted in a thoroughly qualitative examination of the tasks as well as of the textbooks in mathematics and physics. The textbooks were chosen to represent the students' learning history. Certain task variables were identified and the test tasks were compared to the tasks met in the textbooks. The used method is described in Section 7.1 and methodological consideration is discussed in Section 0. The tasks were categorised according to the different kind of mathematical reasoning, FAR, GAR, LCR or GCR, required to reach a solution; or if the task were solvable by only using knowledge from physics, NMR. The analysis was made by the author and often the process was straight forward, but occasionally some border-line cases arose. Thus various examples were continuously discussed in a reference group consisting of the author, one mathematician and one mathematics education researcher. Examples of the different kinds of analyses are presented in Section 7.1.1.

The main results from the analysis of the mathematical reasoning requirements show that 76 % of the tasks required mathematical reasoning in order to be solved, and of these tasks 46 % required CR, which corresponds to 35 % of the total number of tasks. These results answer the first two research questions, to what extent and of what kind mathematical reasoning is required when solving physics tests.

9.2 Paper II

In order to further explore how mathematical reasoning requirements affect upper secondary students' mastering of the physics curricula, RQ II.1 and II.2 (cf. Section 0) were studied. Categorised tasks from the study in Paper I, in the eight tests for which student data were accessible, were used in this study. In order to examine RQ II.1 and II.2 each test was analysed separately. The requirements for the various grades were compared to the total score possible to receive in each category, i.e. IR, CR or NMR. For those tests for which it was possible to get a higher grade than Pass without using CR, the proportion of students who had not solved any CR tasks completely were graphed with respect to their grades on the tests. The method is described more thoroughly in Section 7.2.

The result shows that in three of the eight tests it was possible to receive one of the higher grades without solving any tasks requiring CR. Nevertheless, when graphing the proportions, it turned out this does not occur too frequently. Only in one of the eight tests a larger number of the students got a higher grade than Pass without solving CR tasks. This test, however, differs from the other ones in the way that the number of NMR tasks is larger and this could be a reason for the larger number of students with higher grades.

Viewing the physics tests from the National test bank as an extension of the national policy documents, one can assume that students' results on the tests are a measure of their knowledge of physics. The results above show that a focus on IR when studying physics in upper secondary school will make it hard for the students to do well on the physics tests, thus fully mastering the physics curricula. Therefore, a reasonable conclusion is that a creative mathematical reasoning can be regarded as decisive, which strengthen the outcome from Paper I.

9.3 Paper III

The aim of Paper III is to examine how upper secondary students' ability to reason mathematically affects their success on different kinds of physics tasks. Here "different" refers to the different kinds of mathematical reasoning that are required in order to solve the tasks. To address this aim, RQ III.1 and III.2 (cf. Section 0) are addressed. By analysing students' success in solving physics tasks that require different kinds of mathematical reasoning the study in Paper III deepens the results from Paper I and Paper II, and contributes to the overall aim A, cf. Section 0. The research questions are analysed by comparing ratios between conditional and unconditional probability to solve the different physics tasks. A detailed description of the method is found in Section 7.4.1. The tasks, with corresponding student results, that are used in the analysis are the physics tasks from the same eight national physics tests that are used as data for the study in Paper II.

The main result strongly indicates that mastering creative mathematical reasoning has a positive effect on the success on physics tasks. It is shown that the effect is higher for tasks requiring CR compared to tasks solvable by IR. As previously discussed, former studies have shown that focusing on IR can contribute to learning difficulties and poor results in mathematics. It has then been assumed that creative mathematical reasoning has a positive effect on the learning. The result in Paper III could indicate empirical support for this assumption.

9.4 Paper IV

In the final paper in this thesis, the aim is to explore if/how the presence of a figurative context in mathematics tasks affects upper secondary students' success on the tasks and if the success differs with respect to required mathematical reasoning. This is done by analysing RQ IV.1 to IV.7 (cf. Section 0). The data consist of mathematics tasks from two successional Swedish national tests in each of the courses Mathematics B, C and D, together with student data for each test. The tasks are categorised in a previous study by Palm et al. (2011) with respect to required mathematical reasoning. Both descriptive statistics and significance testing have been used in the analyses, the methods are explained in detail in Section 7.3 and Sections 7.4.2 to 7.4.5.

The main results indicate that there is a greater success on CR tasks if a figurative context is present in the tasks, and that this influence is particular evident for students with lower grades. Furthermore, the presence of context in tasks on national mathematics tests does not seem to influence boys' and girls' success differently when no account for required reasoning is taken. If mathematical reasoning requirements are considered, there are indications that boys with lower grades benefit more from the presence of figurative context in CR tasks than girls with lower grades do. The results suggest that in order to develop an ability to reason mathematically creatively, the presence of a figurative context is beneficial for the students.

10 Discussion of the Results

The importance of mathematics is explicated in the syllabuses for both physics and mathematics, as well as in the curriculum for upper secondary school. In this thesis one of the aspects of mathematics is studied, namely mathematical reasoning, and this aspect is studied in relation to national tests in physics and in mathematics. As mentioned in Section 3.3, the purpose of national tests is to be an assessment support to teachers, and also a guiding of how to interpret the syllabuses/curriculum. The national tests may thus be viewed as an extension of the policy documents, and are in this dissertation used to represent the mathematical reasoning that is required to master/fully master the mathematics and physics courses, according to the policy documents. Due to the same reason, the context analysis of the national mathematics tests are assumed to mirror formal requirements expressed through the policy documents. If the national tests do not align with the documents, the resulting classification of mathematical reasoning consists most likely of more tasks solvable by IR than if the alignment is good. The result is most likely not an overestimate of the requirements of CR.

10.1 Influences of Mathematical Reasoning on Students' Development of Physics Knowledge

Since a part of the present thesis focuses on what kind of mathematics that are *required* of students and not on students' *use* of mathematics in physics, the thesis can be regarded as a complement to the studies categorised by Tuminaro (2010), see Section 2.2. By considering one part of the role of mathematics that students are confronted with when learning physics in upper secondary school, the first three papers are situated within the Mathematics in Physics research field (cf. Section 2.2). In these papers national physics tests are used in the analyses. Because of the way the national tests are constructed (cf. Section 3.3), students who fully master the physics curricula should have the ability to solve any of the tests for the intended course. Therefore, the fact that some of the individual tests in the study in Paper I have a slightly lower proportion than one-third of tasks requiring CR does not weaken the result that creative mathematical reasoning is significant.

The results from the first two papers confirm that the ability to reason mathematically is important and an integral part when solving tasks in physics tests from the National test bank; and thus an integral part of the physics curricula. Mathematical reasoning is according to the definition a process to reach conclusions in tasks solving. When students have the ability to use CR they know how to argue and justify their conclusions and they can draw on previous knowledge. As it is not enough to use only IR to solve a majority of the tasks in a test, but especially CR is required, it is suggested that the ability to use creative mathematical reasoning is necessary to fully master the physics curricula; and thus decisive when students develop their physics knowledge. Further support for this assumption is given by the result in Paper II. The analysis of the score levels for the grades revealed that it was impossible to pass six of the eight tests without reasoning mathematically. This is a main result strengthen the conclusion above.

Also worth commenting on is the result that it is possible for students to attain one of the higher grades without using any kind of CR on three out of the eight tests. However, comparing this result with student data tells that this occurs rarely. Thus the importance of being able to reason mathematically, in particular the ability to use CR, to pass and to do well on physics tests is strengthened further.

The goals and the subject descriptions in the Swedish policy documents of what it means to know physics are quite rich; and are highly in accordance with the content and cognitive domains in the TIMSS Assessment framework (Garden et al., 2006; Swedish National Agency of Education, 2009b); and thus assumed to be shared internationally. The alignment between the TIMSS framework and the Swedish policy documents suggests that the results from Paper I can be regarded as a sort of universal requirement of mathematical reasoning to master a physics curricula. Therefore, the outcome to RQ II.1 and II.2, also says something about how this universal requirement relates to a specific assessment system's formal demands, in this case Sweden's.

This thesis does not claim to say anything about individual students' learning or that mathematical reasoning is the only component that affects students' learning of physics; but viewing the national physics tests as an extension of the national policy documents, one can assume that students' results on the tests are a measure of their knowledge of physics. As mentioned in Section 2.4, individuals' understanding of the relevance of different physics concepts in various contexts has to be examined to discuss what has been learned. To be able to use mathematical concepts learned in another context than the one in the present situation is inherent in the definition of CR. Solving tasks requiring CR may thus reflect a more developed understanding (of the mathematics).

The necessity of being able to reason mathematically is, with respect to the results in Paper I and Paper II, one of the things communicated by the tests to the teachers. According to the well-known saying "*What you test is what you get*", tests stress what is focused on. Thus the necessity of mathematical reasoning is also communicated to the students. The results in this thesis suggests that it is unlikely to attain a higher grade than *Pass* without having some understanding of intrinsic mathematical properties. From the discussion in Section 5.1 about procedural and conceptual knowledge, we know that some intrinsic understanding may follow from working with exercises involving standard procedures. At the same time it is clear that one cannot fully understand the underlying concepts if the focus only is on the procedures. It is well known that a focus on IR can explain some of the learning difficulties that students have in mathematics. The results in Paper II show that a focus on IR when studying physics in upper secondary school will make it hard for the students to do well on the physics tests, thus fully mastering the physics curricula. Therefore, a reasonable conclusion is that focusing on IR can hinder students' development of knowledge of physics, similar to results found about mathematics. Thus, viewing physics only as the mathematical formulas in the handbook is not fruitful for students striving to succeed on physics tests–something most teachers are aware of.

The conclusion above is further strengthened from the result in Paper III. This result gives strong indications of that the ability to reason mathematically creatively has a positive influence on the success on other physics tasks, and that the effect is higher for tasks requiring CR compared to tasks solvable by IR. When students are able use their knowledge in novel situations, they have developed another approach to the task solving process. Their strategy is based on the judgement of plausibility, which means that they analyse the task/assignment and have an idea of plausible conclusions. The ability to reason mathematically creatively is thought to be generalisable to various mathematical areas. Therefore it is reasonable that the effect between success on CR tasks is higher than the effect of success on a CR task and on an IR task.

Nevertheless, there still is a positive effect on IR tasks from the success on CR tasks, and the effect seems to be a bit greater than the corresponding effect from success on IR tasks. This result suggests that students who have developed the ability to reason mathematically creatively also have a better chance to succeed on tasks of a more procedural character. This result could be compared to the findings in Boaler (2002), in which the learning of students who experienced different approaches to mathematics teaching was analysed. The result showed that students who had focused more on learning procedures could rarely use their knowledge in anything other than textbook and specific test situations. On the other hand, students who had experienced a more project based teaching attained significantly higher grades on the national exam. This is thought to be due to that they had developed a more conceptual understanding of the mathematics, and thus a different form of knowledge that are more effective in various kinds of situations.

An interesting conclusion of the analysis of the effect of success on IR tasks, is that the effect is highest on the tasks requiring GCR. This could be seen in relation to Star's (2005, 2007) point of view, in which deep procedural knowledge does not exist without some conceptual understanding of the knowledge. According to the framework for mathematical reasoning, no intrinsic understanding is required in order to solve IR tasks, but this does not exclude the possibility that students could have developed some conceptual understanding; and thus, success on IR tasks positively affects the success on GCR tasks. At the same time, the only characteristics different IR tasks have in common, at least theoretically, are that they should be possible to solve by remembering an answer or a procedure and implement this (cf. Section 4.4). Therefore it is reasonable to expect that the effect on the success on other IR tasks are smaller than the effect on CR tasks. The result gives strong indications of the positive effect of creative mathematical reasoning on task solving. The result might provide empirical evidence for, and contribute to the discussion about, the effect mathematical reasoning has on students' development of knowledge of mathematics as well as of physics. Since the analysis was conducted on physics tasks, continued studies of the dependence should be performed on mathematics tasks in order to deepen and generalise the result.

As shown in Paper I, there were on average more CR tasks in the Physics B tests than in the tests for Physics A. Scrutinising this result for each test shows that the outcome varies over the years and that the variation between tests for the same course sometimes is bigger than between the different courses. Further analysis in Paper II reveals that the average proportions of the scores for the different reasoning categories are the same for tests in Physics A and Physics B. Assuming that the average result is general and drawing on the results from both papers, it seems that solving a CR task in a Physics A test scores higher than solving a CR task in a Physics B test. One conclusion could be that being able to perform more demanding mathematics is more valued in Physics A than in Physics B. Comparing this to the text in the syllabuses, which states that there is a higher demand on the mathematical processing in Physics B, one could ask if not giving as many points for the creative mathematical processing in Physics B as in Physics A is a consequence of the test developers interpretation of the syllabuses.

All students should have the same possibilities to achieve the goals in the physics curricula. Therefore, they ought to be given the opportunity in school to develop and practice this creative mathematical reasoning that is required. As mentioned in Section 7.1, it is common in the physics classes that students solve routine tasks and focus on manipulations on formulas instead of focusing on the conceptual understanding of the underlying principles. A reasonable assumption is that if there is more focus on physics procedures than on the understanding of physics concepts, there is also little focus on creative mathematical reasoning. On the other hand, it is not only the physics classes that might provide students the opportunity to develop a mathematical reasoning ability, this is of course relevant also in the mathematics classes. According to studies about the learning environment in mathematics classes, the focus are on algorithmic procedures and the environment does not provide extensive opportunities to learn and practice different kinds of reasoning (e.g. Boesen et al., 2010). During observations of classroom activities it was shown that opportunities to develop procedural competency was present in episodes corresponding to 79% of the observed time; compared to episodes involving opportunities to develop mathematical reasoning competency, which were present in 32% of the observed time (Boesen et al., 2014). Also tests have an indirect role for students learning, both as formative, when students get feedback on their solutions, and as summative, when the character of the tasks gives students indications of what competences that are sufficient for handling mathematical tasks. Analyses of teacher-made mathematics tests have shown that these focused more on imitative reasoning than the national mathematics tests do (Palm et al., 2011).

Altogether, the above discussions show that CR could be considered formally required to master the physics curricula and thus regarded as decisive when developing physics knowledge. At the same time students seem to be provided limited opportunities to develop his/her creative mathematical reasoning.

10.2 Influences of Figurative Context in Mathematics on Students' Mathematical Reasoning

As outlined previously, Paper IV examines another perspective of mathematical reasoning compared to the first three papers. The perspective in paper IV departs from the relation between upper secondary students' success on tasks in national mathematics tests and the presence of figurative context in the tasks, as well as whether the required mathematical reasoning together with the context has any influence on the success. The results show that it is only after taking account for mathematical reasoning requirements that the presence of figurative context seems to have an effect on students' success. The results indicate that it is easier for students to solve CR tasks if they are embedded in a figurative context. This seems to be reasonable, since if the assignment is embedded in a real-life context that the students can relate to (cf. Sections 0 and 2.5), this helps them to come up with possible ways to solve the problem.

When analysing influence of figurative context, it turned out that the level of difficulty varies for the different tests, and therefore no general conclusion could be drawn. Differences between results on tests from different years are likely due to internal property of the tests, rather than differences between the students' abilities. Instead, in order to study further how context might influence students' success it is desirable to construct intraMath and context tasks with the same level of difficulty and let students in two similar classes solve these tasks.

That especially students with lower ability seem to be in greater need of relating the mathematics to a familiar reality, is noteworthy for teachers' practice. In order to give all students the same possibilities to learn mathematics, it should be desirable to use relevant contexts from students' everyday life when mathematical concepts are introduced. This corresponds to e.g. Boaler's (1994) discussion about the arguments for introducing real-life context into the mathematics education. Furthermore, it is notable that the presence of context in CR tasks seems to affect boys with lower grades more than girls with lower grades. At the same time, when analysing if boys and girls with the same grade succeeded differently on the individual tasks, the presence of context could not alone explain any differences. Since gender differences on tasks not could be explained by the presence of figurative context, the tasks that did show significant differences need to be analysed further from various perspectives.

Finally, one can indeed conclude that the advantage of everyday context in mathematics tasks is very complex and that no general conclusions could be made from the various analyses in the present thesis. It seems to be a positive relation between requirements of CR and figurative context, as well as between the success of students with lower grades and the presence of figurative context.

Bibliography

Ball, D. & Bass, H. (2003). Making mathematics reasonable in school. In J. Kilpatrick, G. Martin & D. Schifter (Eds.), *A Research Companion to Principles and Standards for School Mathematics,* (pp. 27-44). Reston VA: National Council of Teachers of Mathematics.

Baroody, A. J., Feil, Y., Johnson, A. R. (2007). An Alternative Reconceptualization of Procedural and Conceptual Knowledge. *Journal for Research in Mathematics Education,* 38(2), 115-131. Stable URL: http://www.jstor.org/stable/30034952

Bassok, M. & Holyoak, K. (1989). Interdomain Transfer Between Isomorphic Topics in Algebra and Physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(1), 153-166. doi:10.1037/0278-7393.15.1.153

Basson, I. (2002). Physics and mathematics as interrelated fields of thought development using acceleration as an example. *International Journal of Mathematical Education in Science and Technology*, 33(5), 679–690. doi:10.1080/00207390210146023

Bell, A. (1993). Some experiments in diagnostic teaching. *Educational Studies in Mathematics* 24(1), 115-137. doi: 10.1007/BF01273297

Bergqvist, E. (2007). Types of reasoning required in university exams in mathematics. *Journal of Mathematical Behavior*, 26(4), 348-370. doi:10.1016/j.jmathb.2007.11.001

Bergqvist, E. & Lind, A. (2005). *Är det svårare att dela med fyra än med två när man läser matte C?: En jämförelse av svårighetsgrad mellan olika versioner av matematikuppgifter i Nationella kursprov* [Is it more difficult to divide by four than by two when taking the course Mathematics C? A comparision of different versions of mathematics tasks in National tests]. Umeå, Sweden: Department of Educational Measurement, Umeå universitet.

Bergqvist, T., Lithner, J. & Sumpter, L. (2008). Upper Secondary Students' Task Reasoning, *International Journal of Mathematical Education in Science and Technology,* 39 (1), 1-12. doi:10.1080/00207390701464675

Bing, T. (2008). *An epistemic framing analysis of upper level physics students' use of mathematics.* (Doctoral dissertation). Retrieved from http://bit.ly/Bing2008

Björk, L-E. & Brolin, H. (2001). *Matematik 3000, kurs A och B.* Stockholm, Sweden: Natur och Kultur.

Björk, L-E. & Brolin, H. (2006). *Matematik 3000, kurs C och D.* Stockholm, Sweden: Natur och Kultur.

Bliss, J. (2008). Commonsense reasoning about the physical world, *Studies in Science Education*, 44(2), 123-155. doi:10.1080/03057260802264149

Blum, W. & Niss, M. (1991). Applied mathematical problem solving, modelling, applications, and links to other subjects — State, trends and issues in mathematics instruction. *Educational Studies in Mathematics*, 22(1), 37-68. doi:10.1007/BF00302716

Boaler, J. (1994). When Do Girls Prefer Football to Fashion? An analysis of female underachievement in relation to "realistic" mathematic contexts. *British Education Research Journal*, 20(5), 551-564. doi:10.1080/0141192940200504

Boaler, J. (2002). Experiencing school mathematics: Traditional and reform approaches to teaching and their impact on student learning. (Revised and expanded edition.) Mahwah, NJ: Lawrence Erlbaum Associates.

Boesen, J., Lithner, J. & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning student use. Educational studies in mathematics, 75 (1), 89-105. doi:10.1007/s10649-010-9242-9

Boesen, J., Helenius, O., Bergqvist, E., Bergqvist, T., Lithner, J., Palm, T. & Palmberg, B. (2014). Developing mathematical competence: From the intended to the enacted curriculum. Journal of Mathematical Behavior, 33, 72–87. doi:10.1016/j.jmathb.2013.10.001.

Clement, J. (1985). Misconceptions in graphing. In L. Streefland (Ed.), *Proceedings of the Ninth International Conference for the Psychology of Mathematics Education* (pp. 369–375). Utrecht: Utrecht University.

Cohen, E. and Kanim, S. (2005). Factors influencing the algebra "reversal error". *American Journal of Physics*, 73(11), 1072-1078. doi:10.1119/1.2063048

Cooper, B. & Dunne, M. (2000). *Assessing children's mathematical knowledge: social class, sex and problem-solving*. Buckingham, United Kingdom: Open University Press.

Cui, L. (2006). *Assessing college students' retention and transfer from calculus to physics* (Doctoral dissertation). Retrieved from http://bit.ly/Cui2006

Dall'Alba, G., Walsh, E., Bowden, J., Martin, E., Masters, G., Ramsden, P. & Stephanou, A. (1993). Textbook treatments and students' understanding of acceleration. *Journal of Research in Science Teaching*, 30(7), 621-635. doi:10.1002/tea.3660300703

Department of Applied Educational Science (2011-10-06). Some earlier given test in Physics. Retrieved from http://bit.ly/DepAES2011

diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction,* 10 (2-3), 105-225. doi:10.1080/07370008.1985.9649008

diSessa, A. A. (2004). Contextuality and coordination in conceptual change. In E. Redish and M. Vicentini (Eds.), *Proceedings of the International School of Physics "Enrico Fermi:" Research on physics education* (pp. 137-156). Amsterdam, The Netherlands: ISO Press/Italian Physics Society.

Doorman, L. M. & Gravemeijer K. P. E. (2009). Emergent modeling: discrete graphs to support the understanding of change and velocity. *ZDM – The International Journal on Mathematics Education,* 41, 199-211. doi:10.1007/s11858-008-0130-z

Duval, R. (2002). Proof understanding in mathematics: What ways for students? In F.L. Lin (Ed.), *Proceedings of the international conference on mathematics- "Understanding proving and proving to understand"* (pp. 61-77). Taipei, Taiwan: National Science Council and National Taiwan Normal University.

Ekbom, L., Lillieborg, S., Larsson, S., Ölme, A. & Jönsson, U. (2004). *Tabeller och formler för NV- och TE-programmen, 5 ed.* Stockholm, Sweden: Liber.

Engström, S. (2011). *Att vördsamt värdesätta eller tryggt trotsa. Gymnasiefysiken, undervisningstraditioner och fysiklärares olika strategier för energiundervisning* [To respectfully value or confidently defy. Upper secondary physics, teaching traditions and physics teachers' different strategies for energy education] (Doctoral dissertation). Retrieved from http://bit.ly/EngstromDiss2011

Fidalgo, A. M. & Madeira, J. M. (2008). Generalized Mantel-Haenszel Methods for Differential Item Functioning Detection. *Educational and Psychological Measurement*, 68(6), 940-958. doi:10.1177/0013164408315265

Forbus, K. D. (1981). Qualitative reasoning about physical processes. *Proceedings of the 7th IJCAI,* (pp. 326-330). Vancover, BC.

Forbus, K. D. (2004). Qualitative Reasoning. In A. B. Tucker (ed.) *Computer science handbook 2nd ed* (62-1 – 69-19). Boca Raton, Fla.: Chapman & Hall/CRC

Franken, Robert E. (2002). *Human motivation 5 ed*. Belmont, CA: Wadsworth/Thompson Learning.

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O., Mullis, I. V. S., et al. (2006). *TIMSS Advanced 2008 Assessment Frameworks*. Boston:

TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Gray, E. M. & Tall, D. O. (1994). Duality, Ambiguity and Flexibility: A Proceptual View of Simple Arithmetic. *The Journal for Research in Mathematics Education*, 26(2), 115-141.
Stable URL: http://www.jstor.org/stable/749505

Guilera, G., Gómez-Benito, J., Hidalgo, M. D. (2009). Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF. *Psicothema* 21(3) 493-498. ISSN 0214 – 9915

Hanna, G. (2000). Proof and its classroom role: A survey. In M.J. Saraiva et al (Eds.), *Proceedings of Conference en el IX Encontro de Investigacao en Educacao Matematica* (pp. 75-104). Funado.

Hanna, G. & Jahnke, H. N. (2002). Another Approach to Proof: Arguments from Physics. *ZDM – The International Journal on Mathematics Education*, 34 (1), 1–8. doi:10-1007/BF02655687

Hanna, G., de Bruyn, Y., Sidoli, N. & Lomas, D. (2004). Teaching Proof in the Context of Physics. *ZDM – Zentralblatt für Didaktik der Mathematik*, 36(3), 82-90. doi:10.1007/BF02652775

Harel, G. (2006). Mathematics Education Research, Its Nature, and Its Purpose: A Discussion of Lester's Paper. *ZDM – The International Journal on Mathematics Education,* 38 (1), 58-62. doi:10.1007/BF02655906

Haylock, D. (1997). Recognising mathematical creativity in schoolchildren. *ZDM – The International Journal on Mathematics Education*, 29(3), 68-74. doi:10.1007/S11858-997-0002-Y

Haynes, R., Sacket, D. L., Guyatt, G. H., & Tugwell, P. (2006). *Clinical Epidemiology: How to Do Clinical Practice Research 3rd edition*. Philadelphia, PA: Lippincott Williams & Wilkins.

Hirsch, R. (1996). Is Mathematics a Pure Science? *Science & Society*, 60(1), 58-79. Stable URL: http://www.jstor.org/stable/40403533

Holland, P. W. & Thayer, D. T. (1988). *Differential item performance and the Mantel Haenszel Procedure*. In H. Wainer & H. I. Braun (Eds) Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Howson, A. G. (ed.) (1988). *Mathematics as a service subject* ICMI study series, 99-0639469-2 based on the work of a symposium held in Undine (Italy), from 6 to 10 April, 1987. Cambridge, United Kingdom: Cambridge University Press.

Howson, G. (2005). “Meaning” and School Mathematics. In J. Kilpatrick, C. Hoyles, O. Skovsmose & P. Valero (Eds.), *Meaning in Mathematics Education* (pp. 17-38). New York, NY: Springer.

Kim, K. H. (2005). Can only intelligent people be creative? A meta-analysis. *Journal of Secondary Gifted Education*, 16(2/3), 57-66. ISSN: 10774610. Accession Number: 17971216

Klenk, M., Forbus, K., Tomai, E., Kim, H. and Kyckelhahn, B. (2005). Solving Everyday Physical Reasoning Problems by Analogy using Sketches. *American Association for Artificial Intelligence*. Retrieved from http://bit.ly/KlenkEtAl2005

Kulm, G. (1984). The classification of problemsolving research variables. In G. A. Goldin & C. E. McClintock (Eds.), *Task variables in mathematical problem solving* (pp. 1-21). Philadelphia, PA: The Franklin Institute Press.

Leng, M. (2010). Preaxiomatic Mathematical Reasoning: An Algebraic Approach. In G.Hanna et al (Eds.), *Explanation and Proof in Mathematics, Philosophical and Educational Perspectives*. (pp. 47-57). New York, NY: Springer Science + Business Media, LLC 2010

Lesh, R., & Zawojewski, J. (2007). Problem Solving and modeling. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 763-804). Charlotte, NC: Information Age Publishing.

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics,* 67(3), 255-276. doi:10.1007/s10649-00-9104-2

Lobato, J., Rhodehamel, B. & Hohense, C. (2012). “Noticing” as an Alternative Transfer of Learning Process. *Journal ot the Learning Sciences*, 21(3), 433-482. doi:10.1080/10508406.2012.682189

Longford, N. T., Holland, P. W. & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer, (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Lund, T. (2012). Combining Qualitative and Quantitative Approaches: Some Arguments for Mixed Methods Research. *Scandinavian Journal of Educational Research*, 56(2), 155-165. doi:10.1080/00313831.2011.568674

Mannocci, A. (2009). The Mantel-Haenszel procedure. 50 years of the statistical method for confunders control. *Italian Journal of Public Health*, 6(4), 338-340. doi:10.2427/5765

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute,* 22(4), 719-748. doi:10.1093/jnci/22.4.719

Marongelle, K. A. (2004).How Students Use physics to Reason About Calculus Tasks. *School Science and Mathematics,* 104(6), 258-272. doi:10.1111/j.1949-8594.2004.tb17997.x

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models – 2nd edition.* Cambridge: University Press.

