Linear Functions – a Matter of Context? Toward Understanding Students' Learning Processes in Mathematics and Physics

Lynn Knippertz

Vom Fachbereich Mathematik
der Rheinland-Pfälzischen Technischen Universität
Kaiserslautern-Landau
zur Verleihung des akademischen Grades
Doktor der Naturwissenschaften
(Doctor philosophiae naturalis, Dr. phil. nat.)
genehmigte Dissertation

DE-386

Gutachter:
Prof. Dr. Stefan Ruzika
Prof. Dr. Pascal Klein

Tag der Disputation: 18. Juli 2025

Rheinland-Pfälzische
Technische Universität
Kaiserslautern
Landau

Abstract

Learning processes of linear functions in mathematics and physics are widely unclear. In this work, we replicate and extend existing results on learning processes of linear functions in mathematics and physics and use triangulation of gaze data and interview data for a deeper understanding of cognitive processes. Furthermore, we utilize mathematical structures such as networks to algorithmically detect solution strategies from gaze data and to distinguish between correct and incorrect solvers.

Although there are validated tests of linear functions in mathematics and physics, they are not sufficiently validated and remain unused across countries and age groups. Furthermore, these tests are paper-based, making it impossible to investigate visual attention processes. There is also a lack of research on analyzing students' visual data when solving math and physics problems using algorithmic, network-based approaches. In this thesis, we improve tools in learning systems that can guide learning without the help of teachers.

A three-step approach was taken to fill these research gaps:

1. We conducted a large-scale study using the paper-based test instrument on linear functions in mathematics and physics by Ceuppens et al. (2019), which was validated for 9th grade students in Belgium. We evaluated this test instrument for German students in the same grade with a sample size of N = 249. In addition, we used the test instrument in the upper school (grade 11) to check whether the challenges identified by Ceuppens et al. (2019) persist across school years (N = 298). Results show that German 9th graders perform significantly worse than Belgian 9th graders. However, the challenges surprisingly identified by Ceuppens et al. (2019) remain, and are even more serious for German students.

2. To explicitly understand the cognitive processes behind the students' problems, we conducted an eye-tracking study in schools (N = 131), in which we were able to cognitively resolve transfer difficulties in adequately applying mathematical content in a physical context using gaze- and interview data for two forms of representation: formulas and graphs. Here, we showed that cognitive processes are context sensitive.

3. In a final step, we use network analysis methods to detect solution strategies based on geometric patterns (for example slope triangles) from gaze data. It was investigated which network metrics are suitable for distinguishing

correct from incorrect solvers. There exist suitable network metrics—a first step toward the development of a predictor to distinguish between correct and incorrect solvers.

In summary, based on three extensive studies with samples ranging from 131 (eye-tracking studies) to about 250 people (paper-pencil tests), this thesis uses various different methods (paper and pencil, eye-tracking, interviews, network analysis) to empirically analyze learning processes for linear functions regarding mathematics and physics. The results show that students face challenges in interpreting mathematical content adequately in a physical context. The analysis of the cognitive processes form a basis for the sustainable improvement of learning in the future. A teacher training course on this topic has already put this into practice.

Zusammenfassung

Lernprozesse zu linearen Funktionen in Mathematik und Physik sind wissenschaftlich weitgehend unverstanden. In dieser Arbeit reproduzieren und erweitern wir bestehende Ergebnisse zu Lernprozessen linearer Funktionen in Mathematik und Physik. Darüber hinaus triangulieren wir Blickdaten und Interviewdaten, um kognitive Prozesse besser zu verstehen. Wir nutzen mathematische Netzwerke, um Lösungsstrategien aus Blickdaten algorithmisch zu erkennen und zwischen richtigen und falschen Lösern zu unterscheiden.

Es gibt zwar validierte Tests zur Untersuchung von linearen Funktionen in Mathematik und Physik, diese sind jedoch nicht ausreichend validiert und wurden bisher nicht länder- und altersübergreifend eingesetzt. Weiter sind diese Tests papierbasiert, was die Untersuchung visueller Aufmerksamkeitsprozesse unmöglich macht. Es fehlen auch Studien, die die visuellen Daten von Schüler:innen beim Lösen von Mathematik- und Physikaufgaben mit algorithmischen, netzwerkbasierten Ansätzen analysieren. Wir verbessern Tools in Lernsystemen, die in der Lage sind, das Lernen ohne Hilfe von Lehrkräften zu steuern.

Um diese Forschungslücken zu füllen, wurde ein dreistufiger Ansatz gewählt:

1. Wir führten eine große Studie durch, in der wir das papierbasierte Testinstrument zu linearen Funktionen in Mathematik und Physik von Ceuppens et al. (2019) verwendeten, das für Schüler:innen der neunten Klasse in Belgien validiert wurde. Wir haben dieses Testinstrument für deutsche Schüler:innen der gleichen Jahrgangsstufe mit einer Stichprobengröße von $N = 249$ evaluiert. Darüber hinaus wurde das Testinstrument in der Oberstufe eingesetzt, um zu überprüfen, ob die von Ceuppens et al. (2019) festgestellten Schwierigkeiten über die Schuljahre hinweg bestehen bleiben ($N = 298$). Ergebnisse zeigen eine signifikante bessere Leistung für belgische Neuntklässler im Vergleich zu deutschen Neuntklässlern. Überraschenderweise bleiben die Schwierigkeiten der Lernenden jahrgangsübergreifend bestehen und sind für deutsche Schüler:innen sogar noch gravierender.

2. Um die kognitiven Prozesse hinter den Problemen der Schüler:innen besser zu verstehen, haben wir eine Eyetracking-Schulstudie ($N = 131$) durchgeführt, in der wir Transferschwierigkeiten bei der adäquaten Anwendung mathematischer Inhalte in einem physikalischen Kontext anhand von Blickdaten und Interviewdaten für zwei verschiedene Darstellungsformen kognitiv auflösen konnten: Formeln und Graphen. Dabei konnten wir zeigen, dass kognitive Prozesse kontextabhängig sind.

3. In einem letzten Schritt wurden Methoden der Netzwerkanalyse eingesetzt, um Lösungsstrategien auf Basis geometrischer Muster (z.B. Steigungsdreiecke) aus Blickdaten zu erkennen. Es wurde experimentell untersucht, welche Netzwerkmetriken geeignet sind, um Richtig- von Falschlösenden zu unterscheiden. Ergebnisse zeigen hier, dass es geeignete Netzwerkmetriken gibt – ein erster Schritt zur Entwicklung eines Prädiktors zur Unterscheidung zwischen korrekten und falschen Lösern.

Zusammenfassend bietet diese Arbeit auf der Grundlage von drei umfangreichen Studien, mit Stichproben von 131 (Eyetracking-Studie) bis ca. 250 Personen (Paper-Pencil-Tests) und einer Vielzahl unterschiedlicher Methoden (Paper and Pencil, Eyetracking, Interviews, Netzwerkanalyse), ein breites Spektrum zur Analyse von Lernprozessen für lineare Funktionen im Kontext von Mathematik und Physik. Die Ergebnisse zeigen, dass Schüler:innen erhebliche Schwierigkeiten haben, mathematische Inhalte im physikalischen Kontext adäquat zu interpretieren. Die hier angewandte Analyse kognitiver Prozesse bildet eine Grundlage, um das Lernen in diesem Bereich zukünftig nachhaltig zu verbessern. Eine Lehrkräftefortbildung zu diesem Thema hat dies bereits nachhaltig umgesetzt.

Acknowledgement

I thank my supervisor, Prof. Dr. Stefan Ruzika, for his extraordinarily inspiring scientific and personal support with all the impulses, concerns and questions regarding my work and for his encouragement and motivation. I would like to thank Prof. Dr. Pascal Klein for his work as second reviewer of my dissertation. Special thanks to Prof. Dr. Jochen Kuhn for his valuable didactic input, his extremely inspiring motivation and enthusiasm for my topic, and his enriching ideas for this work. Many thanks to Prof. Dr. Claudia Redenbach for her patience and expertise on all my statistical questions and to Jun. Prof. Dr. Bianca Watzka for her great support in the replication study. My sincere thanks also go to Dr. Stefan Küchemann for the inspiring scientific exchange and support of my work on linear formulas and eye-tracking and to Prof. Dr. Sebastian Becker-Genschow for his enriching and intensive support, especially at the beginning of my PhD. I thank StD Petra Hüther for the valuable discussions during the preparation of the teacher training course. Many thanks to all the students and teachers who participated in the studies and the teacher training. I thank all the hardworking proofreaders and all the graduates I supported during my dissertation for their work and inspiration. I am grateful for the financial support provided by the projects "The Future of STEM Learning" of the Deutsche Telekom Foundation, "U.EDU," which is part of the BMBF-funded initiative "Qualitätsoffensive Lehrerbildung," and the project "Schulentwicklung für mathematische Modellierung in MINT-Fächern+, ÜR, 3 (SchuMaMoMint+¨UR3)," which is funded by the Ministry of Education of Rhineland-Palatinate with funds from the European Social Fund Plus (ESF+). Last but not least, I thank my colleagues and my family for their support.

Contents

1. Introduction

1.1. Motivation

Linear functions are the simplest form of functions. They are ombiguous parts to describe, explain, understand and approximate phenomena not only in mathematics, but in science in general. Most students learn about them in lower secondary education. Because of its simplicity, the reader might conclude that this form of function must be the one that students understand best. Nevertheless, it is worrying that students have problems understanding linear relationships. Linear relationships are also found in other subjects. In the context of this work, the representation of a (point) object in the space-time coordinate system as a linear relationship is particularly relevant. If the speed of such an object is to be determined, it is mathematically nothing more than to determine the slope value of the linear space-time function in a physical context. However, students face many difficulties in recognizing and interpreting mathematical content in other contexts. In this thesis, we investigate the influence of contextual embedding in linear function tasks. Specifically: Do students solve linear function problems better when they are purely mathematical or when they are embedded in a physical context? To this end, we examine the learning processes involved in solving these problems. One goal of this work is also to investigate to what extent learning content about linear functions can be embedded in adaptive learning systems using network analysis methods.

Numerous empirical studies confirm that challenges in transferring mathematical knowledge to other contexts, such as physics, are common: Deficits have been identified in this area, especially regarding context-dependent solution strategies (Carli et al. (2020), Ceuppens et al. (2019), Ivanjek et al. (2016), and Van den Eynde et al. (2019)). The challenge is not often due to a lack of mathematical skills, but the construction of an adequate mathematical-physical framework. In addition, these tests—which are often paper-based—have not been used across countries or grade levels. Their transferability is therefore limited. We focus on this research gap. To this end, a paper-based study of linear functions regarding mathematics and physics by Ceuppens et al. (2019) was conducted on Belgian 9th graders ($N = 253$). We replicated this study for German 9th graders ($N = 249$) and extended it across grade levels to German 11th graders ($N = 298$). The results show large differences in performance, but the fundamental challenges identified by Ceuppens et al. (2019) remain valid across countries and year groups.

Nevertheless, the cognitive processes underlying these challenges remain unclear. Validated paper-based test instruments cannot resolve cognitive processes in detail because solution strategies can only be analyzed on the basis of students' written explanations. Therefore, we conducted an eye-tracking (ET) study in schools ($N = 131$). In combination with gaze-directed retrospective think aloud (RTA), we analyzed cognitive processes involved in solving problems in mathematics and physics at a visual level. We did this for two different types of representations: graphs and formulas. Significant context-specific visual differences could be resolved and interpreted, supported by interview data collected from students. Mathematical knowledge is often correctly applied in mathematical context, e.g., when determining the slope of a linear function (equation). In the physics context, mathematical knowledge is not adequately transferred. Entirely new solution strategies are used, even though the problems are designed to be solved using the same solution strategy. For example, the arbitrary linking of coefficients of a linear equation or the linking of segments of an axis dominate. Our results show that mathematics is not interpreted adequately in a physics context.

The differences in gaze data for context-varying tasks can be used to enrich future adaptive learning systems (ALS). These systems support the learner individually, by optimizing the learning process via support at an individual learning pace. To achieve this, two ideas were pursued:

- First, we investigated whether certain solution strategies based on geometric patterns, such as slope triangles, can be detected algorithmically using network analysis methods. We detected slope triangles algorithmically by using temporal networks with a maximum precision of 0.66 and an average recall of 0.60.

- Second, we studied whether network metrics could differentiate between gaze data of correct and incorrect solvers. There are existing studies using network metrics and ET, in which none of them concerns mathematics and physics. We identified suitable network metrics. This is a first step toward developing a predictor that can distinguish between these groups and intervene accordingly. It can recognize incorrect solvers and provide assistance if necessary, or encourage and motivate correct solvers during their learning process.

The key scientific findings on the replication and grade level extension of the study by Ceuppens et al. (2019) and the findings from the (network-based) ET data analysis were communicated in a teacher training course ($N = 4$). In addition, this training was expanded to include numerous materials to raise awareness of transfer difficulties not only between mathematics and physics, but also between mathematics and other contexts. According to the participant survey, awareness was raised and valuable suggestions were taken away for their own lessons.

In summary, this thesis uses various different methods (paper and pencil, ET, RTA, network analysis) to empirically analyze learning processes for linear functions regarding mathematics and physics. The samples used are ranging from 131 (ET) to about 250 people (paper-pencil tests) based on three extensive studies. Students face serious challenges in interpreting mathematical content adequately in a physical context. The analysis of the cognitive processes used here forms a basis for the sustainable improvement of learning in the future. A teacher training course on this topic has already been implemented, at least on a small scale. Network-based approaches can detect solution strategies and distinguish between correct and incorrect solvers, and are a promising step to improve learning systems of the future.

1.2. Outline of Thesis

This introduction is followed by a chapter in which we cover necessary basics needed in the remaining of the thesis. Here, we present the current state of research on challenges with linear functions in mathematics and physics and the state of network analysis methods in ET research and explain the basics of the analysis methods (ET and RTA), statistical methods and network analysis methods used in the course of the dissertation. In the third chapter, we present the results of the ET studies on graphs and formulas. In the fourth chapter, we managed the detection of slope triangles from gaze data using temporal networks and the differentiation of correct and false solvers based on network metrics. This is followed by a final discussion and an outlook.

1.3. Credits and Contributions

During my time in KOMMS and the optimization working group, I had the opportunity to work on three projects: "The Future of STEM Learning" by the Deutsche Telekom Foundation, "U.EDU," which is part of the BMBF-funded initiative "Qualitätsoffensive Lehrerbildung" and "Schulentwicklung für Mathematische Modellierung in MINT-Fächern+, ÜR, 3 (SchuMaMoMint+¨UR3)," funded by the Rhineland-Palatinate Ministry of Education with funds from the European Social Fund Plus (ESF+). I am grateful for the experience and opportunities that working on these projects has given me.
Parts of this thesis are the result of fruitful cooperations, which I like to mention here briefly: Chapter 3 is based on cooperation with Prof. Dr. Jochen Kuhn (Sections 3.1 and 3.2), Prof. Dr. Claudia Redenbach (statistics in Section 3.1), Prof. Dr. Sebastian Becker-Genschow (Section 3.1 and Section 3.2 on graphs),

Junior Professor Dr. Bianca Watzka (Section 3.1), and Dr. Stefan Küchemann (Section 3.2 on formulas). Valuable ideas for the teacher training were developed in a discussion with senior lecturer Petra Hüther (Section 3.3). The ideas for Chapter 4 have been developed together with Anna Münz (Section 4.1) and Simon Schwarzmüller (Section 4.2), which I helped to supervise during their final theses. I like to take this opportunity to thank everyone involved for their valuable cooperation.
Results of Section 3.2 on graphs are based on a joint publication:

- Becker, S., Knippertz, L., Ruzika, S., Kuhn, J. (2023). "Persistence, context, and visual strategy of graph understanding: Gaze patterns reveal student difficulties in interpreting graphs." *Physical Review Physics Education Research* 19.2.

An article on the results about formulas in Section 3.2 is currently in the process of being published:

- Knippertz, L., Ruzika, S., Kuhn, J. "Equations – More a Matter of Context than Text." *Educational Studies in Mathematics* (in preparation).

The publication of the results for Section 3.1 is still planned and main results of Section 4.1 were published at a conference:

- Knippertz, L., Münz, A., Ruzika, S. (2023). "Automated Detection of Geometric Structures in Gaze Data." *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications. Association for Computing Machinery, New York, NY, USA,* 34.1.

In addition, I like to mention the following jointly submitted or published articles, which are thematically related to my work and were also written as part of my doctorate:

- Becker, S., Knippertz, L., Ruzika, S., Kuhn, J. (2022). "Blickdatenanalyse von Prozeduren bei der Interpretation linearer Graphen im mathematischen und physikalischen Kontext." In Baum, M., Eilerts, K., Hornung, G., Roth, J. & Trefzger, T. (Eds.), *Band 1: Perspektiven des (digitalen) MINT-Unterrichts im 21. Jahrhundert. Springer.*

- Knippertz, L, Becker, S., Kuhn, J., Ruzika, S. (2022). "Gaze pattern analysis to reveal student difficulties in interpreting kinematic graphs." In Hodgen, J., Geraniou, E., Bolondi, G. & Ferretti, F. (Eds.), *Proceedings of the Twelfth Congress of the European Society for Research in Mathematics Education (CERME12). Free University of Bozen-Bolzano and ERME.*

- Kennel, K., Knippertz, L., Ruzika, S. "Perspektiven zu blickgesteuerten adaptiven Lernsystemen in Mathematik." In Wehn, N., van Waveren, L. & Gómez Tutor, C. (Eds.), *QLB@RLP - Eine Bestandsaufnahme der Lehrkräftebildung in Rheinland-Pfalz.* (submitted on 06.09.2024).

2. Foundations

This chapter introduces the theoretical background, as well as the methodological methods necessary for the empirical analysis in this thesis. Challenges with linear functions are the primary cause of problems in understanding linear relationships. In the first section of this chapter, we closely examine the challenges that cause problems in understanding linear relationships. We do this by examining the challenges that arise when combining disciplines such as mathematics and physics. Furthermore, we examine the role of multiple representations. In particular, do we read text and equations differently? In addition, we discuss how graphs and formulas are cognitively processed and which steps in the process cause problems of understanding. These fundamentals of understanding graphs and formulas are followed by an introduction to the methodological tools we use to measure visual attention in order to gain deeper insights into cognitive processes related to linear functions in mathematics and physics: ET and RTA. These two methods are used in the ET studies in Section 3.2. The focus here is on the relationship between ET and cognition. Next, an introduction to networks and network metrics is given, which form the basis for the network-based analysis of gaze data in Sections 4.1 and 4.2. We conclude the section on the theoretical background of this work with a derivation of the research questions. In the second section of this chapter, we explain the study design. Furthermore, the data analysis requires qualitative and quantitative statistical methods, which are discussed here.

2.1. Theoretical Background and Research Questions

(Linear) Functions

In the first step, we will look at context-specific problems when solving tasks involving linear functions. We focus on mathematics and physics.

Challenges Connecting Mathematics and Physics

Although there is a close relation between mathematics and physics (Redish, Kuo (2015)), students face challenges connecting both disciplines adequately. The reasons for this include, on one hand, a strong emphasis on domain-specific learning (Leinhardt et al. (1990) and Pollock et al. (2007)), which is also embedded in school, since teaching is only done in one subject and mostly not interdisciplinary. Additionally, students have difficulties transferring knowledge from mathematics to physics (Christensen, Thompson (2012) and Leinhardt et al. (1990)).

A study by Ceuppens et al. (2019) investigated ninth graders' understanding of linear functions regarding kinematics and mathematics. They found difficulties related to kinematics, formulas, and the interpretation of functions with negative slope. In addition, negative signs were observed in the mathematical context but often ignored in the similar kinematic context.

Planinic et al. (2012) investigated the understanding of the slope of linear relationships in mathematics and physics. Contrary to the teachers' assessment, students perform better on math problems than on physics problems. Students are often unaware of the similarity of the problems and use different strategies to solve the problems. Planinic et al. (2012) suggest that this is due to a lack of ability to interpret mathematical quantities in a physical context.

Examining math and physics problems, Van den Eynde et al. (2019) report that math problems were solved correctly more often than physics problems. Context-dependent solution strategies were evident in this study.

Ivanjek et al. (2016) also examined graph interpretation in different contexts (mathematics, physics, and others). They show that graph interpretation strategies are context dependent and underlie transfer difficulties. Furthermore, formulas were preferred when solving physical problems (Ivanjek et al. (2016)).

Carli et al. (2020) developed a test instrument consisting of pairs of problems in mathematics and physics. Their results show that good performance in mathematics does not guarantee the successful application of mathematical skills in physics. They observed context-dependent solution strategies. They found that students tend to get stuck in calculations without understanding the problems. The main problem is not a lack of mathematical skills, but the construction of an adequate mathematical-physical framework. On this basis, Carli et al. (2020) demand the training of the competence of mathematizing in physics, instead of relying only on technical, mathematical procedures.

To summarize, although mathematics and physics are very closely related, problems are solved differently in the two fields. There are also problems in transferring mathematics to physics. The reason for these challenges is often not a lack of mathematical skills, but a lack of competence in mathematization and interpretation of non-mathematical content.

To better understand the cognitive processes behind these problems, understanding the cognitive processing of multiple representations is necessary. In particular, we will highlight the differences between text and formula processing and examine how graphs and formulas are cognitively processed.

Problems With Multiple Representations

Multiple representations, i.e., the joint representation of different representations with the same reference object, e.g., graph and formula, are a major hurdle for students (Ainsworth (2006) and Nieminen et al. (2010)). They find it particularly difficult to switch between different representations (Even (1998)).

People read equations and text differently. There are two main reasons for this: The nonlinear structure of equations and semantically similar equations trigger different cognitive processes in different subjects. As Andrá et al. (2015) already recognized, this non-linear structure must be considered, when decoding the symbols of the mathematical equation. For example, if you read the functional equation $y = 2 \cdot (3x + 4)$ as "y is equal to 2 times three x plus 4", you would also be referring to a different structure, namely $y = 2 \cdot 3x + 4$. Reading and understanding equations varies from discipline to discipline. For example, equations are interpreted differently in mathematics and physics (Redish, Kuo (2015)). Redish, Kuo (2015) illustrated this with a problem: Given a temperature measurement with the result $T(x, y) = k \cdot (x^2 + y^2)$, where k is a constant, physicists and mathematicians are asked to complete the function $T(r, \phi) =?$. The physicist presumably identifies the result of a temperature measurement as a physical function representing the temperature at a given location. He physically interprets $x^2 + y^2$ as the square of the distance from the origin and ϕ and r as polar coordinates. The mathematician completes the function as $T(r, \phi) = r^2 + \phi^2$, emphasizing the mathematical grammar of the expression over the physical meaning. He sees x, y, r and ϕ as dummy variables. Despite the differences, there are still some similarities between reading equations and reading language: "Making of meaning with equations shares (at least) three key commonalities with meaning-making in language: an embodied basis, the use of encyclopedic knowledge, and contextual selection of that encyclopedic knowledge for meaning-making" (Redish, Kuo (2015, p. 573))." Both text and equations with the same structure can be interpreted differently. Interpretation depends on prior knowledge, disciplinary culture, and contextual embedding.