McDonald, J. H. (2009). *Handbook of biological statistics, 2nd edition*. University of Delaware, Baltimore: Sparky House Publishing

Michelsen, C. (1998). Expanding Context and Domain: A cross-Curricular Activity in Mathematics and Physics. *ZDM – The International Journal on Mathematics Education*, 30(4), 100-106. doi:10.1007/BF02653149

Michelsen, C. (2005). Expanding the domain – variables and functions in an interdisciplinary context between mathematics and physics. In Beckmann, A., Michelsen, C., & Sriraman, B (Eds.). *Proceedings of the 1st International Symposium of Mathematics and its Connections to the Arts and Sciences* (pp.201-214). The University of Education, Schwäbisch Gmünd, Germany.

Ministry of Education and Research (2001). *Samverkande styrning: om läroplanerna som styrinstrument* [Interacting governance. About the curricula as means for policy change] Ministry publications series Ds 2001:48. Retrieved from http://bit.ly/DS2001

Ministry of Education and Research (2007). *Tydliga mål och kunskapskrav i grundskolan – Förslag till nytt mål- och uppföljningssystem* [Explicit goals and knowledge requirements in the compulsory school – Proposal for a new system for goals and follow-up]. Ministry publications series SOU 2007:28. Stockholm, Sweden: Fritzes.

Mulhall. P. & Gunstone. R. (2012). Views About Learning Physics Held by Physics Teachers with Differing Approaches to Teaching Physics. *Journal of Science Teacher Education,* 23 (5), 429-449. doi:10.1007/s10972-012-9291-2

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For States, By States (HS-PS)*. USA: Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS.

Nguyen, N-L. & Meltzer, D. (2003). Initial understanding of vector concepts among students in introductory physics courses. *American Journal of Physics*, 71(6), 630-638. doi:10.1119/1.1571831

Niss, M. (1994). Mathematics in Society. In Biehler, R., Scholz, W., Sträβer, R. & Winkelmann, B. (eds.) *Didactics if Mathematical as a Scientific Discipline* (pp. 367 – 378). Dordrecht: Kluwer Academic Publisher.

Niu, W & Sternberg, R. J. (2006). The philosophical roots of Western and Eastern conceptions of creativity. *Journal of Theoretical and Philosophical Psychology*, Vol 26(1-2), 18-38. doi:10.1037/h0091265

Nunes, T., Schliemann, A. & Carraher, D. (1993). *Street mathematics and school mathematics.* Cambridge, United Kingdom: Cambridge University Press.

OECD. (2009). *PISA 2009 Assessment Framework – Key competences in Reading, Mathematics and Science.* Retrieved from http://bit.ly/PISA2009Frame

OECD (2014). *Are boys and girls equally prepared for life?* Retrieved 2015-02-17 from http://bit.ly/oecd2012_gender

Palm, T. (2002). *The Realism of Mathematical School Tasks – Features and Consequences*. (Doctoral dissertation). Umeå, Sweden: Umeå University.

Palm, T., Boesen, J., & Lithner, J. (2011). Mathematical Reasoning Requirements in Upper Secondary Level Assessments. *Mathematical Thinking and Learning, 13*(3), 221-246. doi:10.1080/10986065.2011.564994

Planinic, M., Milin-Sipus, Z., Katic, H., Susac, A. & Ivanjek, L. (2012). Comparison of Student Understanding of Line Graph Slope in Physics and Mathematics. *International Journal of Science and Mathematics Education*, 10(6), 1393-1414. doi:10.1007/s10763-012-9344-1

Pólya, G. (1954). Mathematics and plausible reasoning (Vols. I and II). Princeton, NJ: Princeton University Press.

Pålsgård, J., Kvist, G. & Nilson, K. (2005a). *Ergo Fysik A.* Stockholm: Liber.

Pålsgård, J., Kvist, G. & Nilson, K. (2005b). *Ergo Fysik B.* Stockholm: Liber.

Ramstedt, K. (1996). *Elektriska flickor och mekaniska pojkar. Om gruppskillnader på prov – en metodutveckling och en studie av skillnader mellan flickor och pojkar på centrala prov i fysik* [Electrical girls and mechanical boys. About group differences on tests – a method development and a study about differences between boys and girls on the national tests in physics]. (Doctoral dissertation). Umeå, Sweden: UmU Tryckeri.

Redish, E. F. (2003). *Teaching Physics with the Physics Suite*. Hoboken, NJ: Wiley.

Redish, E. F. & Gupta, A. (2009). *Making Meaning with Math in Physics: A Semantic Analysis*. Paper presented at GIREP 2009, Leicester, United Kingdom. Retrieved from http://bit.ly/RedishGuptaGIREP2009

Ribe, M. (1999). Oddskvoter berättar. [Odds ratios tell]. *Välfärdsbullentinen*, 4, 14-15. Retrieved from http://bit.ly/Ribe1999

Rothman, K.J., Greenland, S., & Lash, T.L. (2008). *Modern Epidemiology, 3rd Edition*. Philadelphia, PA: Lippincott, Williams & Wilkins.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for Research on Mathematics Teaching and Learning* (pp. 334-370). New York, NY: MacMillan.

Sfard, A. (1991). On the dual nature of mathematical conceptions: reflections on process and objects as different sides of the same coin. *Educational Studies in Mathematics*, 22(1), 1-36. doi:10.1007/BF00302715

Shimizu, Y., Kaur, B., Huang, R. & Clarke, D. (2010). The Role of Mathematical Tasks in Different Cultures. In Y. Shimizu, B. Kaur, R. Huang and D. Clarke (Eds.), *Mathematical Tasks in Classrooms Around the World* (pp.1-14). Rotterdam, Netherlands: Sense Publisher.

Silver, E. (1997). Fostering creativity through instruction rich in mathematical problem solving and problem posing. *ZDM – The International Journal on Mathematics Education*, 29(3), 75-80. doi:10.1007/s11858-997-0003-x

Sokal, R. R. & Rohlf, F. J. (1987). *Introduction to biostatics. 2nd ed.* New York, NY: Freeman.

Sriraman, B. (2009). The Characteristics of Mathematical Creativity. *ZDM-The International Journal on Mathematics Education*, 41(1&2), 13-27. doi:10.1007/s11858-008-0162-4

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 36(5), 404-411. Stable URL: http://www.jstor.org/stable/30034943

Star, J. R. (2007). Foregrounding procedural knowledge. *Journal for Research in Mathematics Education*, 38(2), 132-135. Stable URL: http://www.jstor.org/stable/30034953

Sullivan, G. M. & Feinn, R. (2012). Using Effect Size–or Why the *P* Value Is Not Enough. *Journal of Graduate Medical Education*, 4 (3), 279-282. doi:10.4300/JGME-D-12-00156.1

Sumpter, L. (2012). Upper secondary school students´ gendered conceptions about affect in mathematics. *Nordic Studies in Mathematics Education*, 17 (2), 27–47.

Sumpter, L. (2013). Themes and interplay of Beliefs in Mathematical Reasoning. *International Journal of Science and Mathematics Education*, 11(5), 1115-1135. doi:10.1007/s10763-012-9392-6

Sumpter, L. (2015). Investigating Upper Secondary School Teachers' Conceptions: Is Mathematical Reasoning Considered Gendered? *International Journal of Science and Mathematics Education*. doi:10.1007/s10763-015-9634-5

Swedish National Agency for Education. (2000). *Ämne - Fysik, gymnasieskolan* [Syllabuses for upper secondary Physics]. Retrieved from http://bit.ly/Fysik2000

Swedish National Agency for Education. (2001). *Programme Manual (GY 2000:19). Programme goal and structures, core subjects, subject index for upper secondary school*. Stockholm, Sweden: Fritzes.

Swedish National Agency for Education. (2003). *Lusten att lära – Med fokus på matematik: Nationella kvalitetsgranskningar 2001-2002.* Stockholm, Sweden: Swedish National Agency for Education.

Swedish National Agency for Education. (2005). *Skolverkets Provbank. Hur används den och vad tycker användarna?* [The National test bank. How it is used and what do the users think?]. Stockholm, Sweden: Swedish National Agency for Education.

Swedish National Agency for Education. (2006). Curriculum for the non-compulsory school system, Lpf 94. Stockholm, Sweden: Fritzes.

Swedish National Agency for Education (2009a). *TIMSS Advanced 2008. Svenska gymnasieelevers kunskaper i avancerad matematik och fysik i ett internationellt perspektiv* [TIMSS Advanced 2008. Swedish upper secondary students' knowledge of advanced mathematics and physics from an international perspective]. Stockholm, Sweden: Fritzes.

Swedish National Agency for Education (2009b). *Hur samstämmiga är svenska styrdokument och nationella prov med ramverk och uppgifter i TIMSS Advanced 2008?* [How aligned are the Swedish policy documents and national tests with the framework and the tasks in TIMSS Advanced 2008?]. Stockholm, Sweden: Fritzes.

Swedish National Agency for Education (2014-05-22). Tabell [Table] 3B. Retrieved from http://bit.ly/SweNAE2011

Swedish Schools Inspectorate. (2010). *Fysik utan dragningskraft. En kvalitetsgranskning om lusten att lära fysik i grundskolan* [Physics without attraction. An evaluation about the lust to learn physics in elementary school]. Report No. 2010:8. Retrieved from http://bit.ly/skolinsp2010

Tall, D. & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limit and continuity. *Educational Studies in Mathematics,* 12(2), 151-169. doi:10.1007/BF00305619

Tasar, M. F. (2010). What part of the concept of acceleration is difficult to understand: the mathematics, the physics, or both? *ZDM-The International Journal on Mathematics Education,* 42(5), 469–482. doi:10.1007/s11858-010-0262-9

Tuminaro, J. (2002). *How Student Use Mathematics in Physics: A Brief Survey of the Literature*. University of Maryland, Physics Education Research Group. Retrieved from http://bit.ly/Tuminaro2002

Umeå University (n.d.). *Manual för nodgrupper – projekt för uppgiftskonstruktion i fysik och matematik* [Manual for working groups – project for task constructions in physics and mathematics]. Unpublished manual. Umeå, Sweden: Umeå University, Department of Educational Measurement.

Verschaffel, L., De Corte, E. & Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction*, 4(4), 273-294.

Wigner, E. P. (1960). The Unreasonable Effectiveness of Mathematics in the Natural Sciences. *Communication on Pure and Applied Mathematics*, 13, 1-14. doi:10.1002/cpa.3160130102

Winter, H. (1978). Geometrie vom Hebelgesetz aus – ein Beitrag zur Integration von Physik- und Mathematikunterricht der Sekundarstufe I [Geometry from the lever rule - a contribution to the integration of physics- and mathematics education at upper secondary school]. *Der Mathematikunterricht* 24 (5), 88-125.

Wittman, M. C. (2002). The object coordination class applied to wave pulses: Analysing student reasoning in wave physics. *International Journal of Science Education*, 24 (1), 97-118. doi:10.1080/09500690110066944

Wyndhamn, J. (1993). *Problem-solving revisited, On school mathematics as a situated practice*, (Doctoral dissertation). Linköping, Sweden: Linköping University.

Wyndham, J., Riesbeck, E. & Schoult, J. (2000). *Problemlösning som metafor och praktik* [Problem solving as metaphor and as practice]. Linköping, Sweden: Institutionen för tillämpad lärarkunskap, Linköpings universitet.

Yackel, E., & Hanna, G. (2003). Reasoning and proof. In J. Kilpatrick, G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 227–236). Reston, VA: National Council of Teachers of Mathematics.

Paper I – Paper IV

Int J of Sci and Math Educ (2016) 14:1133–1152
DOI 10.1007/s10763-015-9636-3

Mathematical Reasoning Requirements in Swedish National Physics Tests

Helena Johansson

Received: 16 February 2015 / Accepted: 16 February 2015 / Published online: 27 March 2015

Abstract This paper focuses on one aspect of mathematical competence, namely mathematical reasoning, and how this competency influences students' knowing of physics. This influence was studied by analysing the mathematical reasoning requirements upper secondary students meet when solving tasks in national physics tests. National tests are constructed to mirror the goals stated in the curricula, and these goals are similar across national borders. The framework used for characterising the mathematical reasoning required to solve the tasks in the national physics tests distinguishes between imitative and creative mathematical reasoning. The analysis process consisted of structured comparisons between representative student solutions and the students' educational history. Of the 209 analysed tasks, 3/4 required mathematical reasoning in order to be solved. Creative mathematical reasoning, which, in particular, involves reasoning based on intrinsic properties, was required for 1/3 of the tasks. The results in this paper give strong evidence that creative mathematical reasoning is required to achieve higher grades on the tests. It is also confirmed that mathematical reasoning is an important and integral part of the physics curricula; and, it is suggested that the ability to use creative mathematical reasoning is necessary to fully master the curricula.

Keywords Creative mathematical reasoning · Mathematical reasoning · Physics tests · Swedish national assessment · Upper secondary school

Introduction

Mathematics and physics are historically closely intertwined, and many mathematical concepts have been developed to describe the laws of nature. How this relationship becomes apparent in a school context and how it might affect students' learning have been discussed from different points of view in educational research. Some of the discussions focus on how physics can influence the learning of mathematics, referred to below as *physics in mathematics*. Other discussions focus on the learning of physics

H. Johansson (✉)
University of Gothenburg, Gothenburg, Sweden
e-mail: helena.johansson@chalmers.se

and are concerned with various aspects of its relation to mathematics, and this is referred to as *mathematics in physics*.

Physics in Mathematics

Blum & Niss (1991) discuss the great value of maintaining a close relationship between mathematics and physics in school because physics can provide good examples for validating mathematical models. In a paper by Doorman & Gravemeijer (2009), the authors discuss the advantage of learning mathematical concepts through mathematical model building and how examples from physics allow for a better understanding of the concepts. Hanna & Jahnke (2002) refer to e.g. Pólya (1954) and Winter (1978) when they discuss the advantage to use arguments from physics in the teaching of mathematical proofs. The importance of using physics to facilitate students' learning of various mathematical concepts is also discussed by Marongelle (2004), who concludes that using events from physics can help students to understand different mathematical representations.

Mathematics in Physics

Tasar (2010) discusses how a closer relation between the school subjects of mathematics and physics can contribute to the understanding of physics concepts and can help ensure that students already understand the mathematical concepts needed in physics. Similar suggestions are done by Planinic, Milin-Sipus, Katic, Susac & Ivanjek (2012), who, in their study of high school students' success on parallel tasks in mathematics and in physics, concluded that students' knowledge is very compartmentalized and that stronger links between the mathematics and physics education should be established. According to Basson (2002), a closer relationship might also decrease the amount of time physics teachers spend on redoing the mathematics students need in physics. Michelsen (2005) discusses how interdisciplinary modelling activities can help students to understand how to use mathematics in physics and to see the links between the two subjects. Redish & Gupta (2009) emphasised the need to understand how mathematics is used in physics and to understand the cognitive thinking of experts in order to teach mathematics for physics more effectively to students. Basson (2002) mentions how difficulties in learning physics not only stem from the complexity of the subject but also from insufficient mathematical knowledge. Bing (2008) discusses the importance of learning the language of mathematics when studying physics. Nguyen & Meltzer (2003) analysed students' knowledge of vectors and concluded that there is a gap between students' intuitive knowledge and how to apply their knowledge in a formal way, which can be an obstacle when learning physics. Tuminaro (2002) analysed a large body of research, and categorised studies concerning students' use of mathematics in physics according to the researchers approach to the area. The four categories are: (i) the observational approach; (ii) the modelling approach; (iii) the mathematical knowledge structure approach and (iv) the general knowledge approach.

Mulhall & Gunstone (2012) describe two major types of physics teacher, the *conceptual* and the *traditional*. Mulhall & Gunstone conclude that a typical teacher in the conceptual group presumes that students can solve numerical problems in physics without a deeper understanding of the underlying physics theories. A typical opinion among teachers in the traditional group is that physics is based on mathematics and that a student develops an understanding of the physics by working with numerical problems. Doorman & Gravemeijer (2009) notice (with reference to Clement, 1985 and Dall'Alba, Walsh, Bowden, Martin, Masters, Ramsden & Stephanou, 1993) that most of the attention in both physics and mathematics is on the manipulations of formulas instead of focusing on why the formulas work.

Learning Physics

When discussing learning in physics, there is, of course, a large body of additional literature that is relevant to consider depending on what questions one is studying. A lot of research about teaching and learning physics has been conducted by what Redish (2003) refers to as the physics education research (PER) community. When studying how individuals learn physics, certain cognitive principles have to be considered (Redish, 2003). This approach is discussed by diSessa (e.g. in 2004), who emphasises the micro levels but from a knowledge-in-pieces perspective. This perspective is not restricted to the learning of physics, but is also applicable in mathematics. According to this micro-perspective, there are many different levels at which a concept can be understood, and contextuality has to be taken into consideration. Thus, in order to understand a student's learning, his or her understanding of a particular concept has to be studied in a variety of different contexts (diSessa, 2004).

Mathematics in the Syllabuses

The upper secondary school in Sweden is governed by the state through the curriculum and the syllabuses. During the last decades, there has been a gradual change towards a stronger focus on process goals, and they are present in the curriculum from 1994 (Swedish National Agency for Education [SNAE], 2006). These shifts are influenced by and similar to international reforms that aim at enriching both mathematics and physics. Content goals are complemented with process goals as those in the National Council of Teachers of Mathematics Standards (NCTM, 2000), and in the Next Generation Science Standards (NGSS Lead States, 2013) where it, e.g. is explicated that 'emphasis is on assessing students' use of mathematical thinking and not on memorization and rote application of problem solving techniques' when high school students use mathematics in physics (NGSS, 2013, HS-PS1-7, Matter and its Interactions). In the framework for PISA 2009, it is emphasised to focus on the mastery of processes and the understanding of concepts (Organisation for Economic Co-operation and Development [OECD], 2009), and in the Trends in International Mathematics and Science Study (TIMSS) framework, the thinking process is explicated as one of the two dimensions to be assessed (Garden, Lie, Robitaille, Angell, Martin, Mullis et al., 2006). For a more comprehensive discussion about the reforms and their backgrounds, see e.g. Boesen, Helenius, Bergqvist, Bergqvist, Lithner, Palm & Palmberg (2014, pp. 73–74). A central part of the reforms concerns reasoning and its central role in problem solving and in the individual's development of conceptual understanding

In the Swedish syllabuses, the aims and objectives of each specific course are detailed and it is indicated what knowledge and skills students are expected to have acquired upon completion of the various courses. According to the general syllabus in physics, the teaching should aim to ensure that the students, e.g. 'develop their ability to quantitatively and qualitatively describe, analyse and interpret the phenomena and processes of physics in everyday reality, nature, society and vocational life and to develop their ability with the help of modern technical aids to compile and analyse data as well as simulate the phenomena and processes of physics' [1] (SNAE, 2000). Mathematics is thus explicitly required when making quantitative descriptions of phenomena and implicitly required when analysing data. In the particular syllabuses for the two courses, Physics A and Physics B, mathematics is mentioned more explicitly. Physics A is a prerequisite for Physics B, and in the latter course, there are higher demands both on the mathematical processing and on the conceptual understanding of physics phenomena (SNAE, 2000).

The literature review shows that there is a significant amount of educational research on the relation between the school subjects of mathematics and physics that support the necessity of different mathematical competencies when learning physics. However, no studies on what type of mathematical reasoning (see 'Theoretical Framework') is required of physics students were found. The impact of mathematical reasoning *on mathematical* learning has been discussed and studied from multiple perspectives. Schoenfeld (1992), for example, points out that a focus on rote mechanical skills leads to poor performance in problem solving. Lesh & Zawojewski (2007) discuss how emphasising low-level skills does not give the students the abilities needed for mathematical modelling or problem solving, neither to draw upon interdisciplinary knowledge. Students lacking the ability to use creative mathematical reasoning thus get stuck when confronted with novel situations and this hamper their possibilities to learn (Lithner, 2008). Since mathematics is a natural part of physics, it is reasonable to assume that the ability to use mathematical reasoning is an integral part of the physics knowledge students are assumed to achieve in physics courses. Therefore, it should be desirable to get a picture of the mathematical reasoning requirements students encounter and need in order to master or fully master the physics curricula.

Theoretical Framework

The definition of mathematical reasoning and the framework that is used for the analyses in this paper were developed by Lithner (2008) through empirical studies on how students engage in various kinds of mathematical activities. The initial purpose of Lithner's studies was to analyse students' rote thinking and how this may lead to learning difficulties in mathematics. As a result, *reasoning* was defined as 'the line of thought adopted to produce assertions and reach conclusions in task solving' (Lithner, 2008, p. 257). *Mathematical reasoning* is used as an extension of a strict mathematical proof to justify a solution and is seen as a product of separate reasoning sequences. Each sequence includes a choice that defines the next sequence, and the reasoning is the justification for the choice that is made. The mathematical foundation of the reasoning can either be *superficial* or *intrinsic*. The accepted mathematical properties of an object are of different relevance in different situations. This leads to a distinction between *surface* properties and intrinsic properties, where the former have little relevance in the actual context and lead to superficial reasoning and the latter are central and have to be

[1] Author's translation

taken into consideration in the given context (Lithner, 2008, p. 260–261). Depending on whether this reasoning is superficial or intrinsic, the framework distinguishes between *imitative reasoning* and *creative mathematical-founded reasoning*. The framework has been used in previous studies to categorise tasks according to mathematical reasoning (e.g. Palm, Boesen & Lithner, 2011) or to categorise actual students' mathematical reasoning in problematic situations (e.g. Sumpter, 2013).

Creative Mathematically Founded Reasoning

Creativity is an expression often used in different contexts and without an unequivocal definition (for a discussion, see Lithner (2008, p. 267–268)). Creativity within the framework that is used in this paper takes the perspective of Haylock (1997) and Silver (1997) in which creativity is seen as a thinking process that is novel, flexible and fluent. *Creative mathematical reasoning*[2] (*CR*) fulfils all of the following criteria: 'i. Novelty. A new reasoning sequence is created or a forgotten one is recreated. ii. Plausibility. There are arguments supporting the strategy choice and/or strategy implementation motivating why the conclusions are true or plausible. iii. Mathematical foundation. The arguments made during the reasoning process are anchored in the intrinsic mathematical properties of the components involved in the reasoning.' (Lithner, 2008, p. 266).

Imitative Reasoning

Imitative reasoning is categorised as *memorised reasoning* (*MR*) or *algorithmic reasoning* (*AR*). The arguments for the chosen solution method (i.e. the reasoning) can be anchored in surface mathematical properties. 'MR fulfils the following conditions: i. The strategy choice is founded on recalling a complete answer. ii. The strategy implementation consists only of writing it down.' (Lithner, 2008, p. 258).

If some kinds of calculations are required to solve the task, there is often no use in remembering an answer. Instead, it is more suitable to recall an algorithm. The term 'algorithm' is used here in a broad sense and refers to all the procedures and rules that are needed to reach the conclusion of a specific type of task, not just the calculations required to reach a conclusion. 'AR fulfils the following conditions: i. The strategy choice is to recall a solution algorithm. The predicted argumentation may be of different kind, but there is no need to create a new solution. ii. The remaining parts of the strategy implementation are trivial for the reasoned, only a careless mistake can lead to failure.' (Lithner, 2008, p. 259).

Depending on the argumentation for the choice of the used algorithm, AR can be subdivided into the three different categories of *familiar algorithmic reasoning* (*FAR*), delimiting algorithmic reasoning and *guided algorithmic reasoning*, e.g. *text-guided* (*GAR*) or person-guided. In this study, only the categories of FAR and GAR are used. FAR fulfils: 'i. The reason for the strategy choice is that the task is seen as being of a familiar type that can be solved by a corresponding known algorithm. ii. The algorithm is implemented.' (Lithner, 2008, p. 262). GAR fulfils: 'i. The strategy choice concerns identifying surface similarities between the task and an example, definition, theorem,

[2] Originally called *creative mathematical founded reasoning*.

rule or some other situation in a text source. ii. The algorithm is implemented without verificative argumentation.' (Lithner, 2008, p. 263).

Local and Global Creative Mathematical Reasoning

Lithner (2008) introduces a refinement of the category (CR) into *local CR* (*LCR*) and *global CR* (*GCR*) that captures some significant differences between tasks categorised as CR. This subdivision has been further elaborated by other scholars, e.g. Boesen, Lithner & Palm (2010) and Palm et al. (2011). In LCR, the reasoning is mainly MR or AR but contains a minor step that requires CR. If instead there is a need for CR in several steps, it is called GCR, even when some parts contain AR and/or MR.

Non-mathematical Reasoning

The analytical framework in this paper introduces an additional category called *non-mathematical reasoning* (*NMR*). This consists of those tasks that can be solved by using just the knowledge of physics. Physics knowledge here refers to relations and facts that are discussed in the syllabuses and textbooks of the physics courses but not in the mathematics courses, for example, the fact that angle of incidence equals angle of reflection. In the same way, the concept of mathematics refers to school mathematics that is introduced in mathematics courses for students at upper secondary school or the mathematics assumed to already be known according to the curricula.

Research Question

By analysing the mathematical reasoning required to solve tasks in national physics tests, the idea is to capture the mathematical reasoning that is required to master or fully master the physics curricula. It is explicated in the physics syllabuses that the use of mathematics is incorporated in the goals and that the national tests are the government's way of concretising the physics curricula. Based on the definitions in the definitions of the theoretical framework described above, the following research questions were asked:

- Is mathematical reasoning required of upper secondary students to solve national physics tests from the Swedish national test bank?
- If mathematical reasoning is required, what is the distribution of physics tasks requiring CR compared to tasks that are solvable with IR?

Physics Tests from the National Test Bank

About 12 % of all students in upper secondary school in Sweden are enrolled in the Natural Science Programme or the Technology Programme (SNAE, 2011). In both programmes, the course Physics A is compulsory whereas the more advanced course—Physics B—is elective. The aim of the physics courses is that the students should attain various goals specified in the syllabuses. Written tests are commonly used as an assessment of the students' achievements, and a student's grade in a course depends on how well the student has achieved the goals for the course (SNAE, 2000). The descriptions of the goals and the different grade levels are quite brief in the syllabuses, and the intention is that the syllabuses and curriculum should be processed, interpreted and refined locally in each school. In order to accomplish equivalent assessment in physics, the SNAE provides assessment supports, including the National Test Bank in Physics. In this way, the physics tests can be considered as the government's concretisation of the syllabuses for physics. The character and design of the tasks in the tests stress what is covered in the taught curriculum. The tests also influence the teachers' interpretation of the syllabuses, which by extension, stresses what the students focus on (Ministry of Education and Research, 2001; SNAE, 2003).

The material in the National Test Bank is classified and can be accessed via the Internet only by authorised users. The material consists of single tasks to choose from or complete tests that comprise the goals for Physics A or Physics B. In total, there are 847 tasks to choose from and 16 complete tests for each of the Physics A and Physics B courses, all classified. The first tests are from 1998 and the latest is from spring 2011. Besides the classified examples, there are five tests for each course that are open for students to practice on. These give the students an idea of what the tests look like and what is required when taking a test (Department of Applied Educational Science, 2011). As opposed to national tests in mathematics, the teachers are not obligated to use the test from the National Test Bank in physics. However, it is common that these tests are used as a final exam in the end of the physics courses (SNAE, 2005).

Since the beginning of the national testing programme, there has been a change in the design of the tasks on the tests. In the beginning, there was more or less only one correct solution to each task. This has evolved into a higher degree of open tasks that can be solved using different approaches. For the past 10 years, the final task has been an "aspect-task" that is assessed according to the achieved level in different assessment groups. These aspect tasks include initial parts that are easily accessible for most students and parts that are a challenge designed for more proficient students. The task is designed to be easy to start with, but it should also include a challenge to more proficient students. The first 3 years of the testing programme (1998–2000), there was an experimental part included in the tests, but this part is not included in the analysis in this study.

Method

This study analysed the December 1998, May 2002, December 2004, May 2005 and December 2008 tests for the Physics A course and the May 2002, May 2003, May 2005, February 2006 and April 2010 tests for the Physics B course. The first tests chosen were the unclassified tests so that examples could be discussed in the present article. These tests are unclassified by the National Educational Agency to serve as representative interpretations of the syllabuses and the curriculum. To have five tests from each course, the remaining tests were randomly selected among the classified tests.

Categorisation of Mathematical Reasoning Requirements

To categorise physics tasks according to reasoning requirements, solutions to respective task are required. Whether a task is solvable by IR or if the solution requires CR depends on the educational history of the solver (in this case, the test taker) (cf. Björkqvist, 2001; English & Sriraman, 2010; Lesh & Zawojewski, 2007; Schoenfeld, 1985; Wyndham, Riesbeck & Schoult, 2000). The required reasoning refers to what kind of reasoning is sufficient to solve a task, and the framework described above allows for a determination of this.