In summary, there are differences and similarities between equations and text in terms of reading and a number of related open questions: How can cognitive differences and similarities in reading equations and text be analyzed? Are there differences in reading direction when processing text and equation elements?

Cognitive Processing of Graphs and Formulas in Mathematics and Physics

To understand graphs and formulas properly, learners must extract information from a graph or formula and combine it with prior knowledge. Such processes are described by the cognitive theory of multimedia learning (CTML, e.g., Mayer (2009)). Based on the dual coding theory (Paivio (1990)) and the theory of working memory (Baddeley (1992) and Chandler, Sweller (1991)), CTML states that for deep learning, learners must be able to use their cognitive processes to select, organize, and integrate information into existing prior knowledge (Mautone, Mayer (2001), Mayer (2005), and Mayer (2008)). According to Mayer (2005) and Was et al. (2016), the learning and processing process is as follows:

1. Information (graphics, textual content) enters the information processing system through the visual channel. The learner then selects relevant information for further processing or actively engages with it.

2. The selected information is transferred to working memory, where words are organized into a coherent verbal representation and images (e.g., linear graphs) are organized into a pictorial representation.

3. The verbal or pictorial models just mentioned are now integrated with relevant prior knowledge to construct a coherent mental model.

In addition to this theory explaining the cognitive processing of graphs and formulas, there are specific models that deal with the cognitive processing of graphs or formulas. In the following, we will first discuss the cognitive processing of graphs and then that of formulas.

Proper interpretation of graphs requires two central skills: spatial imagination and logical reasoning (Planinic et al. (2012)). Spatial reasoning can be defined as the ability to generate, retain, retrieve, and transform well-structured visual (mental) images (Lohman et al. (1996)). People with low spatial ability have difficulty extracting information from graphs (Vekiri (2002)). Students with logical thinking problems have difficulty creating and interpreting graphs (Berg, Phillips (1994)). They solve problems with graphs based on their perceptions and low-level thinking, which leads to the following major sources of errors in interpreting graphs (Beichner (1994), Berg, Phillips (1994), and McDermott et al. (1987)):

1. **Slope-Height Confusion:** The ordinate intercept is interpreted as a slope.

2. **Graph-as-Picture Error:** The graph is interpreted as a direct representation of the real location.

3. **Point-Interval Confusion:** A point value is considered instead of an interval. For example, the slope is calculated as a quotient of point values instead of intervals.

4. **Slope Error:** A non-zero ordinate intercept is ignored when determining the slope.

To understand the difficulties associated with the representation form formulas, we must first understand how formulas are processed.

In mathematics, when we read equations, we read mathematical symbols that represent mathematical objects. The gap between mathematical symbols and the symbolized mathematical objects leads to comprehension problems (Duval (2006), Frege (1970), and Radford (2002)). Mathematical objects, including equations, can be represented in different ways using symbols, for example, a linear function can be expressed as the mathematical formula $f(x) = 2x + 3$ or as a graph in a coordinate system. Even individual symbols in the mathematical equation can be replaced by other symbols without changing the linear function, e.g., by $\frac{4}{2}v + 2.\bar{9} = g(v)$. The following changes have been made: The dummy variable x has been replaced by the variable v. The function name f has been changed to g. The linear term previously had a coefficient of 2. Since the fraction $\frac{4}{2}$ can be reduced to 2, it was replaced accordingly. The constant 3 can also be written as $2.\bar{9}$, as $2.\bar{9} = 2 + 0.\bar{9} = 2 + \frac{3}{3} = 2 + 1 = 3$. In addition, both sides of the equals sign have been swapped. All possible symbolic representations remain mere reflections of the actual mathematical object, which cannot be fully described by symbols (Duval (2006)).

Besides the difference between mathematical symbols and symbolized objects, mathematical objects are subject to two types of differently cognitively processed transformations: conversations and treatments (Duval (2006)). Conversions are transformations of representations. Only the register of the semiotic representation (Duval (2017)) is changed, not the designated object, such as the transformation of a functional equation into a graphical representation. Treatments are transformations of representations that occur within the same register. For example, a functional equation can be understood dualistically as an object or as a process. The concept *function* can be understood as a set of ordered pairs or as a well-defined method to get from one system to another (Sfard (1991)).

Gray, Tall (1994) go one step further by integrating the symbol into the process-object duality. They define *procept* as a process that produces an object and a symbol that represents either the process or the object. Less successful students remain in procedural thinking and fail to see the object. These difficulties arise, for example, from seeing the function $f(x)$ as a procedure for calculating a particular value for x without seeing $f(x)$ as an object, namely a function for a general x.

Students must find a way to understand these dualities and transformations in order to understand the nature of the mathematical content being represented. Understanding mathematical equations in a scientific context involves different cognitive processes.

In contrast to the phenomena of physics and other natural sciences, mathematical objects cannot be understood with the help of instruments (microscopes, etc.) or human perception, but only through symbols or systems of symbols (Duval (2006)). Although mathematics and physics “linguistically” use the same formal syntax, there are considerable differences in meaning and usage in both fields (Redish, Kuo (2015)).

Solving physical problems with the help of mathematics requires a modeling process (Redish, Smith (2008)):

1. The physical problem in the physical system is translated into the language of mathematics,
2. solved there,
3. then interpreted in the physical system,
4. and then it is checked whether the original physical system is still adequate.

Thus, mathematics plays a threefold role in physics: in mathematizing, in doing pure mathematics, and in interpreting.

Uhden et al. (2012) refine these three roles of mathematics in physics and thus allow to distinguish between different levels of interpretation of mathematical results and pure mathematization. On this basis, they developed a new modeling cycle for the treatment of physical problems, which considers the deeply interwoven connection between mathematics and physics.

In summary, understanding equations in mathematics is a cognitively complex process that becomes even more complicated when formulas are represented as non-mathematical relations.

The next step is to explain how visual attention can be used to analyze cognitive processes in solving graph and formula problems involving linear functions in mathematics and physics.

Visual Attention

When solving linear function problems, we read the task description and examine the presented representations as formulas or graphs. To do this, we move our eyes to focus our concentration, our visual attention. By tracking someone else's eye movements, we can follow the path of the observer's attention. The goal is to gain insight into what the observer finds interesting, or even into the cognitive processes involved in solving the task. In this section, we will introduce the basic concepts of ET and RTA. In particular, we will explain how ET works, closely examine the relationship between cognitive processes and ET, and look at ET research in the context of graphs and formulas. The section concludes with principles of RTA.

Eye-Tracking Fundamentals

ET is a method of studying visual attention by recording eye movements. The interpretation of eye-movement data is based on the following two hypotheses, which are widely accepted in cognitive sciences:

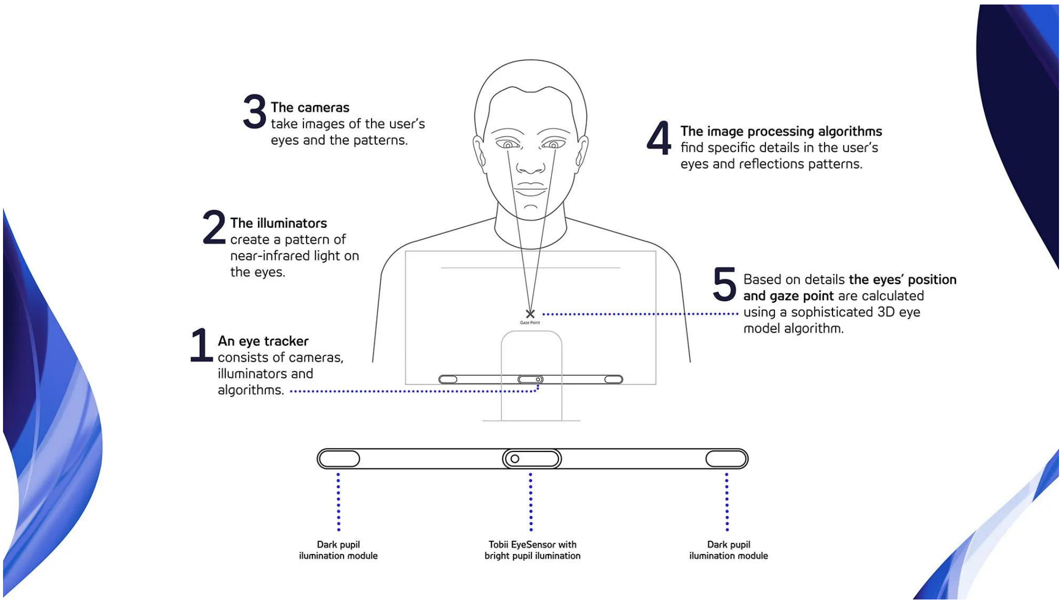


Figure 2.1.: Functionality of ET method, taken from Ieva Miseviciute (2024).

1. The Eye-Mind Hypothesis: The hypothesis of Just, Carpenter (1976) states that what one looks at is cognitively processed. This hypothesis was later confirmed by neuropsychology (Kustov, Robinson (1996)).

2. The Immediacy Assumption: The assumption by Just, Carpenter (1976) is that what you look at is processed without time delay.

Thus, eye movements map the temporal-spatial decoding of visual information (Hoffman, Subramaniam (1995) and Salvucci, Anderson (2001)). Since these hypotheses allow the study of cognitive processes based on gaze data, ET is highly beneficial for educational research. Complex cognitive processes that are difficult to communicate can be observed without interrupting the student and thus without interrupting the solution process (Strohmaier et al. (2020)).

In the following, some basics of ET will be introduced.

The functional principle of a stationary eye tracker can be described as follows: The eyes are exposed to infrared light, which causes reflections on the eye. These reflections are recorded by a camera and the position of the gaze is calculated from the reflections. Figure 2.1 illustrates the principle of operation.

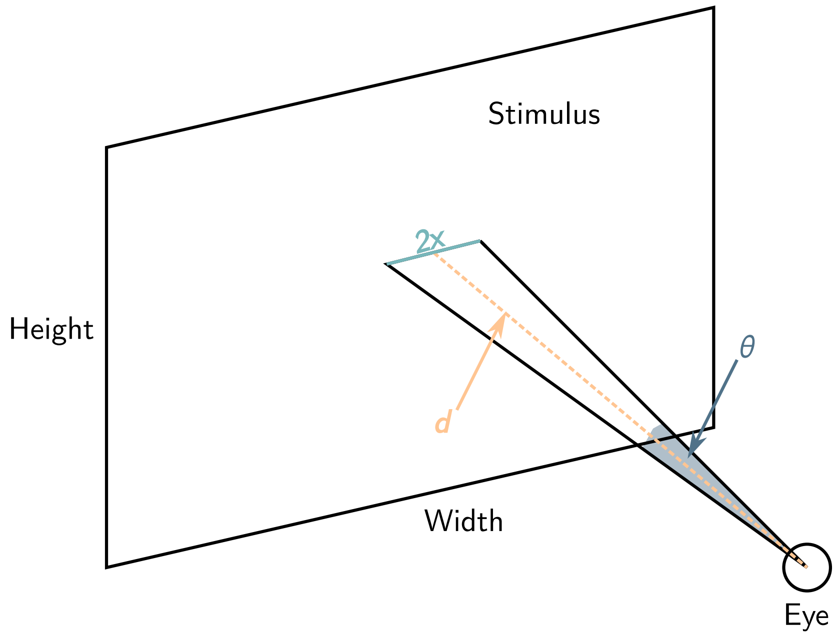


Figure 2.2.: Geometric relationship between stimulus unit x and degrees of visual angle θ, given the viewing distance d. (Adapted from Holmqvist et al. (2011, p.24).)

Figure 2.2 shows the geometric relationship between the stimulus unit x and the visual angle θ, for a given viewing distance d:

$$\tan\frac{\theta}{2} = \frac{x}{d}.$$

This relationship only holds if the stimulus is viewed in the central line of sight (small visual angle θ).

When recording eye movements, different events can be detected, such as fixations. A fixation describes the resting of the eyes on a point for a short period of time where they are relatively still, holding the central foveal vision in place so that the visual system can take in detailed information about what is being looked at (Holmqvist et al. (2011)). The duration of a fixation is highly variable, but is often between 200 and 300 ms (Holmqvist et al. (2011)). In the used ET analysis software, *TobiiStudio*, fixation is a sequence of raw gaze points, where the estimated velocity is below the velocity threshold set in the Identification-Velocity Threshold (I-VT) gaze filter (Tobii AB (2017)). The eye movement that shifts the gaze from one fixation to the next is called a saccade and usually lasts about 30 to 80 ms (Holmqvist et al. (2011)). This movement begins with an initial rapid acceleration until the eye reaches a peak velocity. After the peak is reached, it begins to decelerate until the eye reaches the target location. In the ET software, a saccade is a sequence of raw gaze points where the associ-

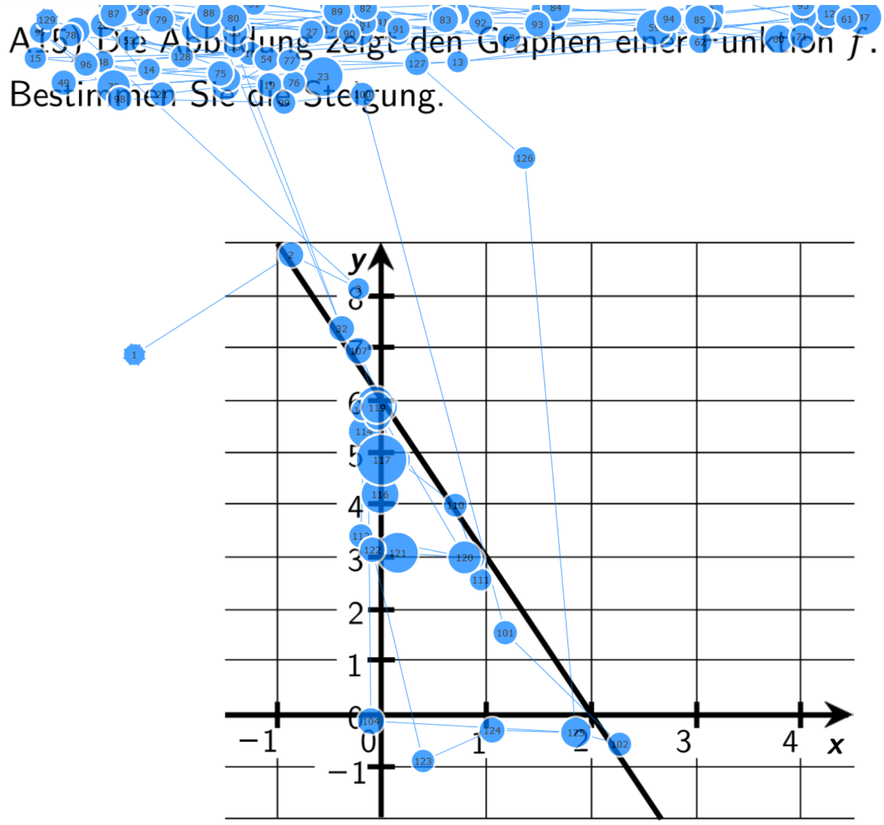


Figure 2.3.: Gazeplot of Person No. 94 solving item M8 from the test instrument by Ceuppens et al. (2019). Fixations are shown as blue points and saccades are blue lines between those points.

ated velocity is above the velocity threshold set in the I-VT gaze filter (Tobii AB (2017)).

Eye movements can be described as a sequence of fixations and saccades and visualized in a gaze plot (see Figure 2.3). In the gaze plot in Figure 2.3, the blue circles represent the fixations and the number within the circles represents the order of the fixations. The diameter of the blue circles represents the duration of the fixations. The larger the diameter, the longer the fixation duration. The advantage of the gaze plot is that it visualizes the sequence of fixations in relation to their position. However, the visualization contains a high level of detail, which can lead to a certain lack of structure. If certain regions of the stimulus are to be examined, areas of interest (AOIs) are a helpful tool to analyze these regions separately (Holmqvist et al. (2011)). The field of view is divided into predefined AOIs depending on the item at hand. The distribution of gaze data over the AOIs is used to study cognitive processes using ET metrics such as total visit duration (TVD, cumulated times between first fixation in and first fixation out

Table 2.1.: ET measures and associated cognitive processes based on Alemdag, Cagiltay (2018), Scheiter et al. (2019), and Hahn, Klein (2022).

ET Measure	Definition	Cognitive Processes	Use
Total visit duration	Duration of all visits within an active AOI (or AOI group)	Selection/organization	Used as a measure of attention or information processing (long durations indicate high attention).
Fixation count	Number of all fixations within an AOI (or within all AOIs belonging to an AOI group).	Selection/organization	Also used as a measure of attention (high numbers indicate high attention).
Transition count	Number of gaze shifts between predefined AOIs.	Integration	Measure of integration processes between different sources of information, e.g., gaze shifts between text and image elements when reading.
Scanpath	Trace of a participant's eye movements in space and time.	Integration	Used to link different pieces of information together, e.g., to determine the number of objects in an image.

of an AOI, AOIs not visited are counted as 0) and transitions (saccades between AOIs).

Connection between Eye-Tracking and Cognitive Processes

The following connection between CTML (cf. Section 2.1) and ET allows for a theory-based interpretation of ET data: Gaze durations (TVD, fixation durations) measure how long information is accessed. High gaze durations indicate more visual attention to certain areas (Hahn, Klein (2022)). They are associated with processes of selection and organization of information extracted from the processed material (e.g., Alemdag, Cagiltay (2018), Scheiter et al. (2019), and Schüler (2017)). Gaze shifts (transitions) are associated with integration processes, e.g., one person studying text-image integration during reading (e.g., Alemdag, Cagiltay (2018). For a more detailed overview on connection between ET metrics and cognitive processes, see Table 2.1.

In summary, ET is a non-intrusive method to obtain information about visual attention and cognitive processing in problem solving.

Research in the Context of Graphs and Equations

In ET studies of line graph problems, it has been shown that students' different strategies for interpreting kinematic diagrams are manifested in their gaze patterns (Becker, Küchemann, et al. (2022)). When determining the slope of line graphs, Klein et al. (2019) found that students who correctly determined the slope directed more visual attention to the graph. For graphs in different contexts, Susac, Bubic, Kazotti, et al. (2018) found that physicists solved physics and finance tasks more successfully than psychology students. Physics students also showed significantly higher fixation durations on graphs than psychology students (Susac, Bubic, Kazotti, et al. (2018)). A replication study of this work by Klein et al. (2019) with business students instead of psychology students showed higher learning success of physics students compared to business students. However, they observed that both cohorts of students directed a similar amount of visual attention to the graph domain. These results by Klein et al. (2019) show that examining visual attention to the entire graph area alone is not sufficient. Therefore, the graph domain should be further subdivided and examined extensively to account for differences in student learning.

ET has been used to study algebraic or arithmetic equations (among others: Susac, Bubic, Kaponja, et al. (2014), simple number comparisons (Merkley, Ansari (2010)), division rules (Potgieter, Blignaut (2017)), addition tasks (Green et al. (2007)), comparisons of different representations (among others: Rello et al. (2013)), and proofs (Inglis, Alcock (2012)).

There is a wealth of research investigating arithmetic and number recognition using ET (Mock et al. (2016), Hartmann (2015), Hegarty et al. (1995), Rello et al. (2013), Suppes et al. (1983), Merkley, Ansari (2010), Green et al. (2007), Susac, Bubic, Kaponja, et al. (2014), and Schneider, Maruyama, et al. (2012)). For example, Hegarty et al. (1995) examined arithmetic word problems. The authors found that unsuccessful students focused on numbers and relational terms, while successful students had a higher number of fixations on variables. Therefore, successful students tended to use a strategy that focused on problem modeling.

Furthermore, Merkley, Ansari (2010) examined numerical comparison tasks with different size ratios (smaller to larger number). They found that when the ratio of numbers was larger, participants were significantly slower and worse at deciding which number was larger. Participants had more fixations on the correct number than on the incorrect number (at all ratios), and fixations were longer and more frequent as the ratio between the numbers increased.

In addition, Green et al. (2007) investigated the dependence of age on the processing of complex addition tasks. They found that young participants chose strategies more adaptively than older participants. Eye movements confirmed the use of strategies by the participants and showed that the strategies of older participants were less clearly differentiated than those of young participants.

In physics education, Susac, Bubic, Kaponja, et al. (2014) investigated the rearrangement of algebraic equations according to x, e.g., $xa = b$. They measured a positive correlation between the number of fixations and task difficulty; the correlation between interviews and gaze plots was partly contradictory.

Gaze-Cued Retrospective Think Aloud Interviews

There are three ways to collect verbal data through thinking aloud (TA): concurrent TA, retrospective TA (RTA), and cued RTA (Holmqvist et al. (2011)). A setup is called concurrent TA when participants speak while performing a task. Alternatively, TA can be recorded after the task has been completed (RTA). RTA can be cued or uncued. In a cued RTA, the participant's memory is cued by showing him or her a recording of his/her own eye movements (gaze-cued RTA) or a pure video recording while his/her verbalization is being recorded. In an uncued RTA, the user provides a description without any cue.

The triangulation of ET measures with gaze-cued RTA data is often requested in educational research as it provides deeper information about the learning process (Kiili et al. (2014)). Furthermore, the combination of ET with gaze-cued RTA has advantages over RTA methods alone, as the collection of verbal data is based on eye movement data observed during replay (Van Gog et al. (2005), Hyrskykari et al. (2008), Olsen et al. (2010), Alhadreti et al. (2017), Van der Weijden et al. (2018), and Prokop et al. (2020)). Especially in mathematical tasks, this method has proven useful for gaining deeper insights into the way students solve problems than by analyzing eye movements alone (Schindler, Lilienthal (2019)). Despite these advantages, only a few works have used this triangulation in the study of solution processes in mathematical tasks (Strohmaier et al. (2020)).

RTA methods are widely used in usability testing. Here, comparisons between gaze-cued RTA and other uncued RTA methods—such as concurrent TA—indicate that using any kind of cue is better than not using it (Hyrskykari et al. (2008) and Olsen et al. (2010)). Cued methods allow participants to identify more problems, a higher number of words and comments, and to mention more problems (Hyrskykari et al. (2008) and Olsen et al. (2010)).

Guan et al. (2006) show that cued RTA is a valid and reliable method that provides a valid account of what people paid attention to while completing tasks. It has a low risk of introducing inventions, and its validity is unaffected by complexity. In addition, it provides information about users' strategies in completing tasks (Guan et al. (2006)).