The solutions used in the analysing procedure were constructed by the researcher. These solutions were determined to be plausible student solutions based on the researcher's experience as a physics teacher together with access to the solution manuals in which proposed solutions are given. Some of the solutions in the manual are authentic student solutions because several of the tasks had been tested on real students before the tasks were included in a test.

Because no students were involved in the present study, there was no actual learning history to consider. Studies on mathematics education suggest that most of the learning activities consist of students working with their textbooks (SNAE, 2003). In an evaluation of physics education in lower secondary school, it was found that the teaching is guided by the textbooks (Swedish Schools Inspectorate, 2010). In addition, The Ministry of Education and Research (2001) has discussed the fact that textbooks and assessments are seen as two of the most important tools in mathematics education. In a qualitative study of a physics class, Engström (2011) showed that the textbook still plays an important role in guiding the education, and the TIMSS Advanced 2008 report showed that teachers mostly use the textbook in physics courses to choose and solve problems from (SNAE, 2009a). Based on the findings described above, the students' prior knowledge in physics and mathematics in this study was equated with the content of the textbooks used in their courses. There are, of course, other factors that play a part in individual students' previous experience, including tasks discussed during classes and/or experience of physical principles outside the classroom. The simplification used in this study was necessary due to the complexity of students' educational history and was reasonable according to the discussion above.

This study considered textbooks in both mathematics and physics. When taking the tests, the students are allowed to use a handbook designed for the physics courses in upper secondary school. The access to formulas and definitions in this handbook had to be taken into account when analysing the tasks in this study. The textbooks and the handbook were chosen among the books commonly used in the physics courses in upper secondary school. The books used for categorisation of the tasks in Physics A tests were 'Ergo Fysik A' (Pålsgård, Kvist & Nilson, 2005a) and 'Matematik 3000 Kurs A och B' (Björk & Brolin, 2001). For tests in Physics B, 'Ergo Fysik B' (Pålsgård, Kvist & Nilson, 2005b) and 'Matematik 3000 Kurs C och D' (Björk & Brolin, 2006) were used. The handbook chosen was 'Tabeller och formler för NV- och TE- programmen' (Ekbom, Lillieborg, Larsson, Ölme & Jönsson, 2004). Even if not all students in the Swedish upper secondary school are using the books above, they are a reasonable assumption for the education history of the average student. The procedure for analysing the tasks was given by the chosen framework, and an analysis sheet was used to structure the procedure. The steps comprised in the procedure are outlined in Table 1 and are used earlier in e.g. Palm et al. (2011).

Table 1 Detailed outline of the analysis procedure

Step I. Analysis of the assessment task—answers and solutions	Step II. Analysis of the assessment task—task variables
The first step in the procedure consisted of constructing a plausible student solution. The solution was then looked at from a mathematical perspective and categorised according to relevant mathematical subject areas that were required for the solution, e.g. asking if the solution included working with formulas, algebra, diagrams, solving equations, etc. Tasks with solutions not including any mathematical object were identified and categorised as NMR tasks. This categorisation is an addition to the original procedure used in previous studies. Mathematical objects refer to entities to which mathematics is applied. The first step also includes the identification of 'real-life' events in the task formulation. This identification is relevant because a described situation in the task could give a clue to a known algorithm that solves the task (see the Weightlifter (a) example (Table 4))	The next step in the procedure was to analyse the solution according to different task variables. The first variable was the explicit formulation of the assignment. The second variable was *what* information about the mathematical objects was given explicitly in the task compared to what information the students need to obtain from the handbook or that they have to assume in order to reach a solution. The third task variable concerns *how* the information was given in the task, e.g. numerically or graphically or whether it was interwoven in the text or explicitly given afterwards. The task could also include keywords, symbols, figures, diagrams or other important hints the student can use to identify the task type and which algorithm to use. These features were gathered into the fourth task variable
Step III. Analysis of the textbooks and handbook—answers and solutions	Step IV. Argumentation for the requirement of reasoning
The third step in the analysis process focused on the textbooks and the handbook. Formulas used in the solution algorithm were looked for in the handbook, and the available definitions were compared to the constructed solution to the task. The textbooks were thoroughly looked through for similar examples or exercises that were solved by a similar algorithm. The theory parts in the textbooks were also examined to see whether they contained any clues as to solve the task	In the final step, the researcher produced an argument, based on steps I to III, for the categorization of the reasoning requirement for every task. In order to be categorised as FAR, there must have been at least three tasks considered as similar in the textbooks. It could then be assumed that the students will remember the algorithm, which might not be the case if there are fewer occasions. Three similar tasks were found to be an appropriate number in the study by Boesen et al. (2010). If the task was similar to a formula or definition given in the handbook, it was assumed that the student could use this as guidance in order to solve the task. Thus, only one similar and previously encountered example or exercise was required for tasks categorised as requiring GAR. To be categorised as requiring MR, tasks with the same answer or solution should have been encountered at least three times in the textbooks. It was then assumed that the student could simply write the same answer on the test. If none of the above reasoning types were sufficient for solving the task and there was a need to consider some intrinsic mathematical property, the task was categorised as requiring some kind of CR

Validity and Reliability

The resulting categorisation of tasks, theoretically established according to the abovementioned procedure, is only meaningful if it represents the reasoning actually used by students while solving the tasks. Meaningful representation can be achieved with the well-documented criteria required for each category along with a routine for agreement and discussions about the categorisation. Higher reliability could also be reached with a less complex phenomenon, e.g. by defining creative mathematical reasoning as solutions consisting of more than three steps. However, this would give a low validity for the meaning of creative mathematical reasoning.

The validity of the analysis is dependent both on the appropriateness of the procedure used for the categorisation and how closely the categorisations resemble students' actual reasoning. The appropriateness is argued for above, and an argument for concordance is based on results from a study by Boesen et al. (2010). In that study, real students' actual mathematical reasoning used to solve tasks on mathematics tests were compared to the theoretically established reasoning requirements for the same tasks based on the same procedure that was used in the present study. It was shown that only 3 % of the tasks were solved with less creative reasoning than what was judged to be required, and 4 % of the tasks were either solved with more creative reasoning or not solved at all. These results indicate that the categorisation of reasoning, as described here, provides meaningful results. The construction of a plausible student solution is one of the four steps in the analysing procedure, and the author's experience with physics students and physics tests can be considered similar to the experience with mathematics students and mathematics tests in Boesen et al. (2010). Therefore, previous results demonstrating the method's validity can be considered to be valid for the present study. The categorisation of all tasks in the present study was made by the author. During the analysis process, there were tasks where the categorisation was straightforward and tasks where the categorisation could be considered as borderline cases. Typical examples of the different kinds of categorisation were continuously discussed in a reference group consisting of a mathematics education researcher well familiar with the analysis procedure and a mathematician. All difficult categorisations were discussed in the group, thus no inter-reliability estimate was calculated.

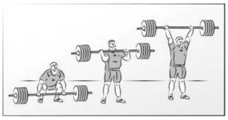
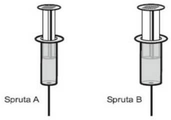
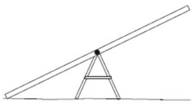
Data and Analyses

The tasks in Table 2 were chosen to represent and illustrate the different types of analysis and the categorisations of the physics tasks. The idea was that the required reasoning would be represented by the constructed solutions. All of the tasks are chosen from publicly available national tests. Normally, subtasks are treated separately because the task variables and the analysis of the textbooks can be different. The first three tasks are examples where no hesitation concerning the categorisations occurred. The three subsequent examples are of tasks were the analysis was not as straightforward. The analyses are displayed in detail in Tables 3, 4, 5 and 6.

Results

The analysis showed that mathematical reasoning was required when solving physics tasks. Of the 209 analysed tasks, there were 76 % that required

Table 2 Six examples of the tasks that were analysed

Task 1 (“The Weightlifter (a)”)	**Task 2 (“The Weightlifter (b)”)**
A weightlifter is lifting a barbell that weighs 219 kg. The barbell is lifted 2.1 m up from the floor in 5.0 s.  a) What is the average power the weightlifter develops on the barbell during the lift?	A weightlifter is lifting a barbell that weights 219 kg. The barbell is lifted 2.1m up from the floor in 5.0 s. b) What is the average power the weightlifter develops on the barbell when he holds it above his head for 3.0 s?
Task 3 (“The Syringes”).	**Task 4 (“Charges on a thread”)**
A patient is going to get an injection. The medical staffs are reading in the instructions that they are supposed to use a syringe that gives the lowest pressure as possible in the body tissue. Which of the syringes A or B shall the staff choose if the same force, F, is applied and the injection needles have the same dimensions?” *Argue for the answer* 	In order to determine the charge on two small, light silver balls, the following experiment was conducted. The balls, which were alike, weighed 26 mg each. The balls were threaded onto a nylon thread and were charged in a way that gave them equal charges. The upper ball levitated freely a little distance above the other ball. There was no friction between the balls and the nylon thread. The distance between the centres of the balls was measured to 2.9 cm. What was the charge on each of the balls?” 
Task 5 (“The seesaw”)	**Task 6 (“Walking in water”)**
How can Lars, 70 kg, and his son Anton, 28 kg, place themselves on a 3.5 m long seesaw so that it stays in equilibrium? 	You are walking out into the water at a beach with a stony bottom. In the beginning, it hurts very much under your feet when you are walking on the stones, but when the water gets deeper it starts to feel less. When the water reaches you up to your chest, the stones do not feel painful anymore. Explain this. Svar: ___

mathematical reasoning. The distribution of tasks categorised as requiring CR (CR tasks) and tasks solvable with IR (IR tasks) were a bit unbalanced. Of the tasks requiring mathematical reasoning, 46 % were CR tasks whereas the remaining ones were IR tasks (Table 7).

The result also showed some differences in the categorisation with respect to the Physics A and Physics B courses. There were slightly more NMR tasks and CR tasks in the Physics B tests than in the Physics A tests. A more distinct difference was seen among the IR tasks, with the greater number of these tasks in Physics A tests (Table 7).

A majority of the IR tasks (78 %) were solvable with FAR and the rest were solvable with GAR. The CR tasks were separated into LCR and GCR. In general, Physics B tests consisted of more GCR tasks than Physics A tests, and the amount of LCR tasks was almost the same (Table 8). When comparing tests from different years, the analysis showed a notable variation in the

Table 3 The first two steps in the analysis procedure for tasks 1 to 4

	I. Analysis of the assessment task—answers and solutions	II. Analysis of the assessment task—task variables
Task 1	A typical solution from an average student could be derived by the relation between power and the change of energy over a specific period of time. In this task, the change in energy is the same as the change of potential energy for the barbell. Multiply the mass of the barbell by the acceleration of gravity and the height of the lift and then divide by the time to get the power asked for. The mathematical subject area is identified as algebra, in this case, working with formulas. The identification of the situation to lift a barbell can trigger the student to use a certain solution method and is, therefore, included in this analysis as an identified 'real-life' situation	The assignment is to calculate the average power during the lift. The mass of the barbell, the height of the lift and the time for the lift are all considered as mathematical objects. As mentioned above, an object is the entity one is doing something with. In this example, all of the objects are given explicitly in the assignment in numerical form. In the presentation of the task, there is also an illustrative figure of the lift
Task 2	It is not necessary to use any mathematical argumentation in order to solve this task, and solution can be derived on physical reasoning alone. There is no lifting and, therefore, no work is done, and this means that no power is developed. This task is a typical example of an analysis resulting in the NMR categorisation	Not a step to consider as this task is categorised as NMR
Task 3	To solve this task, the student can use the relation between pressure, force and area ($p = F / A$). Neglecting the hydrostatic pressure from the injection fluid, if the force applied to the syringe is the same then it is the area of the bottom that affects the pressure. The larger the area, the lower the pressure. The staff should choose syringe B. The mathematical subject area is identified as algebra, such as to work with formulas and proportionality	The assignment is to choose which syringe that gives the minimum pressure and to provide an argument for this choice. Only the force is given as a variable, and this is represented with a letter. Key words for the students can be *force* and *pressure*. The situation is illustrated with a figure in which it appears that syringe B has a greater diameter than syringe A
Task 4	To derive a solution, the forces acting on the upper ball must be considered. Because it is levitating freely, it is in equilibrium and, according to Newton's first law, the net force on the ball is zero. The forces acting on the ball are the downward gravitational force, $F = \mathrm{mg}$, and the upwards electrostatic force from the ball below, $F = k{\cdot}Q_1Q_2 / r^2$. Setting these expressions equal to each other and solving for Q_1 (and assuming that $Q_1 = Q_2$) will give the charges asked for. The mathematical subject area is identified as algebra, such as to work with formulas and to solve quadratic equations	The assignment is to calculate the charges on the balls. The mass of the balls and the distance between their centres are mathematical objects given numerically and explicitly in the assignment. The information about the charges' equal magnitude is textual and is a part of the description of the situation. There is also a figure of the balls on the thread illustrating the experiment

proportions of the different mathematical reasoning types. There was no consistency among the tests with respect to this analysis (Table 8).

Table 4 The last two steps in the analysis procedure for tasks 1 to 4

III. Analysis of the textbooks and handbook—Answers and solutions	IV. Argumentation for the requirement of reasoning
Handbook: Formulas for power, $P = \Delta W/\Delta t$, with the explanation 'ΔW = the change in energy during time Δt'; for 'work during lift', W_l = mg h, with the explanatory text, 'A body with weight mg is lifted to a height h. The lifting work is …'; and for potential energy with the text 'A body with mass m at a height h over the zero level has the potential energy W_p = mg h' Mathematics book[a]: Numerous examples and exercises of how to use formulas, e.g. on pages 28–30 Physics book[b]: Power is presented as work divided by time, and in on example, work is exemplified as lifting a barbell. An identical example is found on page 130. An example of calculating work during a lift in relation to change in potential energy is found on page 136. Exercises 5.05 and 5.10 are solved by a similar algorithm	The analysis of the textbooks shows that there are more than three tasks similar to the task being categorised with respect to the task variables, and these tasks can be solved with a similar algorithm. As mentioned in the method section, if the students have seen tasks solvable with a similar algorithm at least three times, it is assumed that they will remember the solution procedure. This task is then categorised as solvable using IR, in this case FAR
Not a step to consider as this task is categorised as NMR	Not a step to consider as this task is categorised as NMR
Handbook: The relation $p = F/A$ is defined Mathematics book: Proportionalities are discussed and exemplified but are not used for general comparisons Physics book: One example about how different areas affect the pressure and one exercise that is solved in a similar way by using a general comparison between different areas and pressure	There is only one example and one exercise that can be considered similar with regard to the task variables and the solution algorithm. The formula is in the handbook, but there has to be some understanding of the intrinsic properties in order to be able to use the formula in the solution. This task is, therefore, considered to require some CR, in this case GCR, in order to be solved
Handbook: Coulomb's law, $F = k\,Q_1Q_2 / r^2$, with explanation 'r = distance between the charges and … k = … ≈8.99 10^9 N $m^2/(As)^2$' Mathematics book: Numerous examples and exercises of how to use formulas, e.g. on pages 28–30, and of solving quadratic equations on page 269 Physics book: Coulomb's law is introduced and exemplified, and there are at least three exercises of calculating the charge on different objects using this law. One example is of a levitating charge (page 227), but in this case in a homogeneous electrical field instead of due to the electrostatic force from another charged particle. Two exercises of similar situations as in the example. Newton's first law is formulated in the theory text (page 91) where it is shown that the net force has to be zero if an object for example is at rest, and this relation is used on several different occasions in the book. The gravitational force is introduced on pages 92 and is then used throughout the book	Considering the mathematical reasoning, there are more than three examples or exercises in the textbooks where the same algorithm has been used, i.e. to put two expressions equal to each other and then solve for one unknown variable, including taking the square root. However, there are not three or more examples considering the physics context. To solve the task, the student must first identify the force situation in order to know which expressions to equate. After having discussed this task in the reference group, it was concluded that analysing the physics context is not a part of the mathematical reasoning. Although mathematical reasoning is necessary to be able to solve the task, it is not sufficient, and although the mathematical reasoning can be considered as some kind of algorithmic, the task was categorised as requiring LCR, where the minor step is to analyse the physics

[a] The mathematics textbook in all examples is Björk & Brolin (2001)

[b] The physics textbook in all examples is Pålsgård et al. (2005a)

Table 5 The first two steps in the analysis procedure for tasks 5 and 6

	I. Analysis of the assessment task—answers and solutions	II. Analysis of the assessment task—task variables
Task 5	This task can be solved using the equilibrium of torque (moment of force), $M = Fr$, and the knowledge that the torque, with respect to Anton, must have the same magnitude as the torque with respect to Lars. The forces that act on the seesaw are of the same magnitudes as the gravitational forces, $F = \mathrm{mg}$, on Lars and Anton, respectively. Assuming that Anton is placed 1.60 m from the rotation axis, one gets the equation $F_{\mathrm{Lars}}\, r = \mathrm{F}_{\mathrm{Anton}}\, 1.60$, which will give the position Lars must be in when the equation is solved. As in the examples above, the mathematical subject area was identified as algebra, more specifically to work with formulas and equations. A seesaw is a real-life situation often used as an example in mechanics and, therefore, was included in the analysis	The assignment is to show where on the seesaw Anton and Lars can sit when it is in equilibrium. Mathematical objects that are given numerically in the assignment are the masses of Anton and Lars. In addition, the total length of the seesaw is given and there is a picture of a seesaw without any people on it
Task 6	To solve this task, the students are supposed to refer to Archimedes' principle. The greater the volume of the body under the water, the lager the buoyant force from the water. Assuming the body is in equilibrium at each step, the larger the buoyant force becomes, the smaller the normal force from the stones becomes and thus there is less pressure from the stones. Therefore, it hurts less when the water level reaches higher on the body. This relation can be argued for using the formulas for Archimedes' principle, formulas for pressure and the equilibrium of forces. The mathematical area could then be considered to involve formulas and proportionality. Following the solution proposal and the scoring rubric provided with the test, however, there is no need to use any mathematical relations or formulas to argue for the answer	The assignment is to explain why it does not hurt as much when you are in deeper water. No mathematical objects are given explicitly in the task. The situation refers to a real-life event of walking in water. Bathing is a common situation referred to when discussing Archimedes' principle. The depth of the water is also indicated in the assignment as important

Discussion and Implications

The national tests are used in the present study to represent the mathematical reasoning that is required to master or fully master the physics courses according to the syllabuses and curriculum. Because of the way the national physics tests are constructed, students that have fully mastered the physics curricula should have the ability to solve any of the tests for the related course. The fact that slightly less than one third of the tasks on some of the 10 tests in this study require CR (Table 8) does not weaken the overall result that CR is significant for fully mastering the physics curricula.

The fact that a majority of the tasks require mathematical reasoning shows that the ability to reason mathematically is an important competence and an integral part when taking physics tests. Mathematical reasoning is defined as a process to reach

Table 6 The last two steps in the analysis procedure for tasks 5 and 6

III. Analysis of the textbooks and handbook—answers and solutions	IV. Argumentation for the requirement of reasoning
Handbook: Formula for Torque, $M = Fr$, with explanatory text 'r is the perpendicular distance from the rotation axis to the line of action of the force. At equilibrium $\sum Fr = \sum M = 0$' together with a figure of M around a rotation axis with F and r marked Mathematics book: Numerous examples and exercises on how to use formulas (e.g. on pages 28–30) and how to solve equations Physics book: The relation for torque is formulated with words in the theory text. When introducing torque, the theory also refers to a seesaw both in text and with images (page 105). Two examples use the formula for torque as defined in the handbook. One of the examples is similar to this task except that one does not have to assume any distance. There are some exercises using a similar algorithm, but these are for calculating masses (via force) from given distances instead of distances from given masses	The algorithmic procedure to solve a task involving a seesaw has been seen both in the theory text and in the examples. There are plenty of exercises for how to handle expressions and solve equations with one unknown variable. The difference in this case is that none of the distances are given in the task. There are, therefore, two unknown variables in the expression, and one of the distances has to be assumed, by using the information about the total length of the seesaw. After discussion about this task, it is categorised as requiring LCR. The minor step in this case is to realise that one has to make an assumption of one of the distances in order to be able to solve the task, and this is regarded as demanding some intrinsic mathematical understanding
Handbook: Archimedes' principle is formulated with the words, 'The buoyant force on an object is equal to the weight of the displaced fluid' that appear on the same page as the formula for pressure, $p = F/A$ Mathematics book: Numerous examples and exercises on how to use formulas, e.g. on pages 28–30, and exercises on proportionality on pages 73 and 75, but these are not used for general comparison Physics book: Archimedes' principle is formulated with words and as an expression (page 171), and there is one example that relates volume to the buoyant force	Following the scoring rubric of what is demanded of a student to solve this task, there is no need to refer to the formulas or to use them to argue for the given explanation. The student needs to mention Archimedes' principle and that the buoyant force increases when the volume of the body in the water increases, but he/she does not need to explain why or show how the volume increase is related to the force increase. They also have to mention something about how this increased buoyant force decreases the normal force, but according to the scoring rubric, there is no need to use the relation for pressure to show why this decreased normal force makes it hurt less. The space given to write the answer also indicates that a few lines are sufficient as an answer. After discussing this task and the minimum solution that is required of a student, it is decided that the reasoning is mainly physical and that mathematical reasoning is not necessary to solve this task. It is then categorised as solvable with NMR

conclusions when solving tasks. When students have the ability to use creative mathematical reasoning, they know how to argue and justify their conclusions and they can draw on previous knowledge. As it is not enough to only use IR to solve a majority of the tasks in a test, but especially CR is required, a creative mathematical reasoning competency can be regarded decisive when students develop their physics knowledge. At first glance, it might be reasonable to assume that CR is required to get a higher grade on a test, and this hypothesis was tested in a follow-up study (Johansson, 2013). It was shown in that study that in order to get one of the higher grades, students

Table 7 Categorisation results, overview

	Number of tasks	NMR %	CR %	IR %
Physics A	103	21	33	46
Physics B	106	26	38	36
Total	209	24	35	41

had to solve tasks requiring CR in five out of eight national physics tests. For the three tests not requiring CR, students' actual results on these three tests were compared to which tasks they had solved, and it was concluded that even though it was possible to get a higher grade without using CR, this rarely occurred.

The conclusion that CR is vital to students' development of physics knowledge is based on the fact that Swedish national physics tests are a concretisation of the goals in the syllabuses and in the curriculum of what should have been achieved after completing the physics courses. The goals and the subject descriptions in the Swedish syllabuses and curriculum of what it means to know physics are quite rich and highly in accordance with the content and cognitive domains in the TIMSS Assessment framework (Garden et al. 2006; SNAE, 2009b). Although this study deals with the Swedish settings, the alignment with TIMSS suggests that these results are relevant to an international context.

As mentioned in the section 'Learning Physics', individuals' understanding of the relevance of different concepts in various contexts has to be examined in order to discuss what has been learned. The present study does not claim anything about individual students' learning. However, it is shown that mathematical reasoning in

Table 8 Categorisation results, detailed

	Number of tasks	NMR *n*	NMR %	FAR *n*	GAR *n*	IR %	LCR *n*	GCR *n*	CR %	GCR %	IR + LCR %
Physics A Dec 1998	20	3	15	6	6	60	4	1	25	5	80
Physics A May 2002	20	4	20	3	3	30	5	5	50	25	55
Physics A Dec 2004	19	4	21	7	1	42	2	5	37	26	53
Physics A May 2005	19	5	26	6	2	42	4	2	32	11	63
Physics A Dec 2008	25	6	24	12	1	52	4	2	24	8	68
Total Physics A	103	22	21	34	13	46	19	15	33	15	64
Physics B May 2002	18	2	11	7	0	39	5	4	50	22	67
Physics B May 2003	19	5	26	8	1	47	3	2	26	11	63
Physics B May 2005	23	7	30	4	3	30	5	4	39	17	52
Physics B Feb 2006	23	10	43	8	0	35	2	3	22	13	43
Physics B April 2010	23	4	17	5	2	30	4	8	52	35	48
Total Physics B	106	28	26	32	6	36	19	21	38	20	54
Total	209	50	24	66	19	41	38	36	35	17	59

general and CR in particular is vital when students solve tasks in physics. Since CR is based on an intrinsic understanding of a concept and the ability to use the concept in novel situations, this is in line with diSessa's (2004) view of learning as a development of the ability to use a concept in shifting contexts.

This study is situated within the 'Mathematics in Physics' research field (see 'Introduction'). The literature suggests how mathematical knowledge influences the learning of physics and the importance of understanding how mathematics is used in physics. From the results in this study, mathematical reasoning can be concluded to be a central aspect of this mathematical knowledge. In particular, CR is decisive to fully master the physics curricula. To achieve this CR competency, students must be provided opportunities to develop and practice creative mathematical reasoning. This could take place both in the physics classes and in the mathematics classes. According to references discussed in the 'Introduction' as well as in the 'Method' sections, it is common that students in physics classes solve routine tasks and focus on manipulations of formulas instead of focusing on the conceptual understanding. Similar conclusions are drawn regarding the mathematics classes; it is found that the focus is on algorithmic procedures and no extensive opportunities to develop different kinds of CR are provided (e.g. Boesen et al., 2014).

It is known that tests have an indirect role for students learning, both as formative, when students get feedback on their solutions, and as summative, when the character of the tasks give students indications of what competences are sufficient for handling mathematical tasks. Analyses of teacher-made mathematics tests have shown that these focused largely on IR, in contrast to the national mathematics tests, which had a large proportion of tasks requiring CR (Palm et al., 2011). In view of the result of Boesen et al. about the situation in the mathematics classes and of Palm et al. about teacher-made mathematics tests, it is reasonable to assume that the teacher-made tests represent respective mathematics teacher's practice. This assumption is further supported by one of the results in Boesen (2006), where teachers indirectly claim that their assessments align with the instructional practice. In the same way, it is assumed that physics teachers' practices are reflected in the physics tests they construct. As discussed above, the classroom situations in physics and mathematics can be considered similar. Thus, a reasonable conclusion is that there is a similar discrepancy regarding physics tests, i.e. that there is a larger proportion of CR in the national physics tests than there is in the teacher-made tests.

From the discussion above, it seems that although the intense efforts that have been made to change practice through policy changes, discussed in the 'Mathematics in the Syllabuses' section, students are provided limited opportunities to develop the creative mathematical reasoning competency that is required to fully master the physics curricula. It can be assumed that the implementation work of the new curricula in school, concretised through national tests, has not worked as intended. The importance of the relation between mathematics and physics has been known for a long time. What has been found in this study is the fact that the ability to mathematically argue and reason is decisive in order to fully master the physics curricula, and this should have implications on how the education is organised and carried out.

References

Basson, I. (2002). Physics and mathematics as interrelated fields of thought development using acceleration as an example. *International Journal of Mathematical Education in Science and Technology, 33*(5), 679–690. doi:10.1080/00207390210146023.

Bing, T. (2008). *An epistemic framing analysis of upper level physics students' use of mathematics* (Doctoral dissertation). Retrieved from http://bit.ly/Bing2008.

Björk, L.-E. & Brolin, H. (2001). *Matematik 3000, kurs A och B [Mathematics 3000, course A and B]*. Stockholm, Sweden: Natur och Kultur.

Björk, L.-E. & Brolin, H. (2006). *Matematik 3000, kurs C och D* [Mathematics 3000, course C and D]. Stockholm, Sweden: Natur och Kultur.

Björkqvist, O. (2001). Matematisk problemlösning [Mathematical problem solving]. In B. Grevholm (Ed.), *Matematikdidaktik: Ett nordiskt perspektiv* (pp. 115–132). Lund, Sweden: Studentlitteratur.

Blum, W. & Niss, M. (1991). Applied mathematical problem solving, modelling, applications, and links to other subjects—state, trends and issues in mathematics instruction. *Educational Studies in Mathematics, 22*(1), 37–68. doi:10.1007/BF00302716.

Boesen, J. (2006). *Assessing mathematical creativity* (Doctoral dissertation). Umeå, Sweden: Umeå University.

Boesen, J., Lithner, J. & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning student use. *Educational Studies in Mathematics, 75*(1), 89–105. doi:10.1007/s10649-010-9242-9.

Boesen, J., Helenius, O., Bergqvist, E., Bergqvist, T., Lithner, J., Palm, T. & Palmberg, B. (2014). Developing mathematical competence: From the intended to the enacted curriculum. *Journal of Mathematical Behavior, 33*, 72–87. doi:10.1016/j.jmathb.2013.10.001.

Clement, J. (1985). Misconceptions in graphing. In L. Streefland (Ed.), *Proceedings of the ninth international conference for the psychology of mathematics education* (pp. 369–375). Utrecht, The Netherlands: Utrecht University.

Dall'Alba, G., Walsh, E., Bowden, J., Martin, E., Masters, G., Ramsden, P. & Stephanou, A. (1993). Textbook treatments and students' understanding of acceleration. *Journal of Research in Science Teaching, 30*(7), 621–635. doi:10.1002/tea.3660300703.

Department of Applied Educational Science (2011-10-06). *Some earlier given test in physics.* Retrieved from http://bit.ly/DepAES2011.

diSessa, A. A. (2004). Contextuality and coordination in conceptual change. In E. Redish & M. Vicentini (Eds.), *Proceedings of the international school of physics* "*Enrico Fermi:*" *Research on physics education* (pp. 137–156). Amsterdam, Netherlands: ISO Press/Italian Physics Society.

Doorman, L. M. & Gravemeijer, K. P. E. (2009). Emergent modeling: Discrete graphs to support the understanding of change and velocity. *ZDM-The International Journal on Mathematics Education, 41*, 199–211. doi:10.1007/s11858-008-0130-z.

Ekbom, L., Lillieborg, S., Larsson, S., Ölme, A. & Jönsson, U. (2004). *Tabeller och formler för NV- och TE-programmen* [Tables and formulas for the NV- and TE- programs] (5th ed.). Stockholm, Sweden: Liber.

English, L. & Sriraman, B. (2010). Problem solving for the 21st century. In L. English & B. Sriraman (Eds.), *Theories of mathematics education* (pp. 263–290). Berlin, Heidelberg: Springer.

Engström, S. (2011). *Att vördsamt värdesätta eller tryggt trotsa. Gymnasiefysiken, undervisningstraditioner och fysiklärares olika strategier för energiundervisning* [To respectfully value or confidently defy. Upper secondary physics, teaching traditions and physics teachers' different strategies for energy education] (Doctoral dissertation). Retrieved from http://bit.ly/EngstromDiss2011.

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O., Mullis, I. V. S. et al. (2006). *TIMSS advanced 2008 assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Hanna, G. & Jahnke, H. N. (2002). Another approach to proof: Arguments from physics. *ZDM-The International Journal on Mathematics Education, 34*(1), 1–8. doi:10.1007/BF02655687.

Haylock, D. (1997). Recognising mathematical creativity in schoolchildren. *ZDM-The International Journal on Mathematics Education, 29*(3), 68–74. doi:10.1007/s11858-997-0002-y.

Johansson, H. (2013). *Mathematical Reasoning in Physics Tests Requirements. Relations, Dependence* (Licentiate Thesis). Göteborg, Sweden: University of Gothenburg.

Lesh, R. & Zawojewski, J. (2007). Problem solving and modeling. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 763–804). Charlotte, NC: Information Age.

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics, 67*(3), 255–276. doi:10.1007/s10649-007-9104-2.

Marongelle, K. A. (2004). How students use physics to reason about calculus tasks. *School Science and Mathematics, 104*(6), 258–272. doi:10.1111/j.1949-8594.2004.tb17997.x.
Michelsen, C. (2005). Expanding the domain—variables and functions in an interdisciplinary context between mathematics and physics. In A. Beckmann, C. Michelsen & B. Sriraman (Eds.), *Proceedings of the 1st international symposium of mathematics and its connections to the arts and sciences* (pp. 201–214). Schwäbisch Gmünd, Germany: The University of Education.
Ministry of Education and Research (2001). *Samverkande styrning: om läroplanerna som styrinstrument* [Interacting governance. About the curricula as means for policy change] Ministry publications series Ds 2001:48. Retrieved from http://bit.ly/DS2001.
Mulhall, P. & Gunstone, R. (2012). Views about learning physics held by physics teachers with differing approaches to teaching physics. *Journal of Science Teacher Education, 23*(5), 429–449. doi:10.1007/s10972-012-9291-2.
National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.
Next Generation Science Standards Lead States (2013). *Next generation science standards: For states, by states (HS-PS)*. USA: Achieve, Inc. on behalf of the twenty-six states and partners that collaborated on the NGSS.
Nguyen, N.-L. & Meltzer, D. (2003). Initial understanding of vector concepts among students in introductory physics courses. *American Journal of Physics, 71*(6), 630–638. doi:10.1119/1.1571831.
Organisation for Economic Co-operation and Development (2009). *PISA 2009 Assessment framework—key competences in reading, mathematics and science.* Retrieved from http://bit.ly/PISA2009Frame.
Palm, T., Boesen, J. & Lithner, J. (2011). Mathematical reasoning requirements in upper secondary level assessments. *Mathematical Thinking and Learning, 13*(3), 221–246. doi:10.1080/10986065.2011.564994.
Pólya, G. (1954). *Mathematics and plausible reasoning (vols. I and II)*. Princteton, NJ: Princeton University Press.
Planinic, M., Milin-Sipus, Z., Katic, H., Susac, A. & Ivanjek, L. (2012). Comparison of student understanding of line graph slope in physics and mathematics. *International Journal of Science and Mathematics Education, 10*(6), 1393–1414. doi:10.1007/s10763-012-9344-1.
Pålsgård, J., Kvist, G. & Nilson, K. (2005a). *Ergo Fysik A* [Ergo Physics A]. Stockholm, Sweden: Liber.
Pålsgård, J., Kvist, G. & Nilson, K. (2005b). *Ergo Fysik B* [Ergo Physics B]. Stockholm, Sweden: Liber.
Redish, E. F. (2003). *Teaching physics with the physics suite*. Hoboken, NJ: Wiley.
Redish, E. F. & Gupta, A. (2009). *Making meaning with math in physics: a semantic analysis.* Paper presented at GIREP 2009, Leicester, United Kingdom. Retrieved from http://bit.ly/RedishGuptaGIREP2009.
Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic.
Schoenfeld, A. H. (1992). Learning to think mathematically: problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334–370). New York, NY: Macmillan.
Silver, E. (1997). Fostering creativity through instruction rich in mathematical problem solving and problem posing. *ZDM-The International Journal on Mathematics Education, 29*(3), 75–80. doi:10.1007/s11858-997-0003-x.
Sumpter, L. (2013). Themes and interplay of beliefs in mathematical reasoning. *International Journal of Science and Mathematics Education, 11*(5), 1115–1135. doi:10.1007/s10763-012-9392-6.
Swedish National Agency for Education (2000). *Ämne - Fysik, gymnasieskolan* [Syllabuses for upper secondary Physics]. Retrieved 2014-05-22 from http://bit.ly/Fysik2000.
Swedish National Agency for Education (2003). *Lusten att lära – Med fokus på matematik: Nationella kvalitetsgranskningar 2001–2002* [Lust to learn—with focus on mathematics: National quality inspections 2001–2003]. Stockholm, Sweden: Fritzes.
Swedish National Agency for Education (2005). *Skolverkets Provbank. Hur används den och vad tycker användarna?* [The National Test bank. How it is used and what do the users think?]. Stockholm, Sweden: Skolverket.
Swedish National Agency for Education (2006). *Curriculum for the non-compulsory school system Lpf 94*. Stockholm, Sweden: Fritzes.
Swedish National Agency for Education (2009a). *TIMSS Advanced 2008. Svenska gymnasieelevers kunskaper i avancerad matematik och fysik i ett internationellt perspektiv* [TIMSS Advanced 2008. Swedish upper secondary students' knowledge in advanced mathematics and physics from an international perspective]. Stockholm, Sweden: Fritzes.
Swedish National Agency for Education (2009b). *Hur samstämmiga är svenska styrdokument och nationella prov med ramverk och uppgifter i TIMSS Advanced 2008?* [How aligned are the Swedish policy documents and national tests with the framework and the tasks in TIMSS Advanced 2008?]. Stockholm, Sweden: Fritzes.

Swedish National Agency for Education (2014-05-22). *Tabell [Table] 3B*. Retrieved from http://bit.ly/SweNAE2011.

Swedish Schools Inspectorate (2010). *Fysik utan dragningskraft. En kvalitetsgranskning om lusten att lära fysik i grundskolan* [Physics without attraction. An evaluation about the lust to learn physics in elementary school]. Report No. 2010:8. Retrieved from http://bit.ly/skolinsp2010.

Tasar, M. F. (2010). What part of the concept of acceleration is difficult to understand: The mathematics, the physics, or both? *ZDM-The International Journal on Mathematics Education, 42*, 469–482. doi:10.1007/s11858-010-0262-9.

Tuminaro, J. (2002). *How student use mathematics in physics: A brief survey of the literature.* College Park, MD: University of Maryland Physics Education Research Group. Retrieved from http://bit.ly/Tuminaro2002.

Winter, H. (1978). Geometrie vom Hebelgesetz aus – ein Beitrag zur Integration von Physik- und Mathematikunterricht der Sekundarstufe I [Geometry from the lever rule—a contribution to the integration of physics—and mathematics education at upper secondary school]. *Der Mathematikunterricht, 24*(5), 88–125.

Wyndham, J., Riesbeck, E. & Schoult, J. (2000). *Problemlösning som metafor och praktik* [Problem solving as metaphor and as practice]. Linköping, Sweden: Institutionen för tillämpad lärarkunskap, Linköpings universitet.

RELATION BETWEEN MATHEMATICAL REASONING ABILITY AND NATIONAL FORMAL DEMANDS IN PHYSICS COURSES

Helena Johansson

University of Gothenburg

It is widely accepted that mathematical competence is of great importance when learning physics. This paper focuses on one aspects of mathematical competence, namely mathematical reasoning, and how this competency influences students' success in physics. Mathematical reasoning required to solve tasks in physics tests, within a national testing system, is separated into imitative and creative mathematical reasoning. The results show that students lacking the ability to reason creatively are more likely not to do well on national physics test, thus not fully mastering the physics curricula. It is further discussed how the high demands of creative mathematical reasoning in physics tests stand in contrast to what is known about the educational practices in mathematics and physics in upper secondary school.

INTRODUCTION

Many scholars discuss the importance of understanding how mathematics is used in physics and how students' mathematical knowledge affects their learning of physics, e.g., Basson (2002) who mentions how difficulties in learning physics not only stem from the complexity of the subject but also from insufficient mathematical knowledge, Bing (2008), in his discussion of the importance of learning the language of mathematics when studying physics, as well as Redish and Gupta (2009), who emphasise the need to understand the cognitive thinking of experts in order to teach mathematics for physics more effectively to students.

According to the Swedish National Agency for Education (2009a) a common activity in physics classes is students using physics laws and formulas to solve routine tasks. The most common homework is to read in the textbook and/or to solve various tasks posed in the book, and sometimes to memorise formulas and procedures (ibid.). Similar results are described by Doorman and Gravemeijer (2009), who notice that most of the attention in both physics and mathematics in school is paid to the manipulations of formulas instead of focusing on why the formulas work. Redish (2003) states that practice, in the meaning that students just solve various tasks, is necessary but not enough to develop a deeper understanding of the underlying physics concepts. Students must learn both *how* to use the knowledge and *when* to use it.

The impact of mathematical reasoning *on mathematical* learning has been discussed and studied from multiple perspectives. Schoenfeld (1992), for example, points out that a focus on rote mechanical skills leads to poor performance in problem solving in contrast to the performance of mathematically powerful students. Lesh and

2015. In Beswick, K.., Muir, T., & Wells, J. (Eds.). *Proceedings of 39th Psychology of Mathematics Education conference*, Vol. 3, pp. 121-128. Hobart, Australia: PME.

Zawojeskij (2007) discuss how emphasising low-level skills does not give the students the abilities needed for mathematical modelling or problem solving, neither to draw upon interdisciplinary knowledge. Students lacking the ability to use creative mathematical reasoning thus get stuck when confronted with novel situations, and this negatively influences their possibilities to learn (Lithner, 2008). Since mathematics is a natural part of physics, it is reasonable to assume that the ability to use mathematical reasoning is an integral part of the physics knowledge students are assumed to achieve in physics courses.

FRAMEWORK

During studies on how students engage in various kinds of mathematical activities, Lithner (2008) developed a framework for characterising students' mathematical reasoning. The framework distinguishes between *creative mathematical founded reasoning* (CR) and *imitative reasoning* (IR). To be regarded as CR the following criteria should be fulfilled: **i. Novelty**. A new reasoning sequence is created or a forgotten one is recreated. **ii. Plausibility**. There are arguments supporting the strategy choice and/or strategy implementation motivating why the conclusions are true or plausible. **iii. Mathematical foundation**. The arguments made during the reasoning process are anchored in the intrinsic mathematical properties of the components involved in the reasoning (Lithner, 2008, p. 266).

Reasoning categorised as IR fulfils: **i.** The strategy choice is founded on recalling a complete answer. **ii.** The strategy implementation consists only of writing it down (Lithner, 2008, p. 258), or **i.** The strategy choice is to recall a solution algorithm. The predicted argumentation may be of different kind, but there is no need to create a new solution. **ii.** The remaining parts of the strategy implementation are trivial for the reasoned, only a careless mistake can lead to failure (ibid. p. 259).

In the application of the framework for the analyses described in this paper, an additional category, defined in Johansson (2103), is used. This category consists of those tasks that can be solved by only using physics knowledge; and this category is called *non-mathematical reasoning* (NMR). Physics knowledge is here referred to as relations and facts that are discussed in the physics courses and not in the courses for mathematics, according to the syllabuses and textbooks, e.g. that angle of incidence equals angle of reflection.

RESEARCH QUESTIONS

There is a significant amount of educational research on the relation between the school subjects of mathematics and physics that support the necessity of different mathematical competencies when learning physics. However, no studies on what type of mathematical reasoning is required of physics students were found. As an approach to the assumption that students' ability to reason mathematically affects how they master the physics curricula, this study use a previous analysis (Johansson, 2013) of the mathematical reasoning requirements to solve tasks in physics tests together with actual students' results on the same tests.

The Swedish national physics tests are the government's way of concretising the physics curricula. Thus, the requirements of mathematical reasoning to solve tasks in national physics tests should capture the mathematical reasoning that is required to master or fully master the curricula. The goals and the subject descriptions in the Swedish curricula of what it means to know physics are quite rich and are highly in accordance with the content and cognitive domains in the TIMSS Assessment framework (Garden et al. 2006; Swedish National Agency for Education, 2009b). This alignment with TIMSS suggests that the results from this study are relevant to an international context.

By addressing the questions: *Is it possible for a student to get one of the higher grades, Pass with distinction and Pass with special distinction, without using CR?*, and *If it is possible, how common is it?*, this study examines how the universal requirement of a mathematical reasoning competency to master the physics curricula relates to a specific assessment system's formal demands, in this case Sweden's.

METHOD

The empirical data consisted of student data from eight randomly chosen Swedish national physics tests for upper secondary school, and the tasks in the tests. There are mainly two different physics courses in the Swedish upper secondary school. Physics A that is compulsory for all natural science and technology students and Physics B that is an optional continuation. The tasks had previously been categorised according to mathematical reasoning requirements (Johansson, 2013), and together the tests comprised 169 tasks. The tests, which are classified to not authorised users, and the student data were used by permission from Department of Applied Educational Science at Umeå University, the department in charge of the National Test Bank in Physics. Student data come as excel sheets, one sheet for each test. The sheets contain information about individual students' grade, whereas the grade is one of the following: *Not Pass* (IG), *Pass* (G), *Pass with distinction* (VG), and *Pass with special distinction* (MVG). Further information in the sheets are individual student's scores on each task separated in G- and VG-scores, and their total score on the tests. No names of the students are present in the sheets, instead each student has got an ID-number. The IDs are unidentifiable for anyone outside the Department of Applied Educational Science at Umeå University, so data could be considered anonymous. The number of student data for each test varies from 996 to 3666.

For each test there are certain score levels the students need to attain to get a certain grade. To get the grade MVG, students need to fulfil certain quality aspects besides the particular score level. To decide if it is possible for a student to get one of the higher grades, VG or MVG, without using any kind of CR, each test was first analysed separately. This analysis consisted in comparing the score level for each grade with the maximum scores that are possible to obtain, given that the student only has solved (partly or fully) IR- and/or NMR- tasks. The available student data did not give any information about which of the qualitative aspects required for MVG the students have fulfilled, but the data sheets included students grades, thus MVG could be included in the analyses as one of the higher grades. After analysing if it is possible at all to receive the grades VG or MVG without solving any CR-tasks, students' actual results on the categorised tasks for those particular tests are summed up. The proportion of students who only got scores from IR- and/or NMR-tasks is then graphed with respect to the different grades.

RESULTS

Table 1 shows how the scores, possible to obtain on each of the eight tests that were analysed, are distributed among the reasoning categories IR and NMR. The table also includes the levels for the grades G, VG and MVG. The notation for the scores follows the convention G/VG.

Test	Max score (G/VG)	Min required score for G	Min required score for VG	Min required score for MVG	Max scores for IR-tasks	Max scores for NMR-tasks	Max score possible without CR-tasks
Physics A May 02	43 (26/17)	12	25 (with at least 6 VG scores)	25 (with at least 12 VG scores)	12/0	3/3	18 (with 3 VG)
Physics A Dec 04	40 (23/17)	12	24 (with at least 5 VG scores)	24 (with at least 12 VG scores)	14/3	3/3	23 (with 6 VG)
Physics A May 05	38 (22/16)	12	24 (with at least 6 VG scores)	24 (with at least 12 VG scores)	12/3	8/4	27 (with 7 VG)
Physics B May 02	48 (23/25)	12	27 (with at least 7 VG scores)	27 (with at least 13 VG scores)	11/4	2/0	17 (with 4 VG)
Physics B May 03	43 (23/20)	12	25 (with at least 6 VG scores)	25 (with at least 13 VG scores)	12/8	5/1	26 (with 9 VG)
Physics B May 05	44 (22/22)	12	25 (with at least 6 VG scores)	25 (with at least 12 VG scores)	8/5	7/2	22 (with 7 VG)
Physics B Feb 06	43 (22/21)	12	25 (with at least 7 VG scores)	25 (with at least 13 VG scores)	11/7	9/9	36 (with 16 VG)
Physics B April 10	44 (24/20)	12	25 (with at least 6 VG scores)	25 (with at least 12 VG scores)	9/4	4/1	18 (with 5 VG)

Table 1: Analysis of the distribution of G- and VG-scores among IR- and NMR-tasks.

For example, for the Physics A test from May 02 is the maximum score 43, and of these scores are 26 G-scores and 17 VG-scores. To pass this particular test a student has to have at least 12 scores, it does not matter if these scores are G- or VG-scores. To get the higher grade VG, a student has to have at least 25 scores and at least 6 of these scores have to be VG-scores. To get the highest grade, MVG, a student has to have at least 25 scores and at least 12 scores of these have to be VG-scores. As mentioned above, students also have to fulfil some additional quality aspects to achieve the grade MVG. Further, for the Physics A test from May 02, if a student only solves all tasks categorised as IR, he/she can obtain at most 12 G scores. If a student only solves all tasks categorised as NMR, he/she can obtain 3 G-scores and 3 VG-scores. Solving all IR- and NMR-tasks thus result in total 18 scores of which 3 are VG-scores. The scores for the rest of the analysed tests are presented in the same way.

In three of the eight tests (highlighted in Table 1) it is possible to get the grade VG by solving tasks not requiring any CR. In one of these tests, Physics B from February 2006, it is with respect to score level possible to obtain the grade MVG by solving only IR- and NMR-tasks. The analysis does not reveal anything about whether the requirements of the qualitative aspects for MVG are possible to fulfil by solving only these kinds of tasks.

Figure 1 illustrates the proportion of students on the three highlighted tests in Table 1 who only had solved IR- and/or NMR-tasks graphed with respect to their grades on the tests.

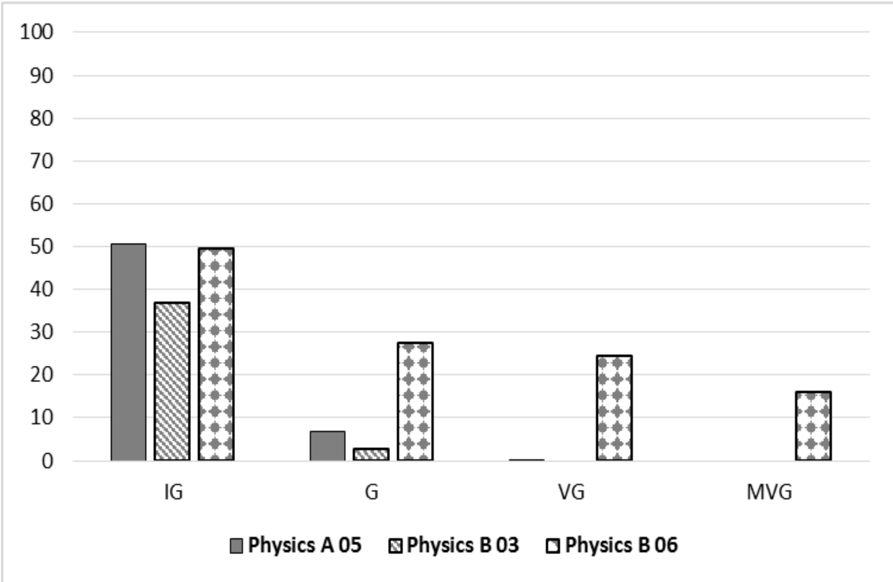


Figure 1: Proportion of students who only solved IR- and/or NMR-tasks with respect to the different grades.

It turned out that it is not frequently occurring that a student gets a higher grade than G by only solving these kinds of tasks. In the test for Physics A from 2005, only 0.17 % of the students got a higher grade; and in the Physics B test from 2003 none of the students got higher grades than G. The Physics B test from 2006 seems to be an exception though, since 25% of the students taking this test got a VG and 17% got a MVG. The analysis of how the scores are distributed among the reasoning categories for the different tests shows that the Physics B test from 2006 contains a lot more scores in the NMR category than any of the other tests (see Table 1). The total scores possible to obtain by only solving NMR-tasks are 18; nine of these are VG-scores, which is more than enough to fulfil the requirement for a VG (minimum 7 VG).

DISCUSSION

The analysis shows that it is possible to receive a higher grade than G by using only IR and NMR on three out of eight tests. When this result is compared with student data it is revealed that not using any CR, still receiving a higher grade, only occurs on one of the eight tests. This particular test, for which this occurs, is slightly different compared to the other tests with respect to how the scores are distributed among the reasoning categories (see Table 1). Further analysis of the test shows that tasks where it is possible to show the qualitative aspects required for the highest grade can be solved without using any mathematical reasoning i.e. these tasks are in the NMR category. This explains the higher frequency of students receiving the higher grades by using only IR and NMR, compared to the other tests.

The analysis of the tests furthermore shows that it is impossible to pass six of the eight tests without solving any tasks requiring mathematical reasoning. As seen in Table 1 it is only on the tests Physics A, May 05 and Physics B, Feb 06 a student can get at least the score 12, which is required to pass a test, by only solving NMR-tasks. These results strengthen the outcome from the author's previous study, which are that the ability to reason mathematically is an important competency and an integral part when taking physics tests (Johansson, 2013).

Mathematical reasoning is defined as a process to reach conclusions when solving tasks (Lithner, 2008). When students have the ability to use creative mathematical founded reasoning, they know how to argue and justify their conclusions and they can draw on previous knowledge. The result in the present study shows that CR is required to succeed on most of the physics tests. The alignment between the TIMSS framework and the Swedish policy documents suggests that this is a universal demand on upper secondary physics students. Viewing the physics tests from the National Test bank as an extension of the national curricula, one can assume that students' results on the tests are a measure of their knowledge in physics. It is well known that a focus on IR can explain some of the learning difficulties that students have in mathematics. The results above show that a focus on IR when studying physics in upper secondary school will make it hard for the students to do well on the physics tests, thus fully mastering the physics curricula. Therefore, a reasonable conclusion is that focusing on IR can hinder students' development of knowledge in physics, similar to results found about mathematics, and a creative mathematical reasoning competency can be regarded decisive.

The argumentative side of mathematics, which is a reasoning based on intrinsic properties of the components involved in the task-solving process, seems to be an inseparable part of mastering physics. All students should have the same possibilities to achieve the goals in the physics curricula. Therefore, they ought to be given the opportunity in school to develop and practice this creative mathematical reasoning competency that is required. As mentioned in the introduction, it is common in the physics classes that students solve routine tasks and focus on manipulations on formulas instead of focusing on the conceptual understanding of the underlying principles (Doorman & Gravemeijer, 2009; Swedish National Agency for Education, 2009a). Although it is the physics perspective that is discussed in the above studies, it is reasonable to assume that if there is more focus on physics procedures than on the understanding of physics concepts, there is also little focus on creative mathematical reasoning.

It is not only the physics classes that might provide students the opportunity to develop a mathematical reasoning competency, this competency is of course relevant also in the mathematics classes. According to studies about the learning environment in mathematics classes, the focus is on algorithmic procedures and the environment does not provide extensive opportunities to learn and practice different kinds of reasoning (e.g., Boesen, Lithner & Palm, 2010). During observations of classroom activities it was shown that opportunities to develop procedural competency was present in episodes corresponding to 79% of the observed time; compared to episodes involving opportunities to develop mathematical reasoning competency, which were present in 32% of the observed time (Boesen et al., 2014). Also tests have an indirect role for students learning, both as formative, when students get feedback on their solutions, and as summative, when the character of the tasks give students indications of what competences that are sufficient for handling mathematical tasks. Analyses of teacher-made mathematics tests have shown that these focused largely on imitative reasoning, in contrast to the national mathematics tests, which had a large proportion of tasks requiring creative mathematical reasoning (Palm, Boesen, & Lithner, 2011). Altogether, the above discussion shows that students are provided limited opportunities to develop the creative mathematical reasoning competency that is formally required to master the physics curricula. The importance of the relation between mathematics and physics has been known for a long time. The result from the present study, that the ability to creatively mathematically argue and reason is decisive in order to fully master the physics curricula, should have implications on how the education is organised and carried out.

References

Basson, I. (2002). Physics and mathematics as interrelated fields of thought development using acceleration as an example. *International Journal of Mathematical Education in Science and Technology*, *33*(5), 679-690.

Bing, Thomas. (2008). *An epistemic framing analysis of upper level physics students' use of mathematics* (Doctoral dissertation). University of Maryland. Retrieved 2010-02-10 from http://bit.ly/Bing2008

Boesen, J., Lithner, J. & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning student use. *Educational studies in mathematics*, *75*(1), 89-105.

Boesen, J. et al. (2014). Developing mathematical competence: From the intended to the enacted curriculum. *Journal of Mathematical Behavior*, *33*, 72– 87.

Doorman, L. M. & Gravemeijer K. P. E. (2009). Emergent modelling: discrete graphs to support the understanding of change and velocity. *ZDM – The International Journal on Mathematics Education*, *41*, 199-211.

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O., Mullis, I. V. S., et al. (2006). *TIMSS Advanced 2008 Assessment Frameworks*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Johansson, H. (2013). Mathematical reasoning in physics tests – Requirements, relations, dependence (Licentiate thesis). Mathematical Science, University of Gothenburg, Gothenburg.

Lesh, R., & Zawojewski, J. (2007). Problem solving and modelling. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 763-804). Charlotte, NC: Information Age Publishing.

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics*, *67*(3), 255-276.

Palm, T., Boesen, J. & Lithner, J. (2011). Mathematical reasoning requirements in Swedish upper secondary level assessments. *Mathematical Thinking and Learning*, *13*(3), 221-246.

Redish, E.F. (2003). *Teaching physics with the physics suite*. USA: John Wiley & Sons, Inc.

Redish, E. F. & Gupta, A. (2009). Making meaning with math in physics: A semantic analysis. *Contributed paper presented at GIREP 2009*, Leicester, UK, August 20, 2009. Retrieved 2012-06-25 from http://bit.ly/RedishGuptaGIREP2009

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334-370). New York: MacMillan.

Swedish National Agency for Education (2009a). *TIMSS Advanced 2008. Svenska gymnasieelevers kunskaper i avancerad matematik och fysik i ett internationellt perspektiv*. Stockholm: Fritzes

Swedish National Agency for Education (2009b). *Hur samstämmiga är svenska styrdokument och nationella prov med ramverk och uppgifter i TIMSS Advanced 2008?*. Stockholm: Fritzes.