The combination of video cued and gaze cued RTA can help participants to detect more problems and remember more details (Alhadreti et al. (2017)). Bicer, Bicer (2022) showed in a qualitative study that mathematical creative processes of young students can be analyzed by ET-stimulated recall techniques. Even the analysis of ET data and TA data from different samples, solving the same test instrument, allows conclusions about the cognitive processes involved in problem solving in geometry (Uygun et al. (2022)). On the other hand, Alhadreti et al. (2017) and Elling et al. (2011) found that participants may be distracted while watching their eye movements. This may have implications for verbalization performance.

In summary, RTA is a valid and reliable method that provides information about users' strategies. In particular, the triangulation of ET with RTA data is suitable for educational research because it provides insights into learning processes. In addition to ET and RTA, we use mathematical networks to detect solution strategies and to distinguish incorrect from correct solvers from gaze data.

Mathematical Networks

We want to apply methods of network analysis to analyze ET data. The goal is to detect geometric structures from gaze data and to find network metrics that distinguish between correct and incorrect solvers. This requires a basic knowledge of graphs, temporal networks, and network metrics, which is presented in this section.

Graph Theory

To use mathematical tools to analyze ET data, we need some basic concepts from graph theory. The goal is to model ET data using mathematical networks. In this section, we introduce mathematical networks and their properties.

Roughly speaking, networks consists of elements and connections between those elements. More precisely, in mathematical way, networks can be described by graphs.

A *directed graph* $G = (V, E)$ consists of a node set V with cardinality n, whose elements are called nodes and an edge set E with cardinality m. The elements of E are called *arcs or directed edges* and they are represented by an ordered pair (u, v) of nodes $u, v \in V$. For an arc $(u, v) \in E$, $u \in V$ is called the *origin* and $v \in V$ is called the *destination*. Two nodes which are connected by an arc $e \in E$ are called *adjacent*. If $u = v$ for an arc $(u, v) \in E$, the arc is called *loop*. In the following, all graphs are assumed to be directed and loop-free unless stated otherwise.

The *in-degree* $\delta^-(v)$ of a node $v \in V$ in a directed graph G is the number of edges with destination v. Analogously, the *out-degree* $\delta^+(v)$ of a node $v \in V$ is the number of edges with origin v.

The *reciprocal degree* of a node v is defined as $\frac{1}{\delta^-(v)+\delta^+(v)}$.

A sequence

$$v_1, e_1, v_2, \ldots, v_k, e_k, v_{k+1}$$

with $e_i = (v_i, v_{i+1})$ for all $i = 1, \ldots, k$ is called a *path* of length k. If $v_1 = v_{k+1}$ the path is called a *cycle*. The adjacency matrix is a common representation of a graph. For the adjacency matrix, the nodes are identified with the numbers

$1, \ldots, n$. The adjacency matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ of a directed graph G is defined as follows

$$a_{ij} = \begin{cases} 1 & \text{if there is an edge}(i,j) \in E, \\ 0 & \text{else,} \end{cases}$$

for $1 \leq i, j \leq n$.

The *distance* $d(u, w)$ between two nodes u and w is the length of the shortest path between the two nodes.

The *neighborhood* $N(v)$ of a node v is defined according to the parameter source:

- if source is "in," then $N(v)$ consists of predecessors of node v and $|N(v)| = \delta^-(v)$.
- if source is "out," then $N(v)$ consists of successors of node v and $|N(v)| = \delta^+(v)$.
- if source is "in+out," then $N(v)$ is both predecessors and successors and $|N(v)| = \delta^-(v) + \delta^+(v)$.

Temporal Networks

To model ET data, temporal networks are utilized. The following definition is a modified version of the definition by Scholtes et al. (2016).

Definition 2.1. (*Temporal Network*)
A *temporal network* is a tuple $G^T = (V, E^T)$, where V denotes the set of nodes. The set $E^T = \{(v_1, v_2, 1), \ldots, (v_T, v_{T+1}, T)\} \subseteq V \times V \times \{1, \ldots, T\}, T \in \mathbb{N}$ consists of arcs $(v, w, t) \in E^T$, and thus, represents a numbering of the edges (v, w).

To summarize, in terms of the definition in Section 2.1 about graph theory, a temporal network is an extension of a graph. The set of nodes V is present in both. However, in the temporal network, the edges are extended by a numbering. This is important for storing temporal information in ET data, such as the numbering of a person's fixations.

Network Metrics

We want to explore the extent to which gaze data from correct and incorrect solvers can be discriminated using network metrics. To do this, we need to define these metrics. The following network metrics and definitions are (unless otherwise indicated) taken from the Python library NetworkX (Hagberg et al. (2008)), which is also used to implement the network metrics in this work. For those metrics that are not defined for an entire graph, but for individual nodes, the mean value was formed over the value of all nodes of the graph. This is a common procedure that is also used for other analyzed metrics that are defined for an entire graph, such as the clustering coefficient.

Definition 2.2. *Average Neighbor Degree, Average Degree Connectivity*
The *average neighbor degree* of a node v in a Graph $G = (V, E)$ is the average degree of the neighborhood of each node:

$$deg_{\varnothing nei}(v) = \sum_{u \in N(v)} \frac{deg(u)}{N(v)}.$$

The *average degree connectivity* is the average nearest neighbor degree of nodes with degree k:

$$con_{\varnothing deg}(G, k) = \sum_{v \in V_k} \frac{deg_{\varnothing nei}(v)}{|V_k|},$$

where V_k is the set of all nodes $u \in V$ for which $deg(u) = k$.

Depending on the choice of neighborhood (cf. Section 2.1), $deg(u)$ denotes the in-degree or out-degree of node u.

If the average degree connectivity is applied to networks of ET data, high metric values result from gaze data with many saccades. Another reason for high values can be an even distribution of node grades and / or only few vacancies in the degrees.

Next, we define a measure that describes the average length of shortest paths in networks.

Definition 2.3. *Average Shortest Path Length*
The *average shortest path length* is defined as:

$$spl_{\varnothing} = \sum_{u,v \in V, u \neq v} \frac{d(u, v)}{|V| \cdot (|V| - 1)}.$$

It is the average length of the shortest paths between all pairs of nodes in the graph G.

It should be noted here that high values of this metric in ET networks can be an indication of a high number of saccades. With many saccades, there are more direct connections between the fixations (nodes), which result in shorter average shortest paths than in gaze data with fewer saccades.

Definition 2.4. *(Normalized) Betweenness Centrality*
The *betweenness centrality* of a node v is the sum of the fraction of all-pairs shortest paths that pass through v:

$$cen_B(v) = \sum_{s,t \in V} \frac{\sigma(s, t|v)}{\sigma(s, t)},$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s,t. If $s=t$, then $\sigma(s,t)=1$, and if $v \in s,t$, then $\sigma(s,t|v)=0$ (Brandes (2008)).

The *normalized betweenness centrality* is defined as

$$\frac{cen_B(v)}{|V-1| \cdot |V-2|}.$$

In the context of ET networks, high values of this metric indicate a high number of central fixations, through which many shortest paths / gaze paths run.

In a next step, we define a measure of centrality in a network.

Definition 2.5. *Closeness Centrality*
The *closeness centrality* of a node u is the reciprocal of the average shortest path incoming distance to v over all $n-1$ reachable nodes u_i:

$$cen_C(v) = \frac{n-1}{\sum_{i=1}^{n-1} d(u_i, v)}.$$

Note, that the closeness centrality is calculated as the inverse of the sum of the lengths of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes.

Next, we define a measure of the degree to which nodes in a network tend to cluster together.

Definition 2.6. *Clustering Coefficient*
For directed graphs, the *clustering coefficient* of a node v is defined as the existing proportion of all possible directed triangles through the node v:

$$cc(v) = \frac{T(v)}{2 \cdot (\delta^-(v) + \delta^+(v)) \cdot ((\delta^-(v) + \delta^+(v)) - 1) - 2 \cdot \delta_{\leftrightarrow}(v))},$$

where $T(v)$ specifies the number of directed triangles that run through node v, $\delta^-(v)$ denotes the in-degree and $\delta^+(v)$ denotes the out-degree of node v, $\delta_{\leftrightarrow}(v)$ is the reciprocal degree of node v.

Low values of the clustering coefficient in gaze data mean that the average proportion of existing triangles in relation to all possible triangles are high. The next metric is called communicability:

Definition 2.7. *Communicability*
The *communicability* between two nodes u and v in a graph $G = (V, E)$ is the sum of the paths (weighted according to their length) of different lengths that start at node u and end at node v. Longer paths are weighted lower than shorter paths:

$$com(u,v) = \frac{NSP(u,v)}{d(u,v)!} + \sum_{k>d(u,v)} \frac{W_k(u,v)}{k!},$$

where $NSP(u, v)$ is the number of shortest paths between nodes u and v. The number of paths between node u and node v with length $k > s$ is denoted with $W_k(u, v)$.

High communicability values in gaze data indicate more gaze paths between fixations. The next metric is called network density. It is often used in passenger transportation. It is a metric that relates the length of the transportation network in an area to the area.

Definition 2.8. *Degree Centrality, Density*
The *degree centrality* for a node v in the graph $G = (V, E)$ is the proportion of nodes in V to which the node v is adjacent:

$$cen_D(v) = \frac{deg(v)}{|V| - 1}.$$

The *density* of a graph $G = (V, E)$ is the mean value of the degree centrality over all accounts of the graph. It indicates the ratio of actually existing edges compared to potentially possible edges and can be calculated as follows:

$$den(G) = \frac{2 \cdot |E|}{|V| \cdot (|V| - 1)}.$$

High density values are found in gaze data from students who focus on relevant parts of the task and ignore irrelevant parts. Next, we define a measure of the influence of a node in a connected network.

Definition 2.9. *Eigenvector Centrality*
The *eigenvector centrality* of a node v_i in the graph $G = (V, E)$ is calculated based on the centrality of its neighbors. The eigenvector centrality for node v_i is the ith element of the vector x given by the following equation:

$$Ax = \lambda x,$$

where A denotes the adjacency matrix of the graph G with the eigenvalue λ. According to the Perron-Frobenius theorem, there is a unique and positive solution if λ is the largest eigenvalue of the adjacency matrix A (Newmann (2010)).

If students look more frequently into the areas relevant for processing the task and make strategic transitions between these areas, the value of eigenvector centrality is high. In a next step, we define the harmonic centrality:

Definition 2.10. *Harmonic Centrality*
The *harmonic centrality* of a node v in the graph $G = (V, E)$ is the sum of the reciprocals of the distances from all other nodes to node v:

$$cen_H(v) = \sum_{u \in V, u \neq v} \frac{1}{d(u, v)}.$$

High values of harmonic centrality in ET networks indicate that the shortest paths between two nodes in the gaze data networks are generally shorter. Next, we define another centrality measure.

Definition 2.11. *Katz Centrality*
The *Katz centrality* of a node v in the graph $G = (V, E)$ is calculated on the basis of the centrality of its neighbors and all other nodes that are connected to node v via the neighbors. It is a generalization of the eigenvector centrality. The Katz centrality for the node v_i is the ith element of the vector x given by the following equation (Newmann (2010)):

$$x = \alpha A x + \beta \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix},$$

where α and β are positive constants and A is the adjacency matrix of the graph G with the eigenvalues λ. The parameter β controls the initial centrality and thus enables an additional weighting of the immediate neighbors. Connections with distant neighbors are weighted by a damping factor α, which should be strictly smaller than the inverse largest eigenvalue of the adjacency matrix so that the Katz centrality can be calculated correctly:

$$\alpha \leq \frac{1}{\lambda_{max}}.$$

In the NetworkX library (Hagberg et al. (2008)), the values $\alpha = 0$ and $\beta = 1$ are used by default. These are also used for the present work.