Dependence between creative and non-creative mathematical reasoning in national physics tests

HELENA JOHANSSON

It is known from previous studies that a focus on rote learning and procedural mathematical reasoning hamper students' learning of mathematics. Since mathematics is an integral part of physics, it is assumed that mathematical reasoning also influences students' success in physics. This paper aims to study how students' ability to reason mathematically affects their success on different kinds of physics tasks. A descriptive statistical approach is adopted, which compares the ratio between conditional and unconditional probability to solve physics tasks requiring different kinds of mathematical reasoning. Tasks from eight Swedish national physics tests for upper secondary school, serve as a basis for the analysis. The result shows that if students succeed on tasks requiring creative mathematical reasoning, the probability to solve the other tasks on the same test increases. This increase is higher than if the students succeed on tasks not requiring creative mathematical reasoning. This result suggests that if students can reason mathematically creatively, they have the ability to use their knowledge in other novel situations and thus become more successful on tests.

Many scholars discuss the importance to understand how mathematics is used in physics and how students' mathematical knowledge affects their learning of physics (e.g. Basson, 2002; Bing, 2008; Nguyen & Meltzer, 2003; Redish & Gupta, 2009). Basson (2002), for example, mentions how difficulties in learning physics not only stem from the complexity of the subject but also from insufficient mathematical knowledge. diSessa (1993) notices how students, who have studied physics, can solve a quantitative task in physics and still give an inconsistent qualitative analysis of the same task. A *quantitative* task refers to when the task is posed in explicitly quantitative terms and the solution can be attained through calculations using appropriate physics laws. A *qualitative* task on the other

Helena Johansson
Mid Sweden University

Johansson, H. (2017). Dependence between creative and non-creative mathematical reasoning in national physics tests. *Nordic Studies in Mathematics Education*, 22 (2), 93–119.

hand refers to when the solution requires an analysis of the posed physical situation i.e. what will occur and/or why. It is shown that students' intuitive understanding of the physical world is quite robust and that their solutions to qualitative problems often contradict the basic physics principles (ibid.). Redish (2003) states that practice, in the meaning that students just solve various tasks, is necessary but not enough to get a deeper understanding of the underlying physics concepts. Students must learn both *how* to use the knowledge and *when* to use it. The same conclusion holds for learning mathematics, shown by e.g. Schoenfeld (1985) in his study of how students become good problem solvers in mathematics; as well as by Lesh and Zawojewski (2007), who discuss how working with mathematical modelling develops students' understanding and learning in mathematics. Michelsen (2005) also addresses benefits from modelling activities. He discusses how interdisciplinary modelling activities can help students to understand how to use mathematics in physics and discover the connections between the two subjects.

During studies of how students are engaging in different mathematical activities, Lithner (2008) has gradually developed a framework for characterising students' mathematical reasoning. The framework distinguishes between *creative mathematical founded reasoning* (CR) and *imitative reasoning* (IR). The former one refers to a reasoning that is anchored in intrinsic mathematical properties and that includes some novelty to the reasoner. If instead the anchoring is in surface properties and the reasoning consists of remembering an answer or following a process step by step, it is IR. Mathematical reasoning is one aspect of mathematical knowledge, and thus assumed to be one competence that influences students' learning of physics. Johansson (2015) shows e.g. that to pass Swedish national physics tests, students have to reason mathematically; and to fully master the Swedish physics curricula students have to be able to use CR. To examine further how students' ability to reason mathematically influences how they succeed in physics, the aim in this study is to analyse if there are any dependencies between students' success on different physics tasks, different with respect to which type of mathematical reasoning that is required to solve the tasks.

Conceptual framework

Lithner define *reasoning* as "the line of thought adopted to produce assertions and reach conclusions in task solving" (Lithner, 2008, p. 257). *Mathematical reasoning* is used as an extension of a strict mathematical proof to justify a solution and is seen as a product of separate reasoning sequences. Each sequence includes a choice that defines the next sequence, and the reasoning is the justification for the choice that is made. The mathematical foundation of the reasoning can either be *superficial* or *intrinsic*. The accepted mathematical properties of an object are of different relevance in different situations. This leads to a distinction between *surface* properties and *intrinsic* properties, where the former have little relevance in the actual context and lead to superficial reasoning and the latter are central and have to be taken into consideration in the given context (Lithner, 2008). As mentioned in the introduction, this framework was developed during empirical studies of how students engage in various mathematical activities. A strength of the framework is that it is not restricted to any specific context. As long as the students have to use some kind of mathematics to come up with a solution, they are assumed to reason mathematically. Therefore, Lithner's framework is considered suitable for categorising the kinds of mathematical reasoning that are required in the physics tests.

Creative mathematically founded reasoning

Creativity is an expression often used in different contexts and without an unequivocal definition (for a discussion see Lithner, 2008, p. 267–268). For the definitions of the different kinds of reasoning, the perspective of Haylock (1997) and Silver (1997) is adopted. This implies that creativity is seen as a thinking process that is novel, flexible and fluent (Lithner, 2008). CR fulfils all of the following criteria:

> i. Novelty. A new reasoning sequence is created or a forgotten one is recreated.
>
> ii. Plausibility. There are arguments supporting the strategy choice and/or strategy implementation motivating why the conclusions are true or plausible.
>
> iii. Mathematical foundation. The arguments made during the reasoning process are anchored in the intrinsic mathematical properties of the components involved in the reasoning.
>
> (Lithner, 2008, p. 266).

Imitative reasoning

The arguments that motivate the chosen solution method (i.e. the reasoning) could be anchored in surface mathematical properties. Reasoning categorised as IR fulfils

> i. The strategy choice is founded on recalling a complete answer.
>
> ii. The strategy implementation consists only of writing it down.
>
> (Lithner, 2008, p. 258)

or

> i. The strategy choice is to recall a solution algorithm. The predicted argumentation may be of different kind, but there is no need to create a new solution.
>
> ii. The remaining parts of the strategy implementation are trivial for the reasoned, only a careless mistake can lead to failure.
>
> (Lithner, 2008, p. 259)

Local and global creative mathematical reasoning

Lithner (2008) introduces a refinement of the category CR into *local CR* (LCR) and *global CR* (GCR) that captures some significant differences between tasks categorised as CR. This subdivision has been further elaborated by other scholars, e.g. Boesen, Lithner and Palm (2010) and Palm, Boesen and Lithner (2011). In LCR, the reasoning is mainly IR but contains a minor step that requires CR. If instead there is a need for CR in several steps, it is called GCR.

Non-mathematical reasoning

In the application of the framework, an additional category, defined in Johansson (2015), is used. This category consists of those tasks that can be solved by only using physics knowledge; and this category is called *non-mathematical reasoning* (NMR). Physics knowledge is here referred to as relations and facts that are discussed in the physics courses and not in the courses for mathematics, according to the syllabuses and textbooks, e.g. that angle of incidence equals angle of reflection.

Related concept

There has been a discussion in the mathematical educational research society whether procedural knowledge should be considered only as superficial and rote learned or viewed from a wider perspective (Baroody, Feil & Johnson, 2007; Star, 2007). Hiebert and Lefevre (1986) defined procedural knowledge as consisting of the formal language of mathematics, as well as of the algorithms and rules for completing mathematical tasks. There is an agreement that procedural knowledge is important, but not enough, when students learn mathematics (e.g. Baroody et al., 2007; Gray & Tall, 1994; Hiebert & Lefevre, 1986; Star, 2007). However, there is also an argumentation about whether *deep* procedural knowledge could exist without involvement of conceptual knowledge (Baroody, Feil & Johnson 2007; Star, 2005, 2007). In the description of the framework used for characterising required mathematical reasoning, Lithner (2008) discusses different aspects of procedures and concepts. Although the definitions of the reasoning categories do not include references to procedural or conceptual knowledge, one could assume some relations between CR and conceptual knowledge on one hand and IR and procedural knowledge on the other hand.

Research questions

Based on Lithner's (2008) framework, physics tasks in Swedish national tests have been categorised with respect to mathematical reasoning requirements in Johansson (2015). The main results showed that students must use some kind of mathematical reasoning to solve three-fourth of the tasks in a test; and that one-third of the tasks require CR. From the outcome, one of the interests that arose was if there is a dependence between how students succeed on physics tasks requiring different kinds of mathematical reasoning. To study a possible dependence the following research questions are posed:

- Does the success on a physics task that requires CR affect the probability to succeed on any other task requiring either IR or CR in the same test?
- Does the success on a physics task solvable with IR affect the probability to succeed on any other task requiring either IR or CR in the same test?

The answers to both questions are intuitively yes, but has to be verified in order to answer the following two research questions:

- How strong is the dependence in each case?
- Are there any difference in effects on tasks requiring different mathematical reasoning?

Method

Physics in the Swedish School

There are mainly two different physics courses in the Swedish upper secondary school. Physics A, which is compulsory for all natural science and technology students, and Physics B that is compulsory for natural science students and an optional continuation for technology students. In the current curricula (from 2011) the names of the courses have changed to Physics 1 and Physics 2, and some of the areas previously included in Physics B are now in Physics 1. During the last decades, there has been a gradual change towards a stronger focus on process goals, and they are present in the curriculum from 1994 (Swedish National Agency for Education, 2006). Content goals are complemented with process goals, and teaching in the subject of physics should for example aim at helping students develop knowledge of the concepts, theories, models and working methods of physics. Students should be given opportunities to develop a scientific approach to the surrounding world, as well as to analyse and solve problems through reasoning based on concepts and models. Mathematics is explicitly required when making quantitative descriptions of phenomena and implicitly required when analysing data (Swedish National Agency for Education, 2000). Similarities between the upper secondary syllabuses for physics in Norway and Sweden are identified and discussed in the TIMSS Advanced 2008 report (Lie, Angell & Rohatgi, 2010). It is further discussed by Grønmo and Onstad (2013), how students' mathematical performance in the Nordic countries form a specific *Nordic profile*, distinct from other countries.

National physics tests for both physics courses are provided by the Swedish National Agency for Education through the National Test Bank as an assessment support to accomplish equivalent assessment for upper secondary physics students throughout the country. Most of the tests are classified to not authorised users. There are a few tests open to the public, which for example students can look upon to get an idea of what is expected. After a test is used, students' results are collected and compiled via the National Test Bank.

Data

The data comprise tasks from eight physics tests from the Swedish National Test Bank. The tests are the May 2002, December 2004 and May 2005 tests for the Physics A course and the May 2002, May 2003, May 2005, February 2006, and April 2010 tests for the Physics B course. These tests were chosen because the tasks in the tests already had been categorised according to mathematical reasoning requirements, and that there were available data about students' results on each of the tasks. Student data were used by permission from Department of Applied Educational Science at Umeå University, the department in charge of the National Test Bank in Physics. No names of the students are present, instead each student has got an ID-number, and thus data could be considered anonymous. The number of students for each test varies from 996 to 3666.

There are in total 119 of the 162 physics tasks on the tests that require mathematical reasoning to be solved, and thus included in this study, i.e. no NMR tasks are included.

Statistical method

To decide whether there exists a dependence between success on a particular task R, the *reference task*, and the success on another task X, it was decided to compare the *conditional probability* $P(X=1|R=1)$ to solve X with the *unconditional probability* $P(X=1)$ to solve X. That is, the ratio

$$\frac{P(X=1|R=1)}{P(X=1)} \qquad (1)$$

was estimated, where X = 1 and R = 1 denote that the tasks have been fully solved, respectively. If this ratio is larger than 1, the probability to succeed on the task X is higher if students successfully have solved the task R than if they have not. The probabilities in (1) are estimated by computing the arithmetic means from the available student data for each test. To estimate $P(X=1|R=1)$, the number of students who had solved both X and R were divided by the number of students who had solved R. The probability $P(X=1)$ was estimated by calculating the number of students who had solved X by the total number of students who had taken the test.

In order to decide if the effect of a calculated dependence is large enough to consider, *odds ratio* is used as a measure of the effect-size. Odds ratio is defined as

$$\frac{\dfrac{P(X=1|R=1)}{1-P(X=1|R=1)}}{\dfrac{P(X=1|R=0)}{1-P(X=1|R=0)}} \qquad (2)$$

where $P(X=1|R=1)$ is as defined above and $P(X=1|R=0)$ denotes the conditional probability to solve X when R *is not* solved, i.e. students have only partly or not at all solved the task R. The effects are divided into small, medium and large, and associated with the calculated magnitudes of odds ratios as follows; small = 1.5, medium = 3.5 and large = 9.0. These values could be considered as a rule of thumb (e.g. Cohen, 1988; and Hopkins, 2002).

The paired sample *t*-test was used for significance testing of the difference between the means of the dependencies that are calculated when

CR-tasks are used as reference tasks and when IR-tasks are used as reference tasks. In order to decide if a significant difference is to be accounted for, Cohen's *d* is used as an index of the effect size. Effect sizes around 0.2 could be considered small, effect sizes around 0.5 are medium and sizes above 0.8 indicate large effects (Cohen, 1988).

Implementation

For each test, a CR task was first chosen as reference task. The CR tasks were chosen so that they should not be the most difficult ones, i.e. they should not require too many analysing steps to be solved. The decisions were based on the formulation of the assignments and on what solutions that were required. For most of the tests the chosen CR task is a GCR task. On two tests the tasks categorised as GCR were judged to only occur among the most difficult tasks, thus LCR tasks were chosen instead. There was now one CR task on each test and this task was used as R in (1). The ratio was then estimated for every other task that required mathematical reasoning in each test, respectively. Since it is the effect on success with respect to mathematical reasoning that is studied, tasks not requiring mathematical reasoning, i.e. NMR tasks, were excluded from the analysis. To analyse the same ratio (1), but with an IR task as the reference task, it was decided to choose an IR task with a position approximately in the middle of the tests and relatively close to the already chosen CR task. This choice are based on that a task's positon in a test indicates how difficult the task is supposed to be to the students. The ratio (1) was estimated in the same way as above for every task requiring mathematical reasoning in each test, respectively, with the chosen IR tasks as R in (1). Below follows two examples of different tasks used as reference tasks in the Physics A 2002 test. Task 6 (figure 1) has previously been categorised as solvable with IR, and task 12 (figure 2) as requiring LCR. The method for the categorisation is thoroughly described in Johansson (2015). Two examples of the categorisation process are provided in appendix, example A-1 and example A-2.

To account for possible effects due to the positions of the tasks, it was decided to do some more calculations of the ratio (1) by choosing a GCR task that occurs earlier in the test than the IR task previously used as reference task. If there were no such GCR task, an IR task that came later in the test than the previously used CR task should be used as reference task. If there were no such IR task, an IR task positioned as close as possible to the previously used CR task was used as reference task. There are now three different ratios, i.e. with different reference tasks, for every test. To be able to statistically compare the ratios, the measure of effect size

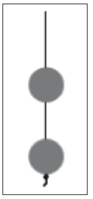
Read the press cutting below.

Watch out!
A study of cause of death and injuries among the population in the South Pacific shows that most accidents are caused by falling coconuts and overturned palm trees. This is nothing to laugh about. A four kilos coconut that comes loose from a 25 meter high palm tree, reaches a speed of 80 km/h and hits the ground – or an unfortunately placed head – with a pressure that corresponds to one ton. The study is performed by Doctor Herman Oberli at the hospital in Honiara, Solomon Islands. (TT-DPA)

Is it true that a coconut can reach 80 km/h after a 25 m high fall?

Figure 1. *Task 6*

In order to determine the charge on two small, light silver balls, the following experiment was conducted. The balls, which were alike, weighed 26 mg each. The balls were threaded onto a nylon thread and were charged in a way that gave them equal charges. The upper ball levitated freely a little distance above the other ball. There was no friction between the balls and the nylon thread. The distance between the centres of the balls was measured to 2.9 cm.



What was the charge on each of the balls?

Figure 2. *Task 12*

(2) is calculated for respective ratio. The two calculated ratios (1) with an IR task and a CR task as close as possible were furthermore used in the paired sample *t*-test in order to decide if the means differ significantly.

Analysis and result

In table 1 the calculations for respective task on the Physics A 2002 test are displayed. The estimated values for the ratios in the first row are the results when the CR task, task number 12 (figure 2), is used as reference. Similarly, the estimated values for the ratios in the second row are the results with the IR task, task number 6 (figure 1), used as reference. By comparing the two estimated ratios for each task it was noticed that the ratio in most cases was larger when the CR-task was the reference, compared to when the IR-task was used in the calculations. For example, the estimated values on row 1 and 2 for task number 1a in table 1 are 1.21 and 1.14, respectively, which shows that it is more likely to succeed on task 1a if you solve the CR task than if you solve the IR task.

Table 1. *Ratios according to (1) for the tasks in the Physics A 2002 test with the tasks 12 and 6 as reference tasks*

Task	1a (IR)	1b (IR)	2a (NMR)	2b (IR)	3 (IR)	4 (LCR)	5 (LCR)	**6 (IR)**	7 (NMR)	8 (NMR)
12	1.21	1.47		1.20	1.44	1.08	1.36	1.49		
6	1.14	1.21		1.12	1.19	1.07	1.15	R		

Task	9 (NMR)	10 (LCR)	11 (IR)	**12 (LCR)**	13 (GCR)	14 G (LCR)	14 VG (GCR)	15 (GCR)	16a (GCR)	16b (GCR)
12		1.29	1.87	R	1.27	1.95	4.76	2.28	1.68	2.81
6		1,19	1.24	1.49	1.17	1.31	1.69	1.37	1.22	1.43

Corresponding values for e.g. task 13 in table 1 are 1.27 and 1.17, respectively, which indicates the same result. The corresponding tables containing the estimated values for the ratios for the rest of the physics tests are provided in appendix, table A-1 to table A-7. Furthermore, by comparing the values in each entry for respective row in table 1 and in the rest of the tables in appendix, the results indicate that the dependence between success on a specific task and on the rest of the tasks on the test increases the later the tasks are positioned in the tests. This tendency seems to be the same for both IR tasks and CR tasks used as reference tasks. For example, consider the values in the first row in table 1, all values are less than 1.5 for the first 10 tasks, and for task 11 and the following six tasks, only one value is less than 1.5. Similar result holds for the values in the second row, where the values for the first 10 tasks except one, are less than 1.20, and all values are larger than 1.20 for all but one of the rest of the tasks.

To analyse further whether there are any differences in the ratios with respect to mathematical reasoning categories, the mean ratio ($\bar{r}$) were calculated for respective category (IR, LCR and GCR) for each reference task in each test (table 2). The first row for every test in table 2 shows the values when the chosen CR task is used as reference task, and the second row shows the values when the IR task is used as reference task. So the top two rows are the respective means of the values in table 1. The values indicate that the success on other tasks is more dependent on students' success on a CR task than on their success on an IR task.

Table 2. *The mean ratio, for each of the mathematical reasoning categories with respect to the different reference tasks in each test*

	R	$\bar{r}$		
		IR	LCR	GCR
Physics A VT 02	12 (LCR)	1.45	1.43	2.56
	6 (IR)	1.18	1.24	1.38
Physics A HT 04	11 (GCR)	1.42	1.68	2.03
	8a (IR)	1.26	1.39	1.31
Physics A VT 05	10 (GCR)	1.28	1.53	1.45
	8a (IR)	1.02	1.03	1.03
Physics B VT 02	10 (GCR)	1.44	1.84	2.21
	7 (IR)	1.27	1.51	1.59
Physics B VT 03	8 (LCR)	1.27	1.44	1.65
	7 (IR)	1.13	1.20	1.27
Physics B VT 05	12b (GCR)	1.17	1.21	1.22
	8b (IR)	1.25	1.46	1.42
Physics B VT 06	12a (GCR)	1.40	1.73	1.94
	10b (IR)	1.38	1.32	1.29
Physics B VT 10	11b (GCR)	1.39	1.52	2.03
	9b (IR)	1.19	1.24	1.49

The tendency discussed above, that the dependence increase with the tasks position in the test, suggests that position has an effect on the dependence. As described in the Method section, a new reference task with another position was chosen for each test and values for the ratios according to (1) were estimated with this new task as reference. The new values for the Physics A 2002 test are displayed in the last row in table 3. The first two rows in the table are the previously calculated and displayed ratios in table 1, with the two other tasks used as reference tasks. As previously, tables for the other physics tests are available in appendix, table A-8 to table A-13. No new task were chosen for the Physics B 2003 test since the two previously used tasks already were positioned next to each other in the test (table A-4).

Comparing the values in row 1 and 3 for each task in table 3 shows that the ratio in all except one case, was larger when the CR task (task 12) was the reference, compared to when the IR task (task 11) was used in the calculations. Consider for example the values in row 1 and 3 for task 1b, these are 1.47 and 1.35, respectively; and corresponding values for task 15 are 2.28 and 2.08, respectively. Similar results were obtained for the rest of the calculations, which indicate that success on a GCR task, even when

Table 3. *Ratios according to (1) for the tasks in the Physics A 2002 test with the new task 11 together with tasks 12 and 6 as reference tasks*

Task	1a (IR)	1b (IR)	2a (NMR)	2b (IR)	3 (IR)	4 (LCR)	5 (LCR)	**6 (IR)**	7 (NMR)	8 (NMR)
12	1.21	1.47		1.20	1.44	1.08	1.36	1.49		
6	1.14	1.21		1.12	1.19	1.07	1.15	R		
11	1.18	1.35		1.16	1.35	1.09	1.33	1.40		

Task	9 (NMR)	10 (LCR)	**11 (IR)**	**12 (LCR)**	13 (GCR)	14 G (LCR)	14 VG (GCR)	15 (GCR)	16a (GCR)	16b (GCR)
12		1.29	1.87	R	1.27	1.95	4.76	2.28	1.68	2.81
6		1.19	1.24	1.49	1.17	1.31	1.69	1.37	1.22	1.43
11		1.23	R	2.11	1.20	1.77	2.13	2.08	1.45	1.23

account for position in the test is taken, in most cases has a larger effect on the success on the rest of the tasks, than success on an IR task has.

The means of the ratios for respective reasoning categories with respect to the latest reference tasks in each physics tasks are displayed in table 4. Comparing the means with the previous calculated ones (table 2), shows that in six out of eight tests the effect on success is higher for all three reasoning categories when students succeed on a CR task compared to if they have succeeded on an IR task. The means of the estimated ratios for the Physics A 2002 test are for example 1.45 for IR tasks, 1.43 for LCR tasks and 2.56 for GCR tasks when the LCR task (task 12) was used as the reference task (table 2). Corresponding means are 1.29, 1.51 and 1.62, respectively (table 4), when the IR task (task 11) was used as the reference task in equation (1).

The significance testing was performed to test the hypothesis that if students solve a CR task they have a higher chance to solve other physics tasks requiring mathematical reasoning, than if they solve an IR task. The hypothesis was tested on the pair of calculated values for the conditional probabilities (1) with the CR task and the IR task as close as possible as reference tasks, or with a CR task occurring earlier in the test than the used IR task. For the Physics A 2002 test this means that it is the values in the first and the last rows for every task (not used as R) that are included (table 3). Since the difference is assumed to be general and not restricted to specific tasks or tests, the *t*-test was performed on all physics tests together. In total 103 pairs of values were included. The result showed that there is a significant difference of the means of the conditional probabilities with CR as R and with IR as R ($\bar{x}_D = 0.15$, $p = 0.000017 < 0.05$). The effect of the significant difference, $d = 0.41$, is considered to be around 0.5 and thus in the lower range of medium.

Table 4. *The mean ratio for each of the mathematical reasoning categories with respect to the added reference tasks in each test*

	R	$\bar{r}$		
		IR	LCR	GCR
Physics A VT 02	11 (IR)	1.29	1.51	1.62
Physics A HT 04	7a (GCR)	1.19	1.31	1.43
Physics A VT 05	12a (IR)	1.19	1.37	1.33
Physics B VT 02	9b (IR)	1.39	1.67	1.97
Physics B VT 05	7 (GCR)	1.27	1.45	1.42
Physics B VT 06	4 (GCR)	1.50	1.45	1.65
Physics B VT 10	7b (GCR)	1.30	1.31	1.67

As described in Method, odds ratio is used as a measure of effect size of the separately calculated dependencies. Estimations of the odds ratio were calculated according to (2) for all tasks in every physics test with the three different reference tasks, respectively. The values of effect sizes for the tasks in the Physics A 2002 test are displayed in table 5, and the effect sizes for the rest of the physics tests are provided in appendix, table A-14 to table A-20. The effect sizes that are interesting to compare for each task are at first-hand the two that are calculated with reference tasks as close as possible to each other, or when the reference task is a CR task occurring earlier in the test than the used IR task. Then, as described above, possible influences of the tasks position in the tests are considered. Thus, for the Physics A 2002 test, the values in the first and the last rows are the one most interesting to compare. For example, the value of the effect size for having solved the CR task 12 (row 1, table 5) is 7.13 for task 1a, and 3.95 for task 1b. Both these could be considered medium effects according to the rule of thumb outlined in the Statistical method section, i.e. the value is between 3.5 and 9. It is also noticed that the effect is larger on task 1a than on 1b. Corresponding values of the effect for having solved the IR task 11 (row 3, table 5) are 6.91 for task 1a and 4.04 for task 1b, which also can be considered as a medium effect, and larger for task 1a than for task 1b.

When analysing the values in row 1 and row 3 for every task, it is noticed that there is a medium effect on 8 of the 14 tasks that are included in the analysis and not used as reference tasks, and a large effect on 2 of the 14 tasks. On the rest of the four tasks the effect is small. When only effect sizes that can be considered at least medium is taken into account, and pair-wise compared for each task in the Physics A 2002 test, it is noticed that the effect is larger on five tasks if the students have solved the

Table 5. *Odds ratios according to (2) for the tasks in the Physics A 2002 test with tasks 12, 6 and 11 as reference tasks*

Task	1a (IR)	1b (IR)	2a (NMR)	2b (IR)	3 (IR)	4 (LCR)	5 (LCR)	**6 (IR)**	7 (NMR)	8 (NMR)
12	7.13	3.95		6.76	4.11	1.56	3.16	4.70		
6	5.06	2.84		4.05	2.72	1.92	2.44	R		
11	6.91	4.04		5.18	3.98	1.96	3.47	5.54		
	9 (NMR)	10 (LCR)	**11 (IR)**	**12 (LCR)**	13 (GCR)	14 G (LCR)	14 VG (GCR)	15 (GCR)	16a (GCR)	16b (GCR)
12		2.67	6.74	R	2.04	3.09	18.9	3.82	4.75	5.44
6		3.42	4.54	4.70	2.31	3.06		3.21	2.50	3.50
11		2.98	R	8.43	2.14	4.93	10.6	6.77	4.11	11.9

CR task, compared to if they have solved the IR task. This holds for task 1a, 2b, 3, 14VG and 16a (table 5). It is also noticed that the effect is higher on five of the tasks if the student instead solves the IR task compared to the CR task, see task 1b, 6, 14G, 15 and 16b in table 5. The values in the table also shows that the effect in most cases are larger for the five tasks with a larger effect for solving the CR task, than on the five task with a larger effect for solving the IR task.

Compiling the above result with the result from the analysis of the effect sizes for the tasks in the rest of the physics tests shows that there are 50 out of 103 tasks that have a medium effect. Of these 50 tasks the effect is larger on 26 of the tasks if the students have solved the CR task used as reference task compared to if they have solved the IR task (table 6). There are four tasks (of the 103 tasks) that have a large effect; and for three of these four tasks, the effect is larger if students solve the CR task than if they solve the IR task. Furthermore, there are one task that no students manage to solve without also solving the CR task used as

Table 6. *Number of tasks with the different effect sizes*

Effect size	Number of tasks
Small effect	48
Medium effect, larger with CR as reference task	26
Medium effect, larger with IR as reference task	24
Large effect, larger with CR as reference task	3
Large effect, larger with IR as reference task	1
Infinite effect with CR as reference	1

reference task, thus the nominator in (2) is 0 and the odds-ratio turns to infinity (task 14VG, table A-14).

When only effects considered at least medium (and the one with infinity is excluded), there are in total 29 tasks for which the effects are higher when a CR task is used as reference task, and 26 tasks for which the effects are higher when an IR task is used as reference. Comparing this with the previous result, that there is a significant difference between the general effect of solving a CR task and an IR task, shows that although the general effect is large enough to consider, the effects on individual tasks do not differ so much.