Note that Katz centrality is a variant of eigenvector centrality. We use this metric to proof internal consistency between eigenvector and Katz centrality. The next measure, reciprocity, is a measure of the likelihood of vertices in a directed network to be mutually linked.

Definition 2.12. *Reciprocity*
The *reciprocity* of a directed graph is defined as the ratio of the number of edges pointing in both directions to the total number of edges in the graph:

$$rec(G) = \frac{|E^{\leftrightarrow}|}{|E|},$$

where $|E^{\leftrightarrow}|$ is the number of edges pointing in both directions in the graph G. For two nodes $u, v \in V$ applies $(u, v) \in E \Rightarrow (v, u) \in E^{\leftrightarrow}$. Edges that have two directions are therefore included twice in the calculation of $|E^{\leftrightarrow}|$.

In the context of ET networks, a high reciprocity value means that many transitions that exist between the nodes in the gaze data networks point in both directions. Following, we define the squares clustering coefficient.

Definition 2.13. *Squares Clustering Coefficient*
The *squares clustering coefficient* for a node v calculates the probability that two neighbors of node v have a common neighbor that is different from node v. For each node, it calculates the proportion of possible squares that exist at this node:

$$cc_4(v) = \frac{\sum_{u=1}^{deg(v)} \sum_{w=u=1}^{deg(v)} q_v(u,w)}{\sum_{u=1}^{deg(v)} \sum_{w=u+1}^{deg(v)} [a_v(u,w) + q_v(u,w)]},$$

where $q_v(u,w)$ is the number of common neighbors of the nodes u and w that are are not the node v (and thus result in a square), and $a_v(u,w) = (deg(u) - (1 + q_v(u,w) + \theta_{uv})) + (deg(w) - (1 + q_v(u,w) + \theta_{uv}))$, where $\theta_{uv} = 1$, if the nodes u and w are adjacent to each other, else $\theta_{uv} = 0$.

Note, that sqaures clustering coefficient is similar to clustering coefficient. We use this metric to show internal consistency among those two metrics. The s-metric is a network metric that measures how interconnected the nodes are in the network.

Definition 2.14. *s-Metric*
The *s-Metric* of a Graph $G = (V,E)$ is defined as the sum of the products of the degrees of the nodes u and v for every edge u, v in E:

$$metric_s(G) = \sum_{(u,v)\in E} deg(u) \cdot deg(v).$$

High values of s-metric occur in gaze data with many transitions. In the following, we define transitivity, a network measure that measures the extent to which the relation that relates two nodes in a network that are connected by an edge is transitive.

Definition 2.15. *Transitivity*
The *transitivity* of a graph $G = (V,E)$ is the proportion of all possible triangles that are present in G:

$$tra(G) = 3 \cdot \frac{T(G)}{Triads(G)},$$

where $T(g)$ is the number of triangles in graph G and $Triads(G)$ denotes the number of pairs of edges in E with a common node.

On this theoretical basis, we now develop the research questions.

Research Questions

Replication Study

To gain more insight into linear relationships in mathematics and kinematics, Ceuppens et al. (2019) considered the following research questions. (We are investigating the replicability of their results on German 9th graders and are therefore asking the same questions.)

RQ1 Are the difficulties diagnosed by Ceuppens et al. (2019) replicable for German students?

RQ1.1 In particular, what differences are there between Belgian and German 9th graders regarding performance and solution strategies?

We are also interested in whether the difficulties identified continue into higher grades:

RQ2 Do the difficulties diagnosed by Ceuppens et al. (2019) also occur in students of the entry phase of the upper secondary level?

RQ2.1 In particular, how do the results from RQ1.1 differ if we examine German 11th graders instead of German (and Belgian) 9th graders?

We answer these research questions in Section 3.1. To gain deeper insights into cognitive processes when solving graph and formula tasks on linear functions, we conducted an ET study.

Eye-Tracking Study

Graphs:

Ceuppens et al. (2019) showed that in similar item pairs in the kinematic context, errors in the interpretation of graphs with negative slope were made more frequently by students than in the mathematical context. However, the question of the cause remains open; we suspect that different solution strategies are applied depending on the context. If this is the case, then these differences should also show up in the gaze data.

RQ3 Given a similar pair of items for interpreting graphs with negative slopes, do eye movements differ significantly between the kinematic and mathematical contexts?

If significant differences in eye movements can be found, gaze cued RTA interviews can be used to determine whether the differences can indeed be attributed to different strategies. This leads to the following research question.

RQ4 Can differences in eye movements be attributed to different strategies in mathematical and kinematic contexts using RTA interviews?

Formulas:
Regarding formulas, we are first interested in whether we can resolve cognitive processes using gaze data. Second, we want to know if we can find contextual differences in gaze data of mathematical and physical formulas, and third, we want to know something about the reading direction of formulas:

RQ5 To what extent can cognitive processes in the interpretation of equations be resolved using ET?

RQ6 What are the contextual differences in gaze data between the same formal representations in mathematics and physics?

RQ7 What influence does the reading direction have on the cognitive processing of equations?

We answer research questions 3 to 7 in Chapter 3.

Network Analysis of Eye-Tracking Data

In the area of network analysis, we will first explore how we can put ET data into the form of mathematical networks in order to apply mathematical structures to ET data. We are also interested in the connection between ET metrics and network-related measures. We will then test whether this mathematical model is suitable for detecting geometric structures in gaze data. Furthermore, we want to test if we can find discriminative network metrics that can distinguish between correct and incorrect solvers.

RQ8 How to model ET data as a network while maintaining sequential information of the data?

RQ9 How can this model be used to detect geometrical structures in gaze data?

RQ10 Are the methods of network analysis suitable for differentiating gaze data from correctly and incorrectly solving persons with regard to item difficulty, context and clustering?

RQ11 If there are differentiating network metrics, which network metrics are particularly suitable for differentiating between correctly and incorrectly solving persons?

These research questions are answered in Chapter 4 of this thesis. In the next section, we will introduce study design and statistical methods that are used in this thesis.

2.2. Study Design and Statistical Methods

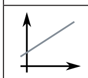
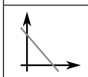
Study Design

In this section, we will explain the study design for the replication study and for the ET study.

Sample

ET study: A total of N=131 upper secondary students (74 male, 56 female, 1 not specified, all with correct to normal vision) from four secondary schools in Germany participated in the study. Secondary school here means a school that provides higher education and constitutes the highest level of the educational system in Germany. Students of this type of school are comparable to U.S. high school students attending college preparatory classes. Students used their free periods or, with teacher permission, regular classes to participate. All students took part voluntarily and were partially rewarded with a 5€-voucher for their participation. At the time the study was conducted, kinematics was already covered in all courses. 60 participants (37 male, 23 female) also agreed to be interviewed about their gaze data.
Replication study: In total, 249 students (108 male, 102 female, 5 divers, 34 not specified, age: 14.71 ± 0.58(SD)) from five different schools in Rhineland-Palatinate (3), Saxony-Anhalt (1) and North Rhine-Westphalia (1) participated in the study. All students attend the 9th grade of a secondary school. Physics was taught to 201 students in year 9. Note that in Germany, physics is not taught continuously in middle school (grades 7-10). It is possible that physics is offered in grades 8 and 10, but not in grade 9. This sample differs only slightly from the sample of the study by Ceuppens et al. (2019). Ceuppens et al. (2019) examined 253 students from seven different schools in Belgium, 117 of whom were male and 136 female.
We also carried out the test with 11th graders. Here, 298 students took part in the test (135 male, 132 female, 8 divers, 23 not specified, age: 16.51 ± 0.73(SD)). Students from grade eleven came from eight different schools in Rhineland-Palatinate (4), Saxony-Anhalt (2), and North Rhine-Westphalia (2). A total of 190 (grade 9) and 252 (grade 11) stated German as their mother tongue. Participation in the study was voluntary. According to the curriculum, the test content in mathematics and kinematics was covered in all courses until the study was conducted. It is striking that this differs from the information provided by the students. According to the students, the concepts of slope and y-intercept in mathematics and speed in kinematics are largely known by more than 80% of the students (slope: 80.7% (9th graders), 92% (11th graders), y-intercept: 81.5% (9th graders), 88.3% (11th graders) and speed: 77.9% (9th graders), 82.9% (11th graders). It is striking that only 18.9% (9th graders) and 48% (11th graders) were familiar with the

Table 2.2.: Test structure of the test instrument by Ceuppens et al. (2019): Two signs for the slope (positive and negative), two contexts (mathematics (M) and kinematics (K)), two concepts (slope, y intercept), two representations (graph, formula) and three question types using two tasks (compare, determine). Items marked in gray require an explanation.

	Compare ..., given a graph		Determine ..., given a graph		Determine ..., given a formula	
	slope	y-intercept	slope	y-intercept	slope	y-intercept
	M K	M K	M K	M K	M K	M K
	M K	M K	M K	M K	M K	M K

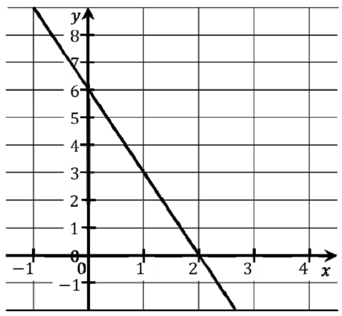
concept of starting position in kinematics. Both levels show average effort in the medium. 9th graders put in slightly more effort on average than 11th graders (9th graders' effort: 0.59 ± 0.26 (SD), 11th graders' effort: 0.48 ± 0.29(SD), effort was measured on a four-point likert scale between 0 and 1: 0 (not at all); 0.33 (little); 0.67 (rather); 1 (very)).

Material

The 24 items of the test instrument are taken from the validated test by Ceuppens et al. (2019), which we translated into German (cf. Appendix A.1). The test instrument comprises of 12 similar item pairs, each consisting of one item in the mathematics context and one item in the kinematic context. The tasks in the test instrument are structured (cf. Table 2.2). There are three different question types: "compare ... given a graph", "determine ... given a graph" and "determine ... given a formula". There are two different types of function (strictly monotonically decreasing and strictly monotonically increasing), two representations (graph, formula), two concepts (y-intercept and slope) and two contexts (kinematics (K) and mathematics (M)). The items were presented in random order (for ET study on a computer screen, for replication study on paper), starting once with the mathematics items and once with the physics items to avoid sequential effects. In the context of this work, we focus on the RTA protocol and ET analysis on two item pairs, where the students determined the slope of a linearly decreasing function, once for a mathematical function graph (M8) or a function equation (M12) and once for a time-position-function graph (K8) or a time-position function equation (K12, cf. Figures 2.4, 2.5). The replication study considers the whole test instrument. It should be noted that, in contrast to the ET study, written explanations were given for some of the tasks in the replication study.

M8

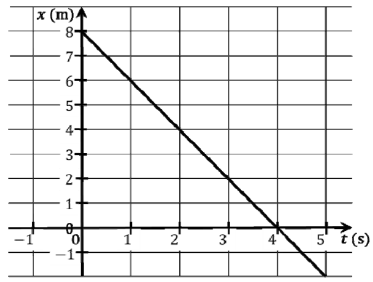
The figure shows the graph a function .



Determine the slope. Explain your answer.

K8

A car is moving along a straight road. The graph shows the position of the car as function of the time . The position is in meter, the time in seconds.



Determine the velocity of the car. Explain your answer.

Figure 2.4.: Similar item pair no. 8 (IP8) from the original test instrument; on the left: kinematic context (K8), on the right: mathematical context (M8).

K12

A car is moving along a straight road. The position x of the car as function of the time t is given by

$$x = 3 - 12t.$$

The position is in meter, the time in seconds.

Determine the velocity of the car. Explain your answer.

M12

A function f is given by

$$f(x) = -6x + 12\,.$$

Determine the slope. Explain your answer.

Figure 2.5.: Similar item pair no. 12 (IP12) from the original test instrument; on the left: kinematic context (K12), on the right: mathematical context (M12).

Procedure

ET study: The study occurred in the school's rooms where two identical ET systems were set up. First, the participants were informed about how the Tobii X3-120 remote ET system works (approx. 3 min). Then, they answered a short questionnaire about their demographics (approx. 2 min). After that, a 9-point calibration was performed on each ET station to obtain a full-customized and accurate gaze point calibration (approx. 3 min without recalibration). Subsequently, the 24 items were shown on the computer screen (1920 x 1080; refresh rate 75 Hz). Each item contained the task followed by a graph or a formula. Some items included answer options, while others included free response formats. After students were ready to give their answer, they pressed a key to advance to the next slide. Here, they were either asked to click on an answer option or to write down their answer on a sheet of paper in front of them. Then, by pressing the key, they were asked on the next slide how confident they were about the correctness of the answer they had just given (4-point Likert-type rating scale, ranging from very high confidence to guessed). The students could take as much time as needed to answer a question. They did not receive feedback after completing a task and could not return to previous tasks. After completing all 24 tasks on the computer screen (median: 19 min 21 s, max.: 40 min 15 s, min.: 11 min 18 s), students could voluntarily participate in gaze cued RTA interviews. Here, the students are shown their own gaze video of a single task. Their voice was recorded during this process. The gaze cued RTA started with a short instruction containing information about the gaze video, the intentions of the data collection, and data protection (approx. 1 min 30 s). Then, the interviewer asked: "Try to describe your solution. It is important that you orientate yourself on the recorded gaze data that you see. You can press the pause button at any time to stop the video. You can also continue speaking during the pauses. Do you have any questions before we start?" The Interwiever was only allowed to use confirmation signs like "aha," "I see," "ok," or reminders to think aloud, such as "Please keep talking," "Please tell me what you did here," to the student's audio (recording duration per item: median: 37 s, max.: 2 min 7 s, min.: 17 s). *Replication study:* The test was conducted as a classic paper-pencil test in order to give the students as much freedom as possible in explaining and creating the answers. Before completing the test, the students were informed that they had one lesson (45 minutes) to complete the test and that they had many options in the explanation fields: explaining with text, table, formula, graph. If they do not complete the task, they should leave the answer and explanation fields blank.

Data Collection and Analysis

Performance Data

The coding of the answers is dichotomous, i.e., 0 for a wrong answer and 1 for a correct answer. Using the confidence rating, guessed answers were marked as incorrect answers. The item difficulty P (proportion of participants who solved the item correctly) was calculated to analyze the difficulty level of each item. For the study including explanations, we rated answers as follows (cf. Ceuppens et al. (2019)): Answers were marked as correct if both the answer and the explanation (when required) are correct or if the answer is correct and the explanation is missing. An answer is marked as incorrect, if the answer is correct and the explanation is not. GEE were used to investigate the dependence of response correctness on covariates such as context, concept, sign of the slope and task type (cf. Section 2.2). The *geeglm* function in the statistical software R was used for the analysis. The statistical model included the covariates context (1D mathematics or kinematics), concept (y-intercept, slope), QType (question type: compare given a graph, determine given a graph, determine given a formula), concept (mathematics (m), kinematics (k)), slope sign (+,-) and gender (female, male, divers). The dependency of the responses of one and the same person was considered in the model using the correlation structure "exchangeable". In comparison to other correlation structures, this one had the lowest QIC value (cf. Section 2.2). With the correlation structure "exchangeable", all of a person's answers are correlated equally in pairs.

Eye-Tracking Data

The data collection and the definition of AOIs was done with the software *Tobii Studio*. The individual viewpoints were recorded at a frequency of 120 Hz. Based on this raw data, fixations and saccades were determined using the default I-VT (Identification by Velocity Threshold) algorithm built into the Tobii software (threshold: 30°/s for the velocity, Salvucci, Goldberg (2000)). To directly compare eye movements on an individual level for one and the same subject, gaze plots were generated using the Tobii software. In such plots, the fixations for each subject are shown as circles; the larger the circle, the longer the corresponding fixation duration. Gaze jumps (saccades) between fixations are represented as straight connecting lines.

For the gaze data analysis of item pair no. 8, the items were restricted to the diagram area only, which was systematically segmented into AOIs in such a way that structures relevant for extracting information are covered by only one AOI at a time. For example, areas of the axis intersection points or areas containing the axis labels are summarized in AOIs. However, it should be noted, that not all AOIs could be kept equal in terms of their area in both items, since the items are not congruent. For example, AOI 5 in the mathematical item covers a larger area of the diagram than in the kinematic item (see Figure 2.6). However, we considered the comparability of the visual information content of the area to be more important than the area when positioning the AOIs. For future developments of test instruments, nevertheless, attention should be paid to an area coverage as equal as possible.

For the gaze data analysis of item pair no. 12, the items were restricted only to the formula area, which was divided into AOIs in two different ways. In the first case (content-specific), the formula area was subdivided in such a way that the structures relevant for information extraction were only covered by one AOI at a time (cf. Figure 2.7 a)). For example, areas of the function name (AOI 1) or areas containing the coefficients (AOIs 4 and 6) are summarized in AOIs. It should be noted that not all corresponding AOIs in both items have the same size, since the items are not congruent. For example, AOI 1 in the mathematical item covers a larger area of the formula than in the kinematic item (see Figure 2.7 a)). However, in this step, we considered the comparability of the visual information content of the area to be more important than the area size when positioning the AOIs. In the second case (position-specific), the formula area was covered with equidistant AOIs in order to investigate position effects (cf. Figure 2.7 b)).The size of the AOI in item M12 corresponds to 43×101 px and in item K12 to 92×94 px. Both AOI sizes are above the spatial resolution limit of the eye tracker.

The ET metrics TVD (selection/organization) and transitions (integration) were considered (cf. Section 2.1). A non-parametric Wilcoxon signed rank test

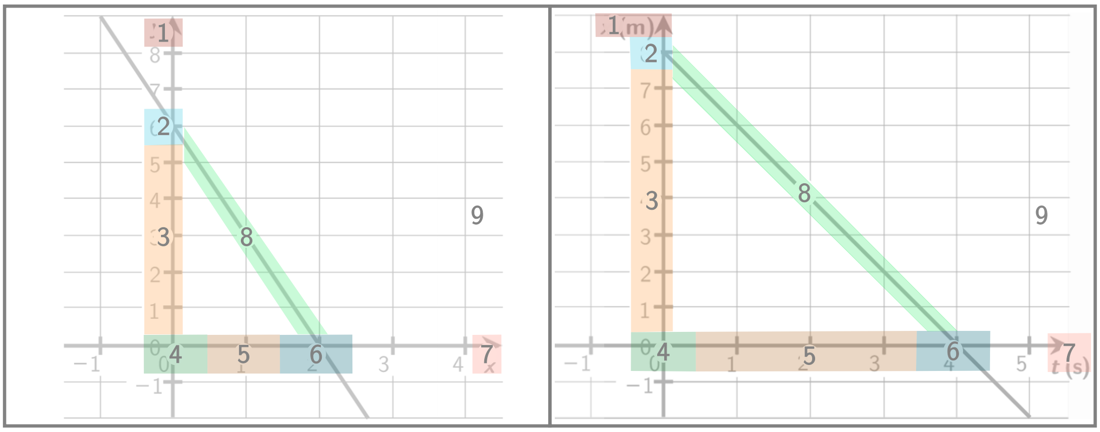


Figure 2.6.: Positioning of the AOIs for the mathematics item M8 (left) and the kinematics item K8 (right).

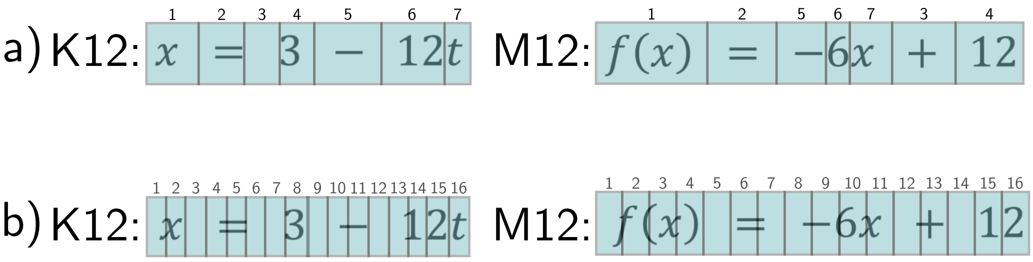


Figure 2.7.: Positioning of the AOIs in two different ways (content-specific (a), and position-specific (b)) for the kinematics item K12 (left) and the mathematics item M12 (right).

was used to test whether the central tendencies of the TVD / number of transitions of the participants were different. The analyzed datasets met the assumptions required to perform the Wilcoxon test. A threshold of $p = 0.05$ was used to determine the effect significance level within all tests performed. To control the false discovery rate due to multiple testing, the p-values were corrected using the Benjamini-Hochberg procedure (cf. Section 2.2). The effect sizes r with 95% confidence interval (calculated using Bootstrapping with 1000 replications) were determined for all Wilcoxon tests with significant results and can be interpreted after Cohen's guidelines (Fritz et al. (2012) and Cohen (1988)).

Gaze-Cued Retrospective Think Aloud Data

For collection of gaze cued RTA data (cf. Section 2.1), students were asked to describe their solution strategy for negative-slope tasks (item pair no. 8 and 12) in a short interview while watching their own eye movements on the task in a video. They could pause and resume the gaze video by pressing a key at any time to explain the eye movements in more detail. The gaze videos of the two item pairs were shown in arbitrary order to avoid learning effects depending on specific item orders. Interviewers were not allowed to comment on the explanations or give hints about the correct solution. Finally, the data collected are videos for each task (M8 and K8/ M12 and K12) separately. These videos show the student's eye movements on the item, which include the student's explanation as an audio track. For the data analysis, the interviews were first transcribed. The RTA transcripts were transcribed with a focus on the content of what was said. This means that all protocols were transcribed as accurately as possible. Linguistic peculiarities such as dialect and abbreviations were retained. Pauses, tenses, comprehension signals (such as "ah", "mhm", "aha"), background noise, laughter and hesitation were not transcribed, since they represent no concret content, need further interpretation and therefore can not be analyzed using content analysis methods (cf. Section 2.2). After the transcription, the interviews where analyzed using qualitative content analysis according to Mayring (2014). For this purpose, each interview was coded by two independent trained individuals according to a triplicate revised coding guide for item pair no. 8 and a twice revised coding guide for item pair no. 12 using the software MAXQDA (meaning the categories were revised in a triple/twice feedback loop and reduced to main categories and checked for reliability). For item pair no. 8, the coding guide contains inductive codes that capture solution strategies of the interviews of the participants. For example, the code *slope triangle* was coded in the interview if the participant expressed that he or she solved the task by forming a slope triangle and/or used the strategy of forming a slope triangle, i.e., the procedure *Then I went one to the right and two down and thus determined the slope* was explained. If a participant indicated in the interview that he or she formed the quotient of distance and time or of the axis intercepts of the linear function to solve the task, the strategy *quotient of axis intercepts* was coded. Solution strategies without counterpart in the respective other context were combined into the code *other*. *No strategy* was coded if participants guessed the answer or if no identifiable solution strategy was used.