Discussion and implications

The outcome of the present study shows that mastering the ability to reasoning mathematically creatively has a positive effect on the success on other physics tasks. It is shown that the effect generally is higher for tasks requiring CR compared to tasks solvable with IR. Going back to the definitions of the reasoning categories, CR tasks require that students can use their mathematical knowledge in novel situations, which in turn implies an intrinsic understanding of the mathematics that is involved. At the same time, when students are able to use their knowledge in novel situations, they have also developed another approach to the task solving process. Their strategy is based on the judgement of plausibility, which means that they analyse the task/assignment and have an idea of plausible conclusions. This ability to reason mathematically creatively is thought to be generalizable to various mathematical areas. IR tasks, on the other hand, could be solved by remembering an algorithm, and no intrinsic understanding of the mathematics is required. Therefore it is reasonable that the effect between success on CR tasks is higher than the effect of success on a CR task and on an IR task. Nevertheless, there is still a positive effect on IR tasks from the success on CR tasks, which suggests that students who have developed the ability to reason mathematically creatively also have a better chance to succeed on tasks of a more procedural character. That no intrinsic understanding is required in order to solve IR tasks does not exclude the possibility that students could have developed some conceptual understanding; and thus, success on IR tasks positively affects the success on CR tasks. At the same time, the only characteristics different IR tasks have in common, at least theoretically, are that they should be possible to solve by remembering an answer or a procedure and implement this. Therefore it is reasonable to expect that the effect on the success on other IR tasks varies depending whether they are solved by similar procedures.

In the analysis, Cohen's *d* was used as a measure of the general effect of the difference in success. This is a recognized effect size for the comparison between means for two different groups. Odds-ratio, was also used as an effect size. Not to compare means, but to determine the association between two variables, which is a common use of odds-ratio. The *t*-test showed significant difference between the successes with respect to the different reasoning types, and Cohen's *d* suggested that the effect of this difference is large enough to consider. Odds-ratios showed that the effects due to the different reasoning types varied quite a lot on the different tasks. Although the effect seemed a bit larger due to CR tasks than due to IR tasks, this was not clearly determined. Further studies of the relation between the various effect sizes and the size of the effect of the dependence are required.

The present analysis of dependence between success on CR tasks and on IR tasks has been conducted on physics tasks, which is a limitation of the study. In order to deepen and generalise the results, continued studies of the dependence should be performed on mathematics tasks. Then, account is taken for possible influence students' understanding of physics has on the result. During the analysis, results indicated that the position of the tasks could influence the dependence of success, and this was accounted for by choosing different reference tasks. It is common in the Swedish test system that tests start with easier tasks and that the difficultness increases with later position. The scores on each task in physics and mathematics tests are labelled to indicate which grades they correspond to. Thus, accounting for the scoring of the tasks may reduce the influence of tasks' difficultness on the result even more.

Another limitation of the study is that it is conducted in a Swedish context. As discussed in the Physics in the Swedish School section, there are alignments with the Norwegian syllabus and a Nordic profile has been identified. Thus the results can be considered interesting to a Scandinavian context. Furthermore, the goals and subject description in the Swedish curriculum are quite rich and highly in accordance with the TIMSS Assessment framework (Garden et al. 2006; Swedish National Agency for Education, 2009). This suggests that the results also are relevant to an international context.

As discussed, there are additional factors to consider in the analysis of the dependence between success on CR tasks and on IR tasks and further studies should be performed in order to make general conclusions. The present results though, give reliable indications of the positive effect of creative mathematical reasoning on task solving. These results might contribute to the discussion about the effect mathematical reasoning has on students' development of knowledge of mathematics as well as of physics.

References

Baroody, A. J., Feil, Y. & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education*, 38 (2), 115–131. Retrieved from http://www.jstor.org/stable/30034952

Basson, I. (2002). Physics and mathematics as interrelated fields of thought development using acceleration as an example. *International Journal of Mathematical Education in Science and Technology*, 33 (5), 679–690. doi: 10.1080/00207390210146023

Bing, T. (2008). *An epistemic framing analysis of upper level physics students' use of mathematics* (Doctoral dissertation). University of Maryland. Retrieved from http://bit.ly/Bing2008

Boesen, J., Lithner, J. & Palm, T. (2010). The relation between types of assessment tasks and the mathematical reasoning student use. *Educational studies in mathematics*, 75 (1), 89–105. doi: 10.1007/s10649-010-9242-9

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction*, 10 (2-3), 105–225. doi: 10.1080/07370008.1985.9649008

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O. et al. (2006). *TIMSS advanced 2008 assessment frameworks*. Boston: TIMSS & PIRLS International Study Center.

Gray, E. M. & Tall, D. O. (1994). Duality, ambiguity and flexibility: a proceptual view of simple arithmetic. *The Journal for Research in Mathematics Education*, 26 (2), 115–141. Retrieved from http://www.jstor.org/stable/749505

Grønmo, L. S. & Onstad. T. (Eds.) (2013). *The significance of TIMSS and TIMSS Advanced mathematics education in Norway, Slovenia and Sweden*. Retrieved from http://bit.ly/TIMSSAdvNo2013

Haylock, D. (1997). Recognising mathematical creativity in schoolchildren. *ZDM*, 29 (3), 68–74. doi: 10.1007/S11858-997-0002-Y

Hiebert, J. & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: an introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: the case of mathematics* (pp. 1–27). Hillsdale: Erlbaum.

Hopkins, W. G (2002). *A scale of magnitudes for effect statistics*. Retrieved from http://www.sportsci.org/resource/stats/ [A new view of statistics]

Johansson, H. (2015). Mathematical reasoning requirements in Swedish national physics tests. *International Journal of Science and Mathematics Education*, 14 (6), 1133–1152. doi: 10.1007/s10763-015-9636-3

Lesh, R. & Zawojewski, J. (2007). Problem solving and modeling. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 763–804). Charlotte: Information Age Publishing.

Lie, S., Angell, C. & Rohatgi, A. (2010). *Fysikk i fritt fall? TIMSS Advanced 2008 i vidaregående skole* [Physics in free fall? TIMSS Advanced 2008 in upper secondary school]. Retrieved from http://bit.ly/TIMSSAdv2008no

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics*, 67 (3), 255–276. doi: 10.1007/s10649-007-9104-2

Michelsen, C. (2005). Expanding the domain – variables and functions in an interdisciplinary context between mathematics and physics. In A. Beckmann, C. Michelsen & B. Sriraman (Eds.), *Proceedings of the 1st international symposium of mathematics and its connections to the arts and sciences* (pp. 201–214). Hildesheim: Verlag Franzbecker.

Nguyen, N.-L. & Meltzer, D. (2003). Initial understanding of vector concepts among students in introductory physics courses. *American Journal of Physics*, 71 (6), 630–638. doi: 10.1119/1.1571831

Palm, T., Boesen, J. & Lithner, J. (2011). Mathematical reasoning requirements in upper secondary level assessments. *Mathematical Thinking and Learning*, 13 (3), 221–246. doi: 10.1080/10986065.2011.564994

Redish, E. F. & Gupta, A. (2009, August). *Making meaning with math in physics: a semantic analysis*. Presented at GIREP 2009, Leicester. Retrieved from https://arxiv.org/pdf/1002.0472.pdf

Redish, E. F. (2003). *Teaching physics with the physics suite*. Hoboken: John Wiley & Sons.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando: Academic Press.

Silver, E. (1997). *Fostering creativity through instruction rich in mathematical problem solving and problem posing*. ZDM, 29 (3), 75–80. doi: 10.1007/s11858-997-0003-x

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 36 (5), 404–411.

Star, J. R. (2007). Foregrounding procedural knowledge. *Journal for Research in Mathematics Education*, 38 (2), 132–135.

Swedish National Agency for Education (2000). *Kursplan för gymnasieskolan – Fysik* [Syllabuses for upper secondary Physics]. Retrieved from http://bit.ly/Fysik2000

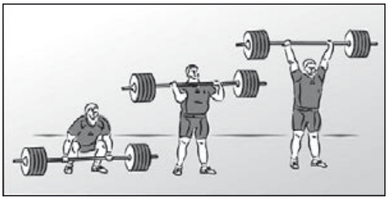
Swedish National Agency for Education (2006). *Curriculum for the non-compulsory school system Lpf 94*. Stockholm: Fritzes.

Swedish National Agency for Education (2009). *Hur samstämmiga är svenska styrdokument och nationella prov med ramverk och uppgifter i TIMSS Advanced 2008?* [How aligned are the Swedish policy documents and national tests with the framework and the tasks in TIMSS Advanced 2008?]. Stockholm: Fritzes.

Appendix

Example A-1

A weightlifter is lifting a barbell that weighs 219 kg. The barbell is lifted 2.1 m up from the floor in 5.0 s.



What is the average power the weightlifter develops on the barbell during the lift?

Analysis

I. Analysis of the assessment task – answers and solutions

A typical solution from an average student could be derived by the relation between power and the change of energy over a specific period of time. In this task, the change of energy is the same as the change of potential energy for the barbell. Multiply the mass of the barbell by the acceleration of gravity and the height of the lift and then divide by the time to get the power asked for. The mathematical subject area is identified as algebra, in this case working with formulas. The identification of the situation to lift a barbell can trigger the student to use a certain solution method and is, therefore, included in this analysis as an identified "real-life" situation.

II. Analysis of the assessment task – task variables

The assignment is to calculate the average power during the lift. The mass of the barbell, the height of the lift, and the time for the lift are all considered as mathematical objects. In this example, all of the objects are given explicitly in the assignment in numerical form. In the presentation of the task, there is also an illustrative figure of the lift.

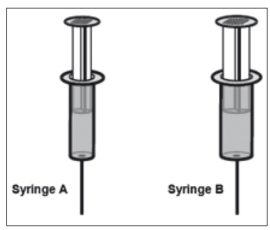
III. Analysis of the textbooks and handbook – answers and solutions

Handbook: Formulas for power, $P = \Delta W/\Delta t$, with the explanation "ΔW = the change in energy during time Δt"; for "work during lift", $W_l = mg \cdot h$, with the explanatory text, "A body with weight mg is lifted to a height h. The lifting work is ..."; and for potential energy with the text "A body with mass m at a height h over the zero level has the potential energy $W_p = mg \cdot h$". *Mathematics book*: Numerous examples and exercises of how to use formulas, e.g. on pages 28–30. *Physics book*: Power is presented as work divided by time, and in on example work is exemplified as lifting a barbell. An identical example is found on page 130. An example of calculating work during a lift in relation to change in potential energy is found on page 136. Exercises 5.05 and 5.10 are solved by a similar algorithm.

IV. Argumentation for the requirement of reasoning
The analysis of the textbooks shows that there are more than three tasks similar to the task being categorised with respect to the task variables, and these tasks can be solved with a similar algorithm. If the students have seen tasks solvable by a similar algorithm at least three times, it is assumed that they will remember the solution procedure. This task is then categorised as solvable using IR.

Example A-2

A patient is going to get an injection. The medical staffs are reading in the instructions that they are supposed to use a syringe that gives the lowest pressure as possible in the body tissue. Which of the syringes A or B shall the staff choose if the same force, F, is applied and the injection needles have the same dimensions? *Argue for the answer*



Analysis
I. Analysis of the assessment task – answers and solutions
To solve this task, the student can use the relation between pressure, force, and area (p=F/A). Neglecting the hydrostatic pressure from the injection fluid, if the force applied to the syringe is the same then it is the area of the bottom that affects the pressure. The larger the area, the lower the pressure. The staff should choose syringe B. The mathematical subject area is identified as algebra, such as to work with formulas and proportionality.

II. Analysis of the assessment task – task variables
The assignment is to choose which syringe that gives the minimum pressure and to provide an argument for this choice. Only the force is given as a variable, and this is represented by a letter. Key words for the students

can be *force* and *pressure*. The situation is illustrated by a figure in which it appears that syringe B has a greater diameter than syringe A.

III. Analysis of the textbooks and handbook – answers and solutions
Handbook: The relation p=F/A is defined. *Mathematics book*: Proportionalities are discussed and exemplified but are not used for general comparisons. *Physics book*: One example about how different areas affect the pressure and one exercise that is solved in a similar way by using a general comparison between different areas and pressure.

IV. Argumentation for the requirement of reasoning
There is only one example and one exercise that can be considered similar with regard to the task variables and the solution algorithm. The formula is in the handbook, but there has to be some understanding of the intrinsic properties in order to be able to use the formula in the solution. This task is, therefore, considered to require some CR, in this case GCR, in order to be solved.

Table A-1. *Ratios according to (1) for the tasks in the Physics A 2004 test with the tasks 11 and 8a as reference tasks*

Task	1 (IR)	2 (IR)	3 (IR)	4 (LCR)	5a (NMR)	5b (IR)	6a (NMR)	6b (NMR)	7a (GCR)	7b (NMR)	8a (IR)	8b (IR)	9 (IR)	10 (NMR)	11 (GCR)	12 (GCR)	13 (LCR)	14 G (GCR)	14 VG (GCR)
11	1.17	1.24	1.39	1.45		1.59			1.27		1.38	1.32	1.83		R	2.28	1.91	1.49	3.06
8a	1.12	1.14	1.23	1.24		1.35			1.19		R	1.37	1.37		1.38	1.18	1.53	1.23	1.59

Table A-2. *Ratios according to (1) for the tasks in the Physics A 2005 test with the tasks 10 and 8a as reference tasks*

Task	1 (IR)	2a (NMR)	2b (IR)	3a (IR)	3b (NMR)	4 (NMR)	5 (LCR)	6a (LCR)	6b (LCR)	7 (NMR)	8a (IR)	8b (IR)	9 (IR)	10 (GCR)	11 (IR)	12a (IR)	12b (LCR)	13 (GCR)	14 (LCR)	15 (NMR)
10	1.20		1.24	1.18			1.20		1.53		1.03	1.16	1.43	R	1.65	1.32	1.50	1.45	1.91	
8a	1.02		1.02	1.02			1.03		1.02		R	1.03	1.03	1.03	1.03	1.02	1.03	1.03	1.03	

Table A-3. *Ratios according to (1) for the tasks in the Physics B 2002 test with the tasks 10 and 7 as reference tasks*

Task	1 (NMR)	2 (IR)	3 (IR)	4 (IR)	5 (IR)	6a (LCR)	6b (LCR)	7 (IR)	8 (LCR)	9a (IR)	9b (IR)	10 (GCR)	11 (LCR)	12a (NMR)	12b (LCR)	13 (GCR)	14 (GCR)	15 (GCR)
10		1.36	1.31	1.23	1.69	1.61	1.67	1.54	1.83	1.33	1.59	R	1.69		2.38	2.12	2.21	2.29
7		1.25	1.20	1.18	1.48	1.39	1.42	R	1.38	1.13	1.39	1.49	1.42		1.81	1.51	1.65	1.61

Table A-4. *Ratios according to (1) for the tasks in the Physics B 2003 test with the tasks 8 and 7 as reference tasks*

Task	1 (IR)	2 (IR)	3 (NMR)	4 (IR)	5 (IR)	6 (NMR)	7 (IR)	8 (LCR)	9 (LCR)	10a (LCR)	10b (NMR)	11a (NMR)	11b (IR)	12 (IR)	13 (IR)	14a (NMR)	14b (IR)	15 (GCR)	16 (GCR)
8	1.25	1.11		0.98	1.28		1.19	R	1.53	1.34			1.27	1.47	1.51		1.34	1.47	1.83
7	1.14	1.07		0.99	1.14		R	1.22	1.22	1.17			1.13	1.21	1.23		1.16	1.24	1.30

Table A-5. *Ratios according to (1) for the tasks in the Physics B 2005 test with the tasks 12b and 8b as reference tasks*

Task	1 (NMR)	2 (NMR)	3 (LCR)	4a (NMR)	4b (NMR)	5 (NMR)	6 (IR)	7 (GCR)	8a (NMR)	8b (IR)	9a (IR)	9b (NMR)	10 (GCR)	11 (IR)	12a (IR)	12b (GCR)	12c (IR)	13 (LCR)	14 (GCR)	15a (NMR)	15b (IR)	15c (LCR)	16 (LCR)
12b			1.11				1.07	1.16		1.11	1.10		1.18	1.31	1.17	R	1.35	1.16	1.32		1.11	1.30	1.27
8b			1.34				1.11	1.30		R	1.21		1.45	1.61	1.11	1.11	1.25	1.31	1.54		1.19	1.57	1.61

Table A-6. *Ratios according to (1) for the tasks in the Physics B 2006 test with the tasks 12a and 10b as reference tasks*

Task	1 (NMR)	2 (IR)	3 (NMR)	4 (GCR)	5 (LCR)	6 (NMR)	7 (NMR)	8a (IR)	8b (IR)	9 (IR)	10a (NMR)	10b (IR)	11 (NMR)	12a (GCR)	12b (GCR)	12c (LCR)	13 (IR)	14a (NMR)	14b (IR)	15 (NMR)	16 G (IR)	16 VG (NMR)
12a		1.21		1.67	1.39			1.36	1.38	1.35		1.27		R	2.21	2.07	1.66		1.64		1.29	
10b		1.08		1.24	1.22			1.21	1.32	1.34		R		1.27	1.35	1.42	1.44		1.38		1.91	

Table A-7. *Ratios according to (1) for the tasks in the Physics B 2010 test with the tasks 11b and 9b as reference tasks*

Task	1 (NMR)	2 (NMR)	3 (IR)	4a (IR)	4b (NMR)	5 (IR)	6 (IR)	7a (IR)	7b (GCR)	8 (LCR)	9a (NMR)	9b (IR)	9c (GCR)	10 (GCR)	11a (LCR)	11b (GCR)	11c (GCR)	12 (GCR)	13 (IR)	14 G (LCR)	14 VG (GCR)
11b			1.24	1.25		1.25	1.50	1.23	1.45	1.39		1.43	1.70	1.93	1.36	R	2.69	1.91	1.85	1.80	2.47
9b			1.10	1.08		1.22	1.16	1.14	1.25	1.25		R	1.59	1.51	1.14	1.43	1.52	1.45	1.44	1.33	1.70

Table A-8. *Ratios according to (1) for the Physics A 2004 test with the new task 7a together with tasks 11 and 8a as reference tasks*

Task	1 (IR)	2 (IR)	3 (IR)	4 (LCR)	5a (NMR)	5b (IR)	6a (NMR)	6b (NMR)	7a (GCR)	7b (NMR)	8a (IR)	8b (IR)	9 (IR)	10 (NMR)	11 (GCR)	12 (GCR)	13 (LCR)	14 G (GCR)	14 VG (GCR)
11	1.17	1.24	1.39	1.45		1.59			1.27		1.38	1.32	1.83		R	2.28	1.91	1.49	3.06
8a	1.12	1.14	1.23	1.24		1.35			1.19		R	1.37	1.37		1.38	1.18	1.53	1.23	1.59
7a	1.10	2.22	1.17	1.24		1.24			R		1.19	1.18	1.31		1.27	1.45	1.37	1.26	

Table A-9. *Ratios according to (1) for the Physics A 2005 test with the new task 12a together with tasks 10 and 8a as reference tasks*

Task	1 (IR)	2a (NMR)	2b (IR)	3a (IR)	3b (NMR)	4 (NMR)	5 (LCR)	6a (LCR)	6b (LCR)	7 (NMR)	8a (IR)	8b (IR)	9 (IR)	10 (GCR)	11 (IR)	12a (IR)	12b (LCR)	13 (GCR)	14 (LCR)	15 (NMR)
10	1.20		1.24	1.18			1.20		1.53		1.03	1.16	1.43	R	1.65	1.32	1.50	1.45	1.91	
8a	1.02		1.02	1.02			1.03		1.02		R	1.03	1.03	1.03	1.03	1.02	1.03	1.03	1.03	
12a	1.15		1.17	1.14			1.14		1.35		1.02	1.12	1.33	1.32	1.42	R	1.41	1.33	1.59	

Table A-10. *Ratios according to (1) for the Physics B 2002 test with the new task 9b together with tasks 10 and 7 as reference tasks*

Task	1 (NMR)	2 (IR)	3 (IR)	4 (IR)	5 (IR)	6a (LCR)	6b (LCR)	7 (IR)	8 (LCR)	9a (IR)	9b (IR)	10 (GCR)	11 (LCR)	12a (NMR)	12b (LCR)	13 (GCR)	14 (GCR)	15 (GCR)
10		1.36	1.31	1.23	1.69	1.61	1.67	1.54	1.83	1.33	1.59	R	1.69		2.38	2.12	2.21	2.29
7		1.25	1.20	1.18	1.48	1.39	1.42	R	1.38	1.13	1.39	1.49	1.42		1.81	1.51	1.65	1.61
9b		1.30	1.25	1.25	1.56	1.53	1.51	1.53	1.51	1.42	R	1.70	1.55		2.25	2.21	2.07	1.68

Table A-11. *Ratios according to (1) for the Physics B 2005 test with the new task 7 together with tasks 12b and 8b as reference tasks*

Task	1 (NMR)	2 (NMR)	3 (LCR)	4a (NMR)	4b (NMR)	5 (NMR)	6 (IR)	7 (GCR)	8a (NMR)	8b (IR)	9a (IR)	9b (NMR)	10 (GCR)	11 (IR)	12a (IR)	12b (GCR)	12c (IR)	13 (LCR)	14 (GCR)	15a (NMR)	15b (IR)	15c (LCR)	16 (LCR)
12b			1.11				1.07	1.16		1.11	1.10		1.18	1.31	1.17	R	1.35	1.16	1.32		1.11	1.30	1.27
8b			1.34				1.11	1.30		R	1.21		1.45	1.61	1.11	1.11	1.25	1.31	1.54		1.19	1.57	1.61
7			1.24				1.14	R		1.30	1.21		1.47	1.66	1.11	1.16	1.28	1.35	1.62		1.19	1.59	1.60

Table A-12. *Ratios according to (1) for the Physics B 2006 test with the new task 4 together with tasks 12a and 10b as reference tasks*

Task	1 (NMR)	2 (IR)	3 (NMR)	4 (GCR)	5 (LCR)	6 (NMR)	7 (NMR)	8a (IR)	8b (IR)	9 (IR)	10a (NMR)	10b (IR)	11 (NMR)	12a (GCR)	12b (GCR)	12c (LCR)	13 (IR)	14a (NMR)	14b (IR)	15 (NMR)	16 G (IR)	16 VG (NMR)
12a		1.21		1.67	1.39			1.36	1.38	1.35		1.27		R	2.21	2.07	1.66		1.64		1.29	
10b		1.08		1.24	1.22			1.21	1.32	1.34		R		1.27	1.35	1.42	1.44		1.38		1.91	
4		1.14		R	1.3			1.25	1.32	1.27		1.24		1.67	1.62	1.54	1.62		1.81		2.37	

Table A-13. *Ratios according to (1) for the Physics B 2010 test with the new task 7b together with tasks 11b and 9b as reference tasks*

Task	1 (NMR)	2 (NMR)	3 (IR)	4a (IR)	4b (NMR)	5 (IR)	6 (IR)	7a (IR)	7b (GCR)	8 (LCR)	9a (NMR)	9b (IR)	9c (GCR)	10 (GCR)	11a (LCR)	11b (GCR)	11c (GCR)	12 (GCR)	13 (IR)	14 G (LCR)	14 VG (GCR)
11b			1.24	1.25		1.25	1.50	1.23	1.45	1.39		1.43	1.70	1.93	1.36	R	2.69	1.91	1.85	1.80	2.47
9b			1.10	1.08		1.22	1.16	1.14	1.25	1.25		R	1.59	1.51	1.14	1.43	1.52	1.45	1.44	1.33	1.70
7b			1.18	1.19		1.19	1.17	1.60	R	1.30		1.25	1.52	1.83	1.18	1.45	1.57	1.64	1.50	1.46	1.98

Table A-14. *Odds-ratios according to (2) for the tasks in the Physics A 2004 test with tasks 11, 8a and 7a as reference tasks*

Task	1 (IR)	2 (IR)	3 (IR)	4 (LCR)	5a (NMR)	5b (IR)	6a (NMR)	6b (NMR)	7a (GCR)	7b (NMR)	8a (IR)	8b (IR)	9 (IR)	10 (NMR)	11 (GCR)	12 (GCR)	13 (LCR)	14 G (GCR)	14 VG (GCR)
11	2.84	4.06	4.44	3.30		3.63			2.65		2.55	2.19	3.38		R	5.23	4.55	2.97	24.3
8a	2.23	2.44	2.63	2.19		2.65			2.31		R	3.15	2.08		2.55	1.39	3.29	1.87	2.88
7a	2.47	2.46	2.58	2.94		2.61			R		2.31	2.15	2.57		2.65	4.08	3.54	2.87	–

Table A-15. *Odds-ratios according to (2) for the tasks in the Physics A 2005 test with tasks 10, 8a and 12a as reference tasks*

Task	1 (IR)	2a (NMR)	2b (IR)	3a (IR)	3b (NMR)	4 (NMR)	5 (LCR)	6a (NMR)	6b (LCR)	7 (NMR)	8a (IR)	8b (IR)	9 (IR)	10 (GCR)	11 (IR)	12a (IR)	12b (LCR)	13 (GCR)	14 (LCR)	15 (NMR)
10	1.73		4.23	2.86			1.80		3.26		3.14	2.94	2.64	R	3.89	2.79	3.26	2.51	5.62	
8a	2.48		2.69	3.40			2.98		1.93		R	6.35	3.29	3.14	3.20	1.98	3.11	2.43	2.38	
12a	1.65		2.87	2.54			1.63		2.56		1.98	2.44	2.48	2.79	2.94	R	3.29	2.27	4.06	

Table A-16. *Odds-ratios according to (2) for the tasks in the Physics B 2002 test with tasks 10, 7 and 9b as reference tasks*

Task	1 (NMR)	2 (IR)	3 (IR)	4 (IR)	5 (IR)	6a (LCR)	6b (LCR)	7 (IR)	8 (LCR)	9a (IR)	9b (IR)	10 (GCR)	11 (LCR)	12a (NMR)	12b (LCR)	13 (GCR)	14 (GCR)	15 (GCR)
10		2.84	3.14	2.53	3.92	4.50	5.43	4.06	4.49	3.67	3.53	R	4.26		6.44	4.32	5.03	5.31
7		2.69	2.66	2.97	4.70	3.90	4.46	R	3.71	2.51	4.18	4.09	3.64		10.2	4.15	5.72	5.95
9b		2.31	2.59	2.99	3.74	4.14	4.11	4.65	3.30	4.36	R	3.91	3.66		7.94	6.72	5.47	4.27

Table A-17. *Odds-ratios according to (2) for the tasks in the Physics B 2003 test with the tasks 8 and 7 as reference tasks*

Task	1 (IR)	2 (IR)	3 (NMR)	4 (IR)	5 (IR)	6 (NMR)	7 (IR)	8 (LCR)	9 (LCR)	10a (LCR)	10b (NMR)	11a (NMR)	11b (IR)	12 (IR)	13 (IR)	14a (NMR)	14b (IR)	15 (GCR)	16 (GCR)
8	2.63	2.95		1.98	4.17		4.34	R	5.94	2.75			4.03	5.68	4.51		4.64	3.80	17.3
7	2.55	2.94		2.61	3.27		R	4.61	3.85	2.62			3.95	4.40	3.47		3.56	6.52	6.97

Table A-18. *Odds-ratios according to (2) for the tasks in the Physics B 2005 test with tasks 12b and 8b and 7 as reference tasks*

Task	1 (NMR)	2 (NMR)	3 (LCR)	4a (NMR)	4b (NMR)	5 (NMR)	6 (IR)	7 (GCR)	8a (NMR)	8b (IR)	9a (IR)	9b (NMR)	10 (GCR)	11 (IR)	12a (IR)	12b (GCR)	12c (IR)	13 (LCR)	14 (GCR)	15a (NMR)	15b (IR)	15c (LCR)	16 (LCR)
12b			1.57				2.18	2.03		1.73	1.79		1.92	2.57	4.88	R	3.41	2.11	2.67		1.77	2.50	2.23
8b			3.25				3.09	3.10		R	2.71		4.03	5.37	2.28	1.73	2.00	3.41	4.20		2.46	4.62	5.09
7			2.00				3.89	R		3.10	2.42		3.27	4.01	2.22	2.03	1.89	3.29	3.65		2.12	3.45	3.38

Table A-19. *Odds-ratios according to (2) for the tasks in the Physics B 2006 test with tasks 12a, 10b and 4 as reference tasks*

Task	1 (NMR)	2 (IR)	3 (NMR)	4 (GCR)	5 (LCR)	6 (NMR)	7 (NMR)	8a (IR)	8b (IR)	9 (IR)	10a (NMR)	10b (IR)	11 (NMR)	12a (GCR)	12b (GCR)	12c (LCR)	13 (IR)	14a (NMR)	14b (IR)	15 (NMR)	16 G (IR)	16 VG (NMR)
12a		3.12		3.40	3.39			3.76	2.58	2.59		2.34		R	19.3	8.35	3.41		3.82		1.67	
10b		1.70		1.93	2.54			2.71	3.06	3.70		R		2.34	3.03	3.42	3.53		3.19		3.66	
4		1.88		R	2.64			2.14	1.95	1.88		1.93		3.40	3.10	2.33	2.54		3.91		3.44	

Table A-20. *Odds-ratios according to (2) for the tasks in the Physics B 2010 test with tasks 11b, 9b and 7b as reference tasks*

Task	1 (NMR)	2 (NMR)	3 (IR)	4a (IR)	4b (NMR)	5 (IR)	6 (IR)	7a (IR)	7b (GCR)	8 (LCR)	9a (NMR)	9b (IR)	9c (GCR)	10 (GCR)	11a (LCR)	11b (GCR)	11c (GCR)	12 (GCR)	13 (IR)	14 G (LCR)	14 VG (GCR)
11b			8.36	6.51		3.77	3.86	2.12	3.03	3.86		4.25	3.34	4.80	16.9	R	41.8	4.46	4.84	5.68	10.4
9b			2.31	1.96		4.60	1.83	1.91	2.41	3.24		R	5.89	4.22	2.66	4.25	4.64	3.36	3.64	2.88	7.51
7b			4.01	3.94		2.86	1.61	11.8	R	3.07		2.41	2.76	5.15	3.02	3.03	3.22	3.32	2.83	2.98	5.61

Helena Johansson

Helena Johansson has a PhD in Mathematics, specialising in Educational Sciences and is a postdoc at Mid Sweden University. Her research interests concern students' mathematical reasoning and how this competence influences students' learning in mathematics and in physics; and how natural language influences students' learning of the symbolic language of mathematics.

helena.johansson@miun.se

Figurative Context and Mathematical Reasoning Requirements – Influences on Students' Success on tasks in National Mathematics Tests

Helena Johansson

Within the field of mathematics education there seem to exist common assumptions concerning the value of introducing real life contexts (figurative contexts) in mathematics tasks. In this study, the influence of figurative context on upper secondary students' success in test tasks is explored. The study also departs from earlier studies concerning mathematical reasoning requirements when students solve national mathematics tests. Data consist of tasks from six Swedish national tests, as well as students' results (number of students $829 \leq n \leq 3481$). Each task is categorised as having a figurative context or not, and if creative mathematical reasoning is required or not. Both descriptive statistics and significance testing have been used for the analyses. The results indicate that context had a positive effect on students' success if the task required creative mathematical reasoning and that this effect was higher for students with lower grades.