For item pair no. 12, the coding guide contains inductive codes and one deductive code (*slope-height-confusion*) that capture solution strategies of the interviews of the participants. For example, the solution strategy *read from formula* was coded in the interview if the participant expressed that she or he has solved the task by reading the slope value from the equation, i.e., the procedure "the slope, it was minus 6 because mx plus b and then m was minus 6 in that case, hence the slope is minus 6." was explained. The strategy *equation rearrangement* was coded when students said that they rearranged the equation to solve the task. If a participant indicated in the interview that he or she formed the quotient, product or sum of the coefficients of the linear function equation to solve the task, the strategy *linking of coefficients* was coded. The category *slope-height-confusion* was coded when the student confused the slope / velocity with the y-axis intercept / initial position i.e. , "[...] that was simply mx plus b and then simply b in the case of twelve" was explained. Solution strategies without counterpart in the respective other context or strategies that have been coded less than four times each in M12 and K12 were combined into the code *other*. *No strategy* was coded if participants mentioned that they guessed the answer or if no identifiable solution strategy was used. Exactly one solution strategy was coded for each interview of a task.

Qualitative and quantitative statistical methods used for data analysis are presented in the next section.

Statistical Methods

First, this section gives an overview of the methods used for qualitative analysis of interview data. We will introduce the qualitative content analysis (QCA) of Mayring (2019) as a suitable, systematic, and rule-based method to analyze large amounts of interview data. Data are analyzed by at least two coders. The most important measure of coding quality is the measure of interrater reliability. Since we follow different coding strategies in this paper, we will present three suitable measures, namely the one according to Cohen (1960), Brennan, Prediger (1981), and an extension according to Kraemer (1980).

Second, the statistical analysis in this thesis requires numerous quantitative statistical methods, which are presented next in this section. To replicate the study of Ceuppens et al. (2019), we need the basics of classical test theory. Generalized estimating equations (GEE) are necessary for further analysis of the clustered data of the replication study. A one-factor analysis of variance (ANOVA) is used to test for significant differences between the different clusters (concept, context, ...). The Wilcoxon test for dependent and independent samples is relevant for investigating differences in contextually different gaze data and for distinguishing between correct and incorrect solvers. All these basics are presented in this section.

Introduction of Qualitative Content Analysis and Distinctions to Similar Concepts

According to Mayring (2019), QCA is a systematic and rule-based method suitable for large amounts of text. Mayring's method of analysis makes it possible to perform qualitative text analysis tasks (summaries, structuring of text material) that also allow quantitative analyses (Mayring (2019)). In contrast to other qualitative content analysis methods, Mayring's QCA is based on pre-defined steps of interpretation (Mayring (2014)). This makes the method transparent, reproducible, and comprehensible for others (Mayring (2014)). It is the most common method in social sciences and psychology (Titscher, Jenner (2000) and Carrera-Fernandez et al. (2014)). In summary, qualitative content analysis according to Mayring (2014) describes various different methods based on clear step-by-step models and analytical rules.

In protocol analysis for evaluating protocols of TA, the material is sequenced on the basis of explanations, descriptions, justifications and rationalizations, but the rules of interpretation are unclear (Ericsson (2017)). Codebook analysis is a content analysis with predefined categories. It is not as precise as a coding guide (Neuendorf (2002)). Qualitative content analysis according to Kukartz is only suitable for smaller amounts of text (Mayring (2019)). QCA according to Schreier (2012) disallows theory-based definitions of categories and is therefore only a slimmed-down version of QCA by Mayring (2014).

To summarize, we are convinced that QCA according to Mayring (2014) is the most suitable method for analyzing the interview data we collected. In a next step, we will describe the procedure of QCA in more detail.

Qualitative Content Analysis: Procedure For Evaluation

The QCA procedure according to Mayring (2014) can be described as a step-by-step model as shown in Figure 2.8. The individual steps of the process model are explained and exemplified using the QCA carried out as part of this work.

The first step is to determine the material on which the analysis is based. It should be noted that the material should not be subsequently expanded or changed unless it is absolutely necessary. The materials of our analyzes are transcribed explanations of students solving the paper-based test of Ceuppens et al. (2019) or transcribed interviews of all learners who participated in the ET-study. In the interview, students were asked to describe their solution strategy while viewing their own gaze data of a previously solved task. The material of the ET studies can be accessed via Becker, Knippertz, et al. (2022) and Knippertz et al. (2024). The material of the paper-based study can be accessed via Knippertz (2024a).

The second step is to describe exactly where, by whom and under what conditions the material was created. Mayring (2014) mentions the following points:

- the author of the material and/or the parties involved in its production;
- the emotional, cognitive and motivational background of the author(s);
- the target group for which the material is intended;
- the concrete circumstances of origin;
- the socio-cultural background.

In our case, three assistant researchers and one more researcher besides myself were involved in the production of the interviews. The transcription of the interview data into text format was done by an assistant researcher and by myself for

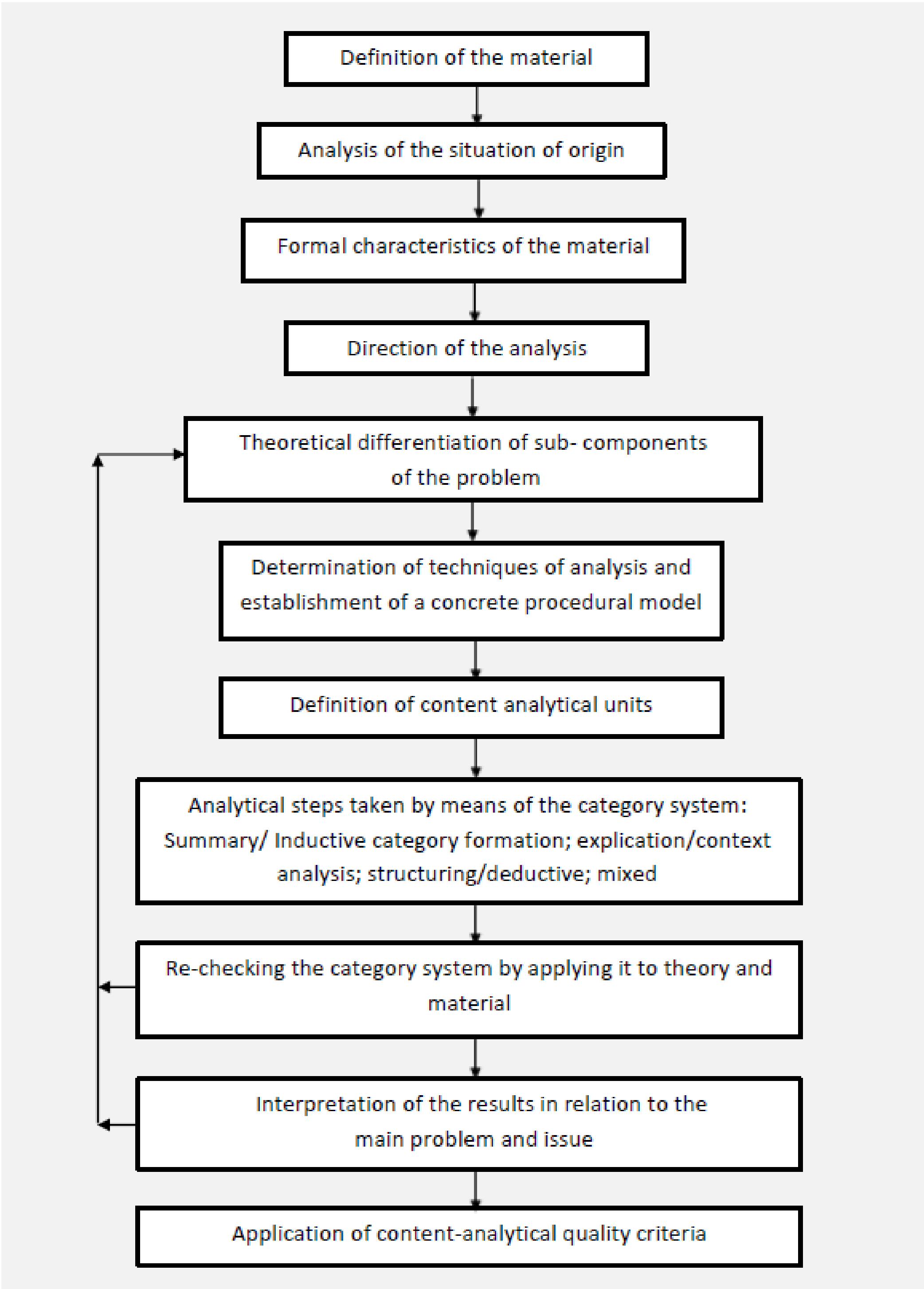


Figure 2.8.: General content-analytical procedural model. From Mayring (2014, p.54)

the ET-studies. For the paper-based replication study, three assistant researchers coded the original written explanations. All researchers were motivated and emotionally and cognitively prepared. The target group of the ET-study consisted of 131 students from four different schools in Rhineland-Palatinate and 249 ninth graders and 298 eleventh graders for the replication study (see Section 2.2).

The next step is to characterize the material in terms of formal criteria. Is the material based on a written text or on audio files, for example? In the case of a transcription, all the rules according to which the transcription was made must be recorded. If applicable, the instructions by which the interview was conducted must also be listed. For our purposes, we transcribed audio files according to the following rules: Transcribe the content as faithfully as possible, preserving linguistic idiosyncrasies (abbreviations, dialect). Pauses, tenses, signals of understanding (ah, mhm, aha), background noise, laughter, and hesitation are not transcribed. For instructions on how to conduct the interviews, see Section 2.2.

The next step is to determine the direction of the analysis. What is the goal of the analysis? Our goal is to make statements about the learners' solution strategies and explanations.

Then, a theory-oriented differentiation of the problem has to be conducted. This means that content analysis is characterized by two features: a rule-based approach and the fact that the analysis follows a precise and theoretically grounded content-related question. This means that the research question based on the analysis has to be well defined in advance and embedded in the context of current research (cf. Sections 2.1, 2.1). This can be followed by a theoretical differentiation of subproblems. In our case, we were only interested in one question: What solution strategies were used to solve the problem? And for the replication study: What explanations did the students use?

The next four steps in Figure 2.8 deal with the development, validation, and refinement of the coding guide to categorize the data. For a detailed description of all steps of inductive category development, see Mayring (2014, pp.80–83). In our case, for the ET-study, a coding guide was created for item pair no. 8 and 12 (IP 8 and IP 12)of the testinstrument by Ceuppens et al. (2019), which was revised three times for item pair no. 8 and twice for IP 12. The revisions always included a clarification of the codes or an explanation of the codes or anchor examples. Solution strategies based on the text material were chosen as codes inductively. Inductive here means that the codes were created from the text and not based on the literature (deductive). The creation of a comprehensive coding guide serves to make the coding process as transparent and reproducible as possible. For the replication study, a translated coding guide of Ceuppens et al. (2019) was adopted. This is followed by an evaluation and interpretation of the analysis and an assessment of the quality of the agreement between the coders (interrater reliability).

Interrater Reliability

Coefficients of rater agreement are statistical measures of the interrater reliability of assessments by (usually) two