Keywords: Differential item functioning, Figurative context, Mathematical reasoning, Upper secondary school.

Introduction

Mathematics is a subject with many fields, and the character of the various fields is in some instances very different. Sometimes it is pure interest (as in basic research) that pushes the development forward, as was the case for many years in Number Theory, and other times it is the applicability that is the driving force of the research, as e.g. in Financial Mathematics. Some of the purely theoretical discoveries can turn out to have practical applications, such as prime numbers have had for encryption that today is a part of the modern man's everyday life. No matter which mathematical field that is considered, Hirsch (1996) discusses in his article that all ideas, even the most abstract ones, ultimately origin from real-life experience. The reason for the strength of mathematics to describe various phenomena is that it has evolved to fit the analysis process, e.g. to separate components and reduce all unnecessary information (Hirsch, 1996). This can be regarded as a decontextualisation of the situation. Nevertheless, at some instances theories need to be experimentally verified and the mathematics is then used in a context.

Formally defined concepts form the basis of mathematics, but before the concepts are defined it is not unusual that they have been experienced in various forms (Tall & Vinner, 1981). This relates to Hirsch's discussion mentioned above. How various people understand a concept depends on the individual's mental pictures and associated properties and processes, which arise from different types of experience. Tall & Vinner use *concept image* to describe the cognitive structure forming the understanding of a concept. Seeing mathematics in context could thus be regarded as helpful when concept images are formed. In school, students have to consider this duality of mathematics. Sometimes it is necessary to see situations in which mathematics is applied in order to understand the concepts. Other times it is necessary to e.g. ignore some context presented in a task and reduce the task to pure mathematic in order to solve it.

Boaler (1994) discusses how bringing everyday contexts to mathematics tasks was motivated by that it would help bridging the abstract world of mathematics and students' real world outside the classroom. She separates the different reasons given for introducing contexts in the learning in three categories; 1. Learning is thought to be more accessible if students are given familiar metaphors. 2. Students are thought to become more motivated to learn if they are provided with examples enriching the curriculum. 3. Transfer of mathematical learning is thought to benefit from linking real world problems to school mathematics. The motivating argument, 2. above, is also noticed by Cooper & Dunne (2000) when they account for how the school mathematics was related to the real world. They also saw another reason, the beneficial aspect, that the mathematics should be relevant to what the students were supposed to need in their upcoming career and life. Motivation as a reason for tasks with realistic and everyday contexts is also used by Howson (2005) when he discusses school mathematics with meaning. The third category of the reasons discussed by Boaler (1994) could be related to Basson's (2002) paper, in which he discusses that it is valuable to relate mathematical concepts to relevant context from the students' real world in order to accomplish a more general understanding of the concepts.

In Boaler (1994) she reviewed a small-scaled piece of research and the findings suggest that contexts in some instances affect boys and girls differently. It is more likely that boys perform better than girls if the context in a task includes real world variables that should not be used nor taken into account to reach a solution. Boaler further discusses how this ignoring of the real world may be a reason to girls disinterest in mathematics. According to the TIMSS Advanced 2008 report, there are significant differences in how boys and girls succeeded on the mathematics test in many of the participating countries. In all except one country, the difference is in boys' favour (Swedish National Agency for Education, 2009). The analysis of PISA 2012 shows similar results; there is a significant difference in boys' favour regarding success on the test in mathematics when accounting for all 65 participating countries. The differences vary though within countries; in 23 of the countries no gender gap is observed and in five of the countries girls outperform boys (OECD, 2014). As mentioned in for example Ramstedt (1996) or Sumpter (2012, 2015), questions regarding gender differences can be viewed from many different perspectives, e.g. psychological, biological, sociological, historical etc., depending on the intention of the studies.

Inspired by the various studies discussed above, it would be interesting to know if and how the context influences students' success in test situations.

Context

Context is one of those concepts that are used with an unequivocal meaning in the education literature. It can for example be used in order to describe the overall situation as a 'school context' or an 'out of school context' (e.g. Verschaffel, De Corte & Lasure, 1994; and Nunes, Schliemann & Carraher, 1993); or for describing various aspects of the learning situation in the classroom (Shimizu, Kaur, Huang & Clarke, 2010; Wyndhamn, 1993). Other times context is used to denote a subject (Hanna et al. 2004) or various areas of a subject (Doorman & Gravemeijer's, 2009). Context is further used to denote students' expectations (e.g. Bassok & Holyoak, 1989; and Bing, 2008).

In the present study, the focus is on how context influences success in solving mathematical tasks. Thus, how context is used by scholars discussing tasks in relation to mathematics education is of primarily interest. Shimizu et al. (2000) discuss how tasks have a central place in the mathematics classroom instructions. Individual teachers' choice of various tasks thus influences students' understanding of the mathematics that is taught. 'Figurative context' is used when an authentic situation is described in a task, see e.g. Palm (2002). Lobato, Rhodehamel and Hohensee (2012) use the single word 'context' to describe the situation posed in the task; for example, a hose is used to fill a pool and the amount of water is graphed with respect to time. The task is to find and interpret the slope in the graph. Marongelle (2004) refers to Kulm (1984) and concludes that when context is used in the literature about mathematical problem solving, it often refers to the non-mathematical meanings that are present in the problem situation. This is similar to what Verschaffel et al. (1994) call 'problem context', a task embedded in some kind of described reality.

Verschaffel et al. (1994), as well as Boaler (1994) and Cooper and Dunne (2000), conclude that text-tasks in school mostly are artificial, and, contrary to the intention, that it sometimes may be negative for students to use their common-sense knowledge as one usually does in real-life problems. Contexts in tasks that are posed in such way that students have to ignore what would have happened in real-life in order to provide the correct answer are called 'pseudo-real contexts' by Boaler (1994). She further proposes a somewhat different approach to how context may be used from what was common at that time. Instead of using context to present students various specific real-life situations in which certain mathematics can be used, context can be valuable for giving student real-life situations that they have to reflect upon. Boaler concludes from her study that context describing real-life situations only is valuable if the described real-life variables have to be taken into account to solve the posed question.

Palm (2002) discusses how different answers to tasks describing real-life events could be considered correct/reasonable, depending on how the students interpret the purpose of a particular task/question. For example, the decision of how many buses that are required in order to go on a school-excursion if each bus have 40 seats and there are 540 students and teachers in total at the school. Palm argues for how three possible answers (13, 13.5 and 14) can be regarded as correct solutions depending on how the purpose is interpreted. Is for instance the purpose to use the solution as information to decide how many buses that are required or is it the actual number of buses that one wants to order from the bus company that should be given as an answer. Further, should the student account for if more than one child can sit in one seat; and other real life considerations? Palm modified this "bus-task" to a more authentic variant by including an order slip to the bus company, which the students should fill out as an answer. The result showed that 97 % of the students reflected about and discarded the "half" bus answer, compared to 84 % of the students who solved the original task.

Bergqvist & Lind (2005) investigate whether a change in context on some mathematical tasks affects how students succeed on the same tasks. In their study, tasks are categorised as either 'intra-mathematical' or as 'having contexts'.

Central Concepts

Henceforth in this paper, tasks are said to have a context when a figurative context is present, and tasks are said to be intra-mathematical when there are no real-life situations described in the tasks.

Mathematical reasoning

The impact of mathematical reasoning on mathematical learning has been discussed and studied from multiple perspectives. Schoenfeld (1992), for example, points out that a focus on rote mechanical skills leads to poor performance in problem solving. Lesh & Zawojeskij (2007) discuss how emphasising low-level skills does not give the students the abilities needed for mathematical modelling or problem solving, neither to draw upon interdisciplinary knowledge. Students lacking the ability to use creative mathematical reasoning thus get stuck when confronted with novel situations, and this hamper their possibilities to learn (Lithner, 2008).

Lithner (2008) has, through empirical studies on how students engage in various kinds of mathematical activities, developed a framework that defines mathematical reasoning and different reasoning categories. This framework distinguishes between *creative mathematical founded reasoning* (CR) and *imitative reasoning* (IR). To be regarded as CR the following criteria should be fulfilled "i. Novelty. A new reasoning sequence is created or a forgotten one is recreated. ii. Plausibility. There are arguments supporting the strategy choice and/or strategy implementation motivating why the conclusions are true or plausible. iii. Mathematical foundation. The arguments made during the reasoning process are anchored in the intrinsic mathematical properties of the components involved in the reasoning" (Lithner, 2008, p.266).

Reasoning categorised as IR fulfils "i. The strategy choice is founded on recalling a complete answer. ii. The strategy implementation consists only of writing it down" (Lithner, 2008, p. 258), or "i. The strategy choice is to recall a solution algorithm. The predicted argumentation may be of different kind, but there is no need to create a new solution. ii. The remaining parts of the strategy implementation are trivial for the reasoned, only a careless mistake can lead to failure" (Lithner, 2008, p.259).

Methods

The tasks used in the analyses in this study come from six Swedish national mathematics tests for three consecutive mathematics courses for upper secondary school; Mathematics B, C and D from December 2003; and Mathematics B, C and D from May 2004. Each task has previously been categorised with respect to mathematical reasoning requirements, i.e. IR or CR, (Palm, Boesen & Lithner, 2011). In addition to the tests, various information about the students' that have taken the tests are available. The student data contain information about students' score on each task, their total test score, their grade on the test and their course grade, as well as their school, their gender, if Swedish is mother-tongue or not and their attained programme. The number of students varies for the six different tests, from 829 to 3481.

Below a description of the national testing programme first follows, and after that an outline of the categorisation of the tasks according to context and mathematical reasoning. Thereafter comes a specification of the research questions, followed by the adherent methods and analyses.

National Tests in Mathematics

In the curricula from 1994 the goals for mathematics was changed to focus more on students' competency to argue for their solutions and to draw conclusions. The descriptions of the goals and the different grade levels are quite brief and the intention is that the curricula should be processed, interpreted and refined locally at each school. By reflecting on how knowledge is viewed in the curricula, the national tests have several aims and two of them are to concretise the governmental goals and grade criteria, and to support equal assessment and fair grading (Ministry of Education, 2007). One aspect of knowledge that the curricula focus on, is that school should take advantage of the knowledge and the experience students bring from "out-of-school" reality (Swedish National Agency of Education, 2006). Thus, it is desirable that the tasks in the national tests should contain a realistic and/or motivating context.

National tests in mathematics are compulsory for upper secondary students. The tests are developed by the Department of Applied Educational Science at Umeå University, which has had this commission since shortly after the new national curricula was implemented in 1994. The national mathematics tests begin with seven to eight tasks that are to be solved without using any equipment other than a pencil. For the following eight to nine tasks a calculator is allowed. Students are provided with a "formula sheet", containing some mathematical formulas the students do not have to remember. One of the tasks in the tests, often the last one, is an "aspect-task" that is assessed according to different aspects, e.g. choice of method, accomplishment, mathematical reasoning and use of concepts. This task should be easy to start with, but it should also include a challenge to more proficient students.

Categorising Tasks According to Context

In order to say anything about whether contexts in tasks affect students' success, it is desirable to analyse the tasks from various perspectives and with different methods. Before the analyses of the success could start, the tasks had to be divided into different groups according to if a figurative context was present or if the tasks were intra-mathematical. The tasks were further grouped with respect to required mathematical reasoning, CR or IR. The categories, with respect to which the tasks will be analysed with respect to, are: *context task(s)*–tasks with a figurative context, *intraMath task(s)*–tasks without a figurative context, *CR-C task(s)*–context tasks requiring CR, *CR-M task(s)*–intraMath tasks requiring CR, *IR-C task(s)*–context tasks solvable with IR, and *IR-M task(s)*–intraMath-tasks solvable with IR. It is further noticed whether the figurative context in respective context task is a real context or a pseudo-real context (cf. Boaler, 1994).

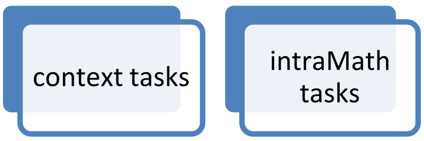


Figure 1. Tasks grouped according to the presence of figurative context**.**

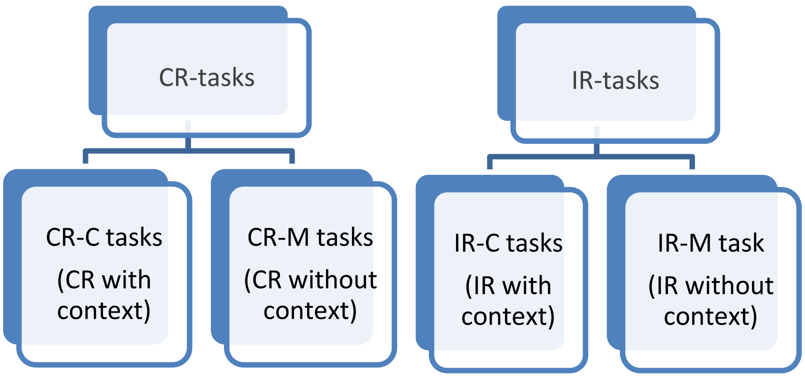


Figure 2. Overview of the subdividing of tasks according to mathematical reasoning and presence of figurative context.

All the tasks that are analysed are solved in a test situation in a school context. The test situations are assumed to be approximately the same for all students. It is further assumed that the setting as a test situation influences average students in similar ways, i.e. it is a test situation and the students' intentions are to manage as well as they can. Therefore, these aspects of the settings are considered to be fixed in this study.

There are settings/factors that do vary, and thus could be important to consider in the analyses. One of them is the mathematics course respective test assesses. It is assumed that students prepare themselves by studying the relevant areas of mathematics they know will be tested. Because of what is known about the testing system, it is further assumed that students expect that there will not be any tasks assessing any other areas of mathematics than are specified beforehand. Another one of the factors is a task's position in the test. It is known that the position influences students' expectations regarding whether the task is assumed to be easy or more difficult. The character of the tasks vary depending on whether a calculator is allowed or not, and if the task is an aspect task. These are further factors worth considering. During the categorisation of the tasks, notes are thus taken about "mathematics course", "test year", "task placement", "calculator" or "no calculator". At the same time it is also identified which mathematical area is involved in the task, e.g. if it is to solve a quadratic equation or maybe to estimate the probability of an event.

With the tasks grouped as described above and illustrated in Figure 1 and Figure 2, the analysis process started with the first two research questions specified below. The outcome from these questions gave rise to the succeeding questions, 3 to 7. Since students' grades and gender were available through the data, these were the factors decided to account for.

Research Questions

1. Does the presence of figurative context influence the solution rates on mathematics tasks?
2. Does the presence of figurative context influence the solution rates on mathematics tasks when mathematical reasoning requirements are taken into account?
3. Are there significant differences in students' solving rate on CR-C tasks and on CR-M tasks, and solving rate on IR-C tasks and on IR-M tasks?
4. Does the presence of figurative context have different influences on students' success depending on their grades?

5. **5.** Does the presence of figurative context have different influences on students' success depending on their grades if required mathematical reasoning is taken into account?
6. **6.** Does the presence of figurative context have different influences on students' success depending on their grades if gender is taken into account?
7. **7.** Does the presence of figurative context influence girls and boys significantly differently with account taken for that they have the same mathematical ability?

Different methods are used in order to answer the questions above. Below follows a description of respective method and performed analysis. The numbers in the headlines refer to the research questions (RQ).

Comparing solution rates (RQ 1 and 2)

Each test was analysed separately. For every task, the number of students who had solved the task partly or completely was divided by the total number of students who had tried to solve the task. This resulted in a solution rate for each task. For example, consider a task with the maximum score 2. Assume that for this particular task there are 978 students who have got 1 or 2 scores, and there are 1257 students in total, who have tried to solve the task, i.e. students who have 0, 1 or 2 scores. Then the solution rate for that task is 978/1257=0.778.

The tasks, separated by test, were then grouped according to the six categories listed above, and a mean solution rate for every category was calculated. If there are for example 7 intraMath tasks on a test, the solution rates for these tasks are summed and divided by 7. For every test there were now a mean solution rate for each of the six categories; context tasks, intraMath tasks, CR-C task, CR-M tasks, IR-C tasks and IR-M tasks. These different mean solution rates were then compared in order to see if the presence of figurative context could be a reason to any differences.

Paired Sample T-Test (RQ 3)

For the quantitative analysis and significance testing of how figurative context might influence students' success on tasks, the results for individual students were required. For each student, individual solving rates were calculated with respect to the four different sub-categories; CR-C task, CR-M tasks, IR-C tasks and IR-M tasks. For example, to calculate the solving rate on CR-C tasks for a particular student, the student's scores on all CR-C tasks is summed and then divided by the total scores possible to obtain by solving all CR-C tasks. So if the student has got 15 out of 18 of the scores for the CR-C tasks, the student's solving rate for CR-C tasks is 15/18=0.83.

The paired T-test was used for hypothesis testing of the difference between students' means of the solving rates for the pairs CR-C and CR-M tasks, and IR-C and IR-M tasks. The tested null hypothesis is: H_0: the mean value of the differences between the pairs is zero. In order to use a parametric test, such as the paired T-test, data have to be normally distributed. Since the *t* distribution tends to a normal distribution for large sample size, the normality condition could be neglected if the sample size is at least 30 (Sokal & Rohlf, 1987 p. 107). The sample sizes in the present study, and thus the differences (the data), fulfils the criteria, therefore the T-test can be used. At the same time, large sample size always tends to give significant differences, even though they are very small in practice. In order to decide if the significant differences are to be accounted for, Cohen's *d* is used as an index of the effect size. This number is defined as

$$d = \frac{\bar{x}_D}{s_D}, \quad (1)$$

where $\bar{x}_D$ is the difference of the group means and s_D is the standard deviation of the difference. The effect size is classified as small if $d = 0.2$, as medium if d 0.5, and as large if $d \geq 0.8$ (Sullivan & Feinn, 2012).

Accounting for grades and gender, descriptively (RQ 4, 5 and 6)

To be able to say more about if/how figurative context influences students' success, it is desirable to account for students' ability. As a measure of the ability, students' course grades were used. An alternative would have been to use their test scores, which would have provided a more fine grained scale for the ability. On the other hand, tests score is dependent on the observed task itself, which could lead to circle dependencies. In any case, the course grade is enough for this study, and since the course grade is based on both the national test result and other performances made during the course, possible circle dependencies can be reduced.

There are four grades possible to obtain in a course, *Not Pass* (IG), *Pass* (G), *Pass with distinction* (VG) and *Pass with special distinction* (MVG). After grouping students with respect to their grades, each grade-group's total score on the Mathematics B 2004 test was summed and compared to the scores respective group had received on each of the categories intraMath tasks and context tasks. After this, the different grade-groups' total scores on CR tasks and IR task were summed and compared to respective group's scores on the CR-M and CR-C tasks, and on the IR-M and IR-C tasks respectively.

For example, sum the scores on all CR-M tasks that students with the grade MVG have received. Divide this sum by the sum of the total scores the students with the grade MVG has received on all the CR tasks. Make the same calculations for the scores students with the grade MVG has received on all the CR-C tasks in the test. Then repeat this for the remaining grade-groups. By graphing the obtained proportions, a descriptive comparison of the influence of figurative context is obtained.

As mentioned in the introduction, there exist some differences in boys' and girls' success on various mathematics tests. Therefore it is desirable to take gender into account when studying the influence of figurative context. To do this, students were grouped by gender and kept sub-grouped by grades. For every subgroup, the logarithmic differences between the odds for students' success on intraMath tasks and on context tasks on the Mathematics B 2004 test were calculated. Letting f_M denote an individual student's proportion of intraMath scores and $1 - f_M$ denote the proportion of the intraMath scores not received, the individual student's odds for intraMath-tasks is $f_M/(1 - f_M)$. The odds for context tasks is calculated in the same way, which gives the logarithmic differences as

$$\log\left(\frac{f_M}{1-f_M}\right) - \log\left(\frac{f_C}{1-f_C}\right). \quad (2)$$

The logarithmic differences were then graphed for boys and girls espectively and a local regression line (LOESS) was fitted to each graph.

Differential Item Functioning (RQ 7)

To significantly test if the presence of figurative contexts might influence boys and girls differently despite that they have the same ability, the tasks are tested for differential item functioning (DIF). DIF exists if people with the same knowledge/ability, but belonging to different groups, have different probabilities to give the right answer to an item/task. Group belongings could be with regard to, for example, gender (as in present study), ethnicity, culture or language. A widely used method for detecting DIF is the Mantel-Haenszel procedure (MH) (Guilera, Gómez-Benito & Hidalgo, 2009). MH was originally developed for data analyses from retrospective studies in the clinical epidemiology area. The purpose of MH was to test if there were any relations between the occurrence of a disease and some factors. The disease could for instance be lung cancer and one factor could be cigarette smoking (Mantel & Haenszel, 1959). Holland and Thayer (1988) were the first ones to use MH to detect DIF. The original method is based on 2x2 contingency tables (Table 1) were the rows indicate group belongings, reference (R) or focal (F) group, and the columns indicates success or not on the dichotomous task that is analysed for DIF. There is one table for every measure *i* of the ability, which in the present study is students' course grades.

Table 1. Contingency table for repeat *i*. a_i, b_i, c_i and d_i represent the frequencies for right and wrong for the groups R and F. $n_{Ri} = a_i + b_i$, is the number of students in the reference group, and $n_{Fi} = c_i + d_i$, is the number of students in the focal group, $n_i = n_{Ri} + n_{Fi}$.

Group	Score on the task		Total
	1	0	
R	a_i	b_i	n_{Ri}
F	c_i	d_i	n_{Fi}
Total	m_{1i}	m_{0i}	n_i

The method consists in estimating the *common odds ratio*, α_{MH}, for the different contingency tables. The number α_{MH} is the so-called MH index of the DIF and it is estimated as the sum of the weighted odds ratios for the individual contingency tables $i = 1,...,k$, where k is the number of contingency tables. That is,

$$\alpha_{MH} = \frac{\sum_i a_i d_i / n_i}{\sum_i b_i c_i / n_i} = \frac{\sum_i w_i \alpha_i}{\sum_i w_i}, \tag{3}$$

Where

$$\alpha_i = \frac{a_i/b_i}{c_i/d_i} = \frac{a_i d_i}{b_i c_i} \tag{4}$$

is the odds ratio for table *i* and

$$w_i = \frac{b_i c_i}{n_i} \tag{5}$$

is the weight associated to α_i. The index α_{MH} is a measure of the effect size; if $\alpha_{MH} = 1$ there is no difference between the groups' success on the task, if $\alpha_{MH} >$

1 the task is in favour of the reference group, and if $\alpha_{MH} < 1$ the task is in the focal group's favour.

To decide whether α_{MH} differs significantly from 1, a MH test statistic, χ^2_{MH}, is calculated. This test statistic is approximately chi-square distributed, and is compared to a chi-square distribution with one degree of freedom (Mantel & Haenszel, 1959; Ramstedt, 1996). The definition of χ^2_{MH} is

$$\chi^2{}_{MH} = \frac{(|\sum_i a_i - \sum_i E(a_i)| - ½)^2}{\sum_i Var(a_i)}, \tag{6}$$

where $E(a_i) = {}^{n_{Ri}m_{1i}}/_{n_i}$ is the expected value for a_i under H_0 and

$$Var(a_i) = \frac{n_{Ri}n_{Fi}m_{1i}m_{0i}}{n_i^2(n_i - 1)} \tag{7}$$

is the variance for a_i (Mantel & Haenszel, 1959).

Since the χ^2_{MH} test statistic is dependent on sample size, large sample size tends to always give significant differences, even though they are very small in practice, and small sample size can result in large differences though the result is not significant (also discussed in the Paired T-test section). To decide whether the DIF is practically significant, Ramstedt (1996) refers to the ETS (Educational Testing Service) DIF classification of α_{MH}. Depending on different values of α_{MH}, the effect of DIF is divided into three groups, A: negligible, B: slight to moderate and C: moderate to large. Since the sample sizes in this study can be considered large, only significant results will be considered. Thus DIF classified as B occurs for $0.53 < \alpha_{MH} < 0.65$ or $1.54 < \alpha_{MH} < 1.89$, and DIF classified as C occurs for $\alpha_{MH} < 0.53$ or $\alpha_{MH} > 1.89$.

To be able to use MH for detecting DIF on polytomous tasks, Ramstedt (1996) introduced a modified version of MH in his study about differences in boys' and girls' success on national physics tests. Instead of letting the frequencies in the contingency tables represent number of boys and girls, they represent number of "boy-scores" and "girl-scores" for the different cells. The analogues of the frequencies in Table 1 are calculated according to $a_i = p_{Ri} \cdot n_{Ri}$ and $b_i = (1-p_{Ri}) \cdot n_{Ri}$, where p_{Ri} is the proportion solved tasks (scores) and $1-p_{Ri}$ is the proportion non-solved tasks (non-scores) for group R. The generalised solution proportion p_{Ri} is defined as

$$p_{Ri} = \frac{\sum_j n_{Ri} \cdot i}{M \cdot n_{Ri}}, \tag{8}$$

where the summation runs over $j = 1, \ldots, M$, and M is the maximum score on the task. The frequencies c_i and d_i for group F is computed in the same way. With this generalised meaning of a_i, b_i, c_i and d_i, one can calculated the equalities (3) to (7) and MH can be used for detecting DIF on tests with a mixture of dichotomous and polytomous tasks (Ramstedt, 1996).

Results

The categorisation of tasks according to if they are intraMath tasks or context tasks, suggests that the number of context tasks decreases the more advanced the mathematics courses become. In average, 52 % of the tasks in a Mathematics B test are context tasks (54 % and 50 %), compared to 15 % in a Mathematics D test (16 % and 14 %) (Table 2, column 2 and 3).

Comparing students' solution rate (RQ 1 and 2)

Comparing the solution rate on intraMath tasks and on context tasks did not reveal any trend on which type of task students succeeded better on (Table 2, column 4 and 5). For both of the Mathematics D tests, the solution rate are higher for context tasks than for intraMath task, 71 % and 62 % vs. 59 % and 55 %, respectively. This is not the case for the Mathematics B and the Mathematics C tests; instead the solution rate for intraMath tasks is higher than for context tasks in the Mathematics B test from 2003 and in the Mathematics C test from 2004, and lower in the other two tests. It is further noticed in Table 2 that the solution rates for intraMath tasks seem to be lower on the 2004 tests compared to the tests from 2003 for the same math courses. It is reasonable to assume that students' knowledge and skills have not changed over a year. Instead, the differences are an indication of different level of difficulty in the different tests. Calculations for the solution rates for the subcategories; CR-M tasks, CR-C tasks, IR-M tasks and IR-C tasks, are outlined in Table 2, columns 7, 9, 11 and 13, respectively.

The results indicates that the presence of a figurative context can influence students' success when respect is taken to required kind of mathematical reasoning. For most of the tests, the solution rate is higher for CR-C tasks than for CR-M tasks, cf. columns 9 and 7 in Table 2. This suggests that students do better on a task requiring CR if the task includes a figurative context. If a task is solvable by IR, it seems that the influence is the opposite, i.e. the solution rates are higher for IR-M tasks than for IR-C tasks in most of the tests (Table 2, columns 11 and 13).

Table 2. Number of students, number of tasks and the mean solution rates for the categories context tasks and intraMath tasks, as well as the number of tasks in each sub-category with the respective mean solution rate. Each row in the table displays the values for one of the Mathematics tests.

Test	Stud-ents n	context-tasks n (%)	Intra Math-tasks n (%)	mean sol.rate context-tasks %	mean sol.rate intra Math-tasks %	CR-M-tasks n (%)	mean sol.rate CR-M-tasks %	CR-C-tasks n (%)	mean sol.rate CR-C-tasks %	IR-M-tasks n (%)	mean sol.rate IR-M-tasks %	IR-C-tasks n (%)	mean sol.rate IR-C-tasks %
Math B 2003	1600	13 (54)	11 (46)	48	62	5 (21)	49	7 (29)	50	6 (25)	75	6 (25)	46
Math B 2004	3481	11 (50)	11 (50)	59	50	6 (27)	36	9 (41)	59	5 (23)	68	2 (9)	56
Math C 2003	1504	9 (37)	15 (63)	75	63	9 (37)	55	4 (17)	65	6 (25)	80	5 (21)	85
Math C 2004	1014	7 (27)	19 (73)	51	55	9 (35)	52	3 (12)	37	10 (38)	57	4 (15)	62
Math D 2003	829	3 (16)	16 (84)	71	59	12 (63)	52	2 (11)	70	4 (21)	88	1 (5)	75
Math D 2004	861	3 (14)	19 (86)	62	55	14 (64)	50	3 (13)	62	5 (23)	71	0 (0)	No tasks of this kind

Significance testing of the influence of figurative context (RQ 3)

The significance testing was made to test the hypothesis that had arose from the previous results. The hypothesis was that the presence of figurative context has an effect on students' success if mathematical reasoning is taken into account such that: if it is a CR task, figurative context has a positive effect, and if it is an IR task, students succeed better if the task does not have a figurative context. The hypothesis was first tested in the Mathematics B 2003 test. The result showed significant differences both between CR-M and CR-C tasks and between IR-M and IR-C tasks (first row, Table 3). The computed mean of students' solving rate on CR-M tasks was lower than the mean for CR-C tasks ($\bar{x}_D$ = - 0.11) and since the effect size can be considered slightly larger than medium (*d* = 0.6), it is most likely that the means are different. The means of the solving rates for IR-M and IR-C tasks are most likely very different ($\bar{x}_D$ = 0.23, *d* = 0.8).

Table 3. Results from the paired sample T-tests. The first column shows the direction of the difference, as well as the estimators of the means of the solving rates for CR-M and CR-C tasks. The second column shows the estimated mean difference between CR-M and CR-C tasks with corresponding confidence interval, as well as the p-value. The fourth and the fifth column shows the same entities, but with respect to IR-M and IR-C tasks. The third and the sixth column display Cohen's *d* for the respective estimated mean differences.

Test	Directions of the difference in success on CR-tasks (means (%) for CR-M and CR-C)	Mean difference, (confidence interval), and p-value, with respect to CR tasks	Cohen's *d* for CR pairs	Directions of the difference in success on IR-tasks (means (%) for IR-M and IR-C)	Mean difference, (confidence interval), and p-value, with respect to IR tasks	Cohen's *d* for IR pairs
Math B 2003	CR-M < CR-C (0.22; 0.33)	-0,11 (-0.12 : -0.10) p < 0,005	0.6	IR-M > IR-M (0.56; 0.34)	0.23 (0.21 : 0.24) p < 0.005	0.8
Math B 2004	CR-M < CR-C (0.23; 0.47)	-0.24 (-0.24 : -0.23) p < 0.005	1.2	IR-M > IR-M (0.57; 0.52)	0.05 (0.04 : 0.06) p < 0.005	0.14
Math C 2003	CR-M < CR-C (0.40; 0.47)	-0.07 (-0.08 : -0.06) p < 0.005	0.3	No significant difference	--- --- p = 0.7	---
Math C 2004	**CR-M > CR-C (0.36; 0.27)**	0.09 (0.08 : 0.1) p < 0.005	0.4	IR-M > IR-M (0.53; 0.47)	0.07 (0.05 : 0.08) p < 0.005	0.3
Math D 2003	CR-M < CR-C (0.41; 0.50)	-0.08 (-0.10 : -0.07) p < 0.005	0.3	IR-M > IR-M (0.74; 0.61)	0.13 (0.10 : 0.15) p < 0.005	0.4
Math D 2004	CR-M < CR-C (0.39; 0.59)	-0.19 (-0.21 : -0.18) p < 0.005	0.7	No such pair	---	---

To see if these results for Mathematics B 2003 hold in general, the analysis was replicated for the other five tests and the results are displayed in Table 3. Results for the Mathematics B 2004 test verified that the mean of solving rates for CR-M tasks was lower than for CR-C tasks, and the difference was in this case large ($\bar{x}_D$ = -0.24, *d* = 1.2). The difference of the means of the solving rates for IR-M and IR-C tasks could on the contrary be considered small ($\bar{x}_D$ = 0.05, *d* = 0.14). The results from the paired T-test on the Mathematics C 2003 test gave only significant difference for the CR-pair, and this difference is most likely quite small ($\bar{x}_D$ = -0.07, *d* = 0.3). For the Mathematics C 2004 test, the opposite difference for CR tasks was found significant, and large enough to account for ($\bar{x}_D$ = 0.09, *d* = 0.4). Both the Mathematics C 2004 test and the Mathematics D 2003 test gave significant difference between the solving rates for the IR-pairs. Since there were no IR-C tasks in the Mathematics D 2004 test, no difference could be tested for IR-pairs. The difference between solving rate means for the CR tasks could be considered quite large though ($\bar{x}_D$ = -0.19, *d* = 0.7) (Table 3).

Descriptive analyses of figurative context's influence on success, accounting for students' grades (RQ 4 and 5)

Figure 3 shows the proportion of scores for the different categories intraMath tasks and context tasks each grade group has received compared to respective groups total score on the Mathematics B 2004 test.

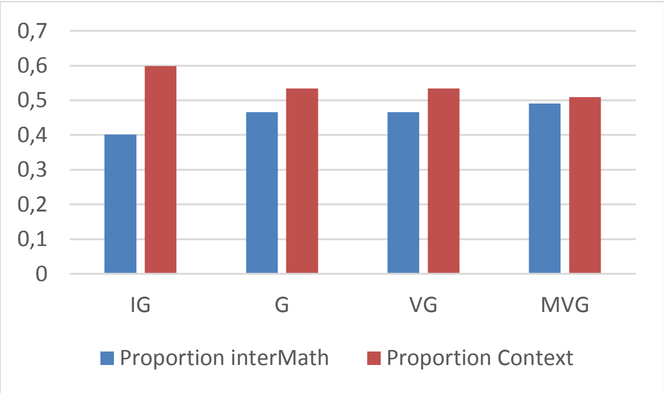


Figure 3. The proportion intraMath scores (left bars) and proportion context scores (right bars) of respective grade-groups total scores on the Mathematics B 2004 test.

It is noticed that for students with the highest grade there is almost no difference; approximately half of the total score of the MVG-group comes from intraMath tasks and the other half from context tasks. For students with the grades G and VG, the proportion context scores are a bit higher than the proportion intraMath scores. For the students with the lowest grade, the difference has increased. These students solve proportionally more context tasks than intraMath tasks.

Subdividing tasks according to mathematical reasoning requirements reveals that there are almost no differences between the students in the different grade-groups regarding scores on IR-M tasks and IR-C tasks (Figure 4).

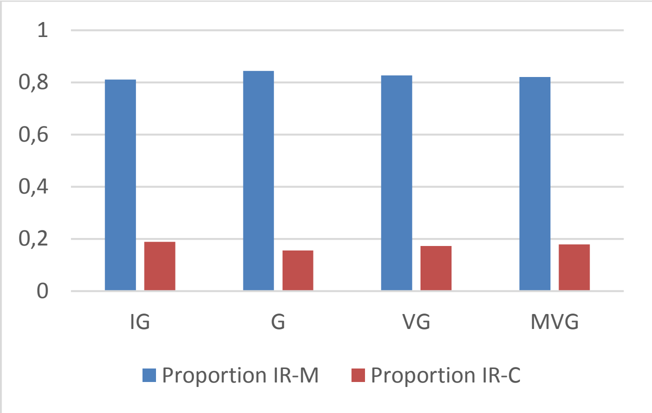


Figure 4. The proportion IR-M scores (left bars) and proportion IR-C scores (right bars) of respective grade-groups total IR scores on the Mathematics B 2004 test.

On the other hand, comparing the proportion of scores for CR-M and for CR-C tasks indicates a difference in students' success depending on their grade. It seems that students with the lower grades get a higher proportion of the scores for CR tasks among the tasks that include a figurative context (Figure 5).

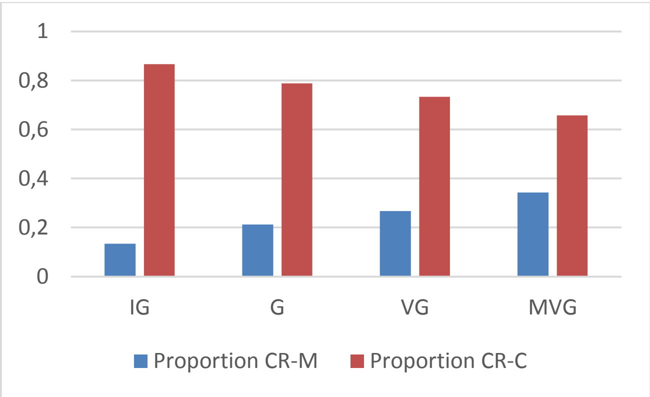


Figure 5. The proportion CR-M scores (left bars) and proportion CR-C scores (right bars) of respective grade-groups total CR scores on the Mathematics B 2004 test.

Descriptive analyses of figurative context's influence on success, accounting for students' grades and gender (RQ 6)

The logarithmic differences of boys' and girls' odds on intraMath tasks and context tasks are illustrated in Figure 6. The LOESS line indicates that the differences for boys and girls are approximately the same for students with the higher grades. For the lower grades though, it seems that the difference is larger for boys than for girls, which should indicate that boys with lower grades do better on tasks embedded in a figurative context.

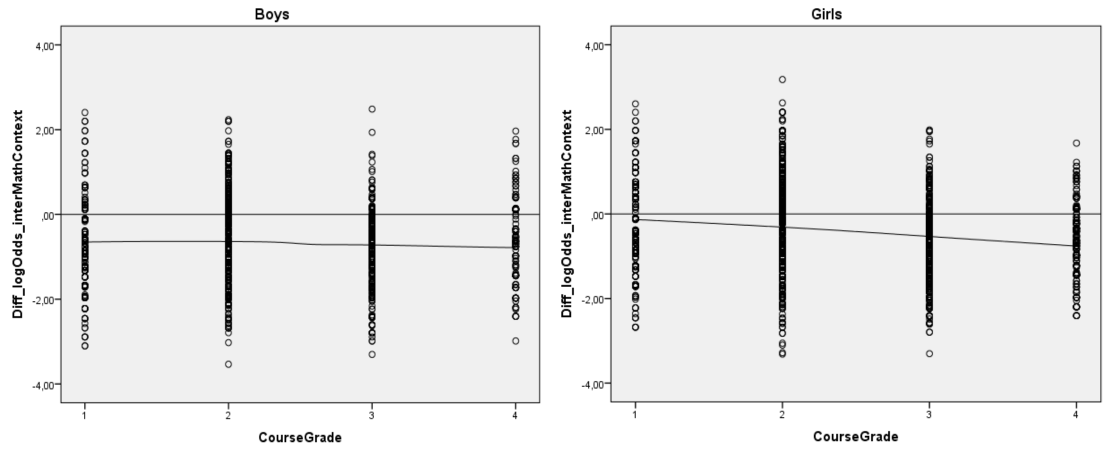


Figure 6. The logarithmic differences (y-axis) between boys' and girls' odds on intraMath tasks and context tasks, with respect to their grades (x-axis), 1 = IG, 2 = G, 3 = VG, 4 = MVG. Both graphs contain a reference line at y = 0 and a LOESS line.

Testing tasks for DIF that could be explained by the presence of figurative context (RQ 7)

The index for effect, α_{MH}, for each task in the Mathematics B 2004 test is given in Table 4, together with the values for the respective test statistic, χ^2_{MH}. The critical value for the χ^2 test with respect to 5% significance level is 3.84, i.e. if the calculated value is greater than 3.84 H_0 can be rejected. According to ETS DIF classification accounted for in the method section, α_{MH} should be lower than 0.65 or larger than 1.54 if the DIF is to be accounted for. Of the 22 tasks in the test, four showed DIF with respect to gender; these tasks are highlighted in Table 4. As noticed, all the significant DIF indices are greater than 1, i.e. boys do better than girls on the tasks, even though they have the same mathematical abilities. Three of the four tasks are CR tasks, the other one is an IR task. Two tasks include a figurative context and two are intra-mathematical.

Table 4. Values for the Mantel-Haenszel DIF index, α_{MH}, and for the test statistic, χ^2_{MH}, as well as the p-value, for the tasks in the Mathematics B 2004 test.

	Task 1 IR-M	Task 2 IR-M	Task 3 IR-M	Task 4 IR-M	Task 5a IR-C	Task 5b CR-C	Task 6 IR-M	Task 7 CR-M	Task 8a CR-C	Task 8b CR-C	**Task 9 CR-C**
α_{MH}	0,843	0,867	0,842	0,926	1,045	1,354	1,028	1,364	1,454	1,444	**1,639**
χ^2_{MH}	4,122	3,080	4,692	0,766	0,302	15,006	0,099	11,232	22,206	19,578	**34,673**
p	0,042	0,079	0,030	0,382	0,583	0,0001	0,753	0,0008	$2,4\cdot10^{-06}$	0,0000	**$3,9\cdot10^{-06}$**
	Task 10a CR-C	Task 10b CR-C	Task 11a CR-M	**Task 11b CR-M**	**Task 12a IR-C**	Task 12b CR-C	Task 13 CR-C	Task 14 CR-C	**Task 15a CR-M**	Task 15b CR-M	Task 16 CR-M
α_{MH}	1,426	1,529	1,198	**1,876**	**1,736**	1,205	0,973	1,310	**1,973**	1,486	0,997
χ^2_{MH}	18,544	29,622	4,545	**25,414**	**42,114**	3,619	0,084	9,612	**28,913**	7,476	0,000
p	0,00002	$5,3\cdot10^{-08}$	0,033	**$4,6\cdot10^{-07}$**	**$8,6\cdot10^{-11}$**	0,057	0,772	0,002	**$7,6\cdot10^{-08}$**	0,006	0,989

The results from calculating DIF on this Mathematics B test do not show that figurative context affects boys' and girls' success differently. To see if similar results are obtained, DIF is calculated for the tasks on the other five mathematics tests, the result is displayed in Table 5. These calculations resulted in DIF for two tasks out of 24 in the Mathematics B 2003 test, one CR-C task in favour of the boys and one IR-M task in favour of the girls. In the Mathematics C 2003 test there were two tasks indicating DIF, one CR-M task and one CR-C task, and in the Mathematics C 2004 test there were three tasks with DIF, two CR-M tasks and one IR-M task, all DIF tasks in the Mathematics C tests were in favour of the boys. In the Mathematics D 2003 test there was one CR-M task with significant DIF, and in the Mathematics D 2004 test there were two tasks with DIF, one CR-M and one CR-C task, all three tasks in favour of the boys.

Table 5. Values for the significant DIF indices, together with the values for the tests statistics and the p-value for the remaining mathematics tests.

	α_{MH}, χ^2_{MH}, p (type of task)	$\propto_{MH}$, χ^2_{MH}, p (type of task)	α_{MH}, χ^2_{MH}, p (type of task)
Math B 2003	1.79, 14.96, 0.0001 (CR-C)	0.63, 10.95, 0.0009 (IR-M)	
Math C 2003	1.60, 11.98, 0.0005 (CR-M)	1.63, 13.55, 0.0002 (CR-C)	
Math C 2004	1.62, 7.49, 0.0062 (CR-M)	1.58, 7.99, 0.0047 (CR-M)	1.62, 10.16, 0.0014 (IR-M)
Math D 2003	1.71, 9.50, 0.0021 (CR-M)		
Math D 2004	1.89, 10.04, 0.0015 (CR-M)	1.96, 16.18, 0.00006 (CR-C)	

Discussion

As the results from comparing solution rates show, it is first after taking account for mathematical reasoning requirements that the presence of figurative context seems to have an effect on students' success. The results indicate that it is easier for students to solve CR tasks if they are embedded in a figurative context. The solution to a CR task requires that the solver uses his/her mathematical knowledge in a novel situation, which in turn requires an understanding of the assignment. It is therefore reasonable to believe that if the assignment is embedded in a real-life context students can relate to, this helps them to come up with possible ways to solve the problem.

An interesting outcome from the analyses is that the number of context tasks decreases the more advanced the mathematics course is. As seen in Table 2, from around 50 % context tasks in mathematics B, the proportion decreases to around 15 % in Mathematics D. At the same time, most of the tasks in the Mathematics D tests are CR tasks, on which according to the previously discussed results, students likely do better if such tasks are embedded in a figurative context. A reason could be that it is easier to find real-life situations students can relate to and that not require any knowledge of science for more basic mathematics courses. One could also assume that the motivating argument for real-life context in tasks (discussed in the Introduction) is considered more important for students in the less advanced courses, since those mathematics courses are compulsory to most upper secondary students whereas the more advanced courses are intended for students in programmes with specialisation in mathematics and science.

Since the level of difficulty varies for the different tests, no general conclusion can be drawn. Differences between results on tests from different years are likely due to internal property of the tests, rather than differences between the students' abilities. Instead, in order to study further how context might influence students' success it is desirable to construct intraMath and context tasks with the same level of difficulty and let two similar classes solve these tasks.

The indication that the presence of context in tasks affects students differently, with respect taken to their ability, is noteworthy for teachers' practice. Students with lower ability seem to be in greater need of relating the mathematics to a familiar reality. Thus, in order to give all students the same possibilities to learn mathematics, it is desirable to use relevant contexts from those students' everyday life when mathematical concepts are introduced. This corresponds with one of the arguments for bringing context to the mathematics education, discussed in Introduction.

It is interesting to note from the results accounting for gender that the presence of context in CR tasks seems to affect boys with lower grades more than girls with lower grades. At the same time, from the tasks flagged for DIF, the presence of context could not alone explain the differences in boys' and girls' success. The tasks that did show significant DIF need to be analysed further from various perspectives in order to find a reason for DIF. What can be noticed in Table 5 is that most of the tasks flagged for DIF are in the boys' favour and these tasks are, in all cases but one, CR tasks. This is interesting to compare with the results from Sumpter's (2015) study, which analysed teachers' conception about whether mathematical reasoning is gendered. A conclusion from that study is that CR is thought of as neutral. It is also discussed that girls are thought of as process focused and that boys are more chance taking. Sumpter further discusses how boys in previous studies have been observed to have a more guessing strategy than girls have. Since solutions to CR tasks require novelty, some guessing and chance taking could be considered necessary in the solution procedure.

Finally, one can indeed conclude that the advantage of everyday context in mathematics tasks is very complex and that no general conclusions could be made from the various analyses in the present paper. It seems to be a positive relation between requirements of creative mathematical reasoning and figurative context, as well as between the success of students with lower grades and the presence of figurative context. Gender differences in the success could not be explained only by the presence of context, but other analyses are required. Since the presence of context in a task in many cases involves more text to read and to interpret, it would be desirable to study how context affects students' success depending on their literacy.

References

Bassok, M. & Holyoak, K. (1989). Interdomain Transfer Between Isomorphic Topics in Algebra and Physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 15(1), 153-166. doi:10.1037/0278-7393.15.1.153

Basson, I. (2002). Physics and mathematics as interrelated fields of thought development using acceleration as an example. *International Journal of Mathematical Education in Science and Technology*, 33(5), 679-690. doi:10.1080/00207390210146023

Bergqvist, E. & Lind, A. (2005). *Är det svårare att dela med fyra än med två när man läser matte C?: En jämförelse av svårighetsgrad mellan olika versioner av matematikuppgifter i Nationella kursprov* [Is it more difficult to divide by four than by two when taking the course Mathematics C? A comparision of different versions of mathematics tasks in National tests]. Umeå, Sweden: Department of Educational Measurement, Umeå universitet.

Bing, T. (2008). *An epistemic framing analysis of upper level physics students' use of mathematics.* (Doctoral dissertation). Retrieved from http://bit.ly/Bing2008

Boaler, J. (1994). When Do Girls Prefer Football to Fashion? An analysis of female underachievement in relation to "realistic" mathematic contexts. *British Education Research Journal*, 20(5), 551-564. doi:10.1080/0141192940200504

Cooper, B. & Dunne, M. (2000). *Assessing children's mathematical knowledge: social class, sex and problem-solving*. Buckingham, England: Open University Press.

Doorman, L. M. & Gravemeijer K. P. E. (2009). Emergent modeling: discrete graphs to support the understanding of change and velocity. *ZDM – The International Journal on Mathematics Education*, 41, 199-211. doi:10.1007/s11858-008-0130-z

Guilera, G., Gómez-Benito, J. & Hidalgo, M. D. (2009). Scientific production on the Mantel-Haenszel procedure as a way of detecting DIF. *Psicothema* 21(3), 493-498.

Hanna, G., de Bruyn, Y., Sidoli, N. & Lomas, D. (2004). Teaching Proof in the Context of Physics. *ZDM – Zentralblatt für Didaktik der Mathematik*, 36(3), 82-90. doi:10.1007/BF02652775

Hirsch, R. (1996). Is Mathematics a Pure Science? *Science & Society*, 60(1), 58-79. Stable URL: http://www.jstor.org/stable/40403533

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel Haenszel Procedure. In H. Wainer & H. I. Braun (Eds), *Test validity* (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Howson, G. (2005). "Meaning" and School Mathematics. In J. Kilpatrick, C. Hoyles, O. Skovsmose & P. Valero (Eds.) *Meaning in Mathematics Education* (pp.17-38). New York: Springer.

Kulm, G. (1984). The classification of problemsolving research variables. In G. A. Goldin & C. E. McClintock (Eds.), *Task variables in mathematical problem solving* (pp. 1-21). Philadelphia: The Franklin Institute Press.

Lesh, R., & Zawojewski, J. (2007). Problem Solving and modeling. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp.763-804). Charlotte, NC: Information Age Publishing.

Lithner, J. (2008). A research framework for creative and imitative reasoning. *Educational Studies in Mathematics*, 67(3), 255-276. doi:10.1007/s10649-007-9104-2

Lobato, J., Rhodehamel, B. & Hohense, C. (2012). “Noticing” as an Alternative Transfer of Learning Process. *Journal ot the Learning Sciences*, 21(3), 433-482. doi:10.1080/10508406.2012.682189

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719-748. doi:10.1093/jnci/22.4.719

Marongelle, K. A. (2004).How Students Use physics to Reason About Calculus Tasks. *School Science and Mathematics*, 104(6), 258-272. doi:10.1111/j.1949-8594.2004.tb17997.x

Ministry of Education and Research (2007). *Tydliga mål och kunskapskrav i grundskolan – Förslag till nytt mål- och uppföljningssystem* [Explicit goals and knowledge requirements in the compulsory school – Proposal for a new system for goals and follow-up]. Ministry publications series SOU 2007:28. Stockholm, Sweden: Fritzes.

Nunes, T., Schliemann, A. & Carraher, D. (1993). *Street mathematics and school mathematics.* Cambridge, United Kingdom: Cambridge University Press.

OECD (2014). *Are boys and girls equally prepared for life?* Retrieved 2015-02-17 from http://bit.ly/oecd2012_gender

Palm, T. (2002). *The Realism of Mathematical School Tasks – Features and Consequences*. (Doctoral dissertation). Umeå, Sweden: Umeå University.

Palm, T., Boesen, J., & Lithner, J. (2011). Mathematical Reasoning Requirements in Upper Secondary Level Assessments. *Mathematical Thinking and Learning, 13*(3), 221-246. doi:10.1080/10986065.2011.564994

Ramstedt, K. (1996). *Elektriska flickor och mekaniska pojkar. Om gruppskillnader på prov – en metodutveckling och en studie av skillnader mellan flickor och pojkar på centrala prov i fysik* [Electrical girls and mechanical boys. About group differences on tests – a method development and a study about differences between boys and girls on the national tests in physics]. (Doctoral dissertation). Umeå, Sweden: UmU Tryckeri.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for Research on Mathematics Teaching and Learning* (pp.334-370). New York, NY: MacMillan

Shimizu, Y., Kaur, B., Huang, R. & Clarke, D. (2010). The Role of Mathematical Tasks in Different Cultures. In Y. Shimizu, B. Kaur, R. Huang and D. Clarke (Eds.), *Mathematical Tasks in Classrooms Around the World* (pp.1-14). Rotterdam, Netherlands: Sense Publisher.

Sokal, R. R. & Rohlf, F. J. (1987). *Introduction to biostatics. 2nd ed.* New York, NY: Freeman.

Sullivan, G. M. & Feinn, R. (2012). Using Effect Size–or Why the *P* Value Is Not Enough. *Journal of Graduate Medical Education*, 4 (3), 279-282. doi:10.4300/JGME-D-12-00156.1

Sumpter, L. (2012). Upper secondary school students´ gendered conceptions about affect in mathematics. *Nordic Studies in Mathematics Education*, 17 (2), 27–47.

Sumpter, L. (2015). Investigating Upper Secondary School Teachers' Conceptions: Is Mathematical Reasoning Considered Gendered? *International Journal of Science and Mathematics Education*, Published online 04 Mar 2105. doi:10.1007/s10763-015-9634-5

Swedish National Agency for Education. (2006). Curriculum for the non-compulsory school system, Lpf 94. Stockholm, Sweden: Fritzes.

Swedish National Agency for Education (2009). *TIMSS Advanced 2008. Svenska gymnasieelevers kunskaper i avancerad matematik och fysik i ett internationellt perspektiv* [TIMSS Advanced 2008. Swedish upper secondary students' knowledge in advanced mathematics and physics from an international perspective]. Stockholm, Sweden: Fritzes.

Tall, D. & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, 12(2), 151-169. doi:10.1007/BF00305619

Verschaffel, L., De Corte, E. & Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction*, 4(4), 273-294.

Wyndhamn, J. (1993). *Problem-solving revisited, On school mathematics as a situated practice*, (Doctoral dissertation). Linköping, Sweden: Linköping University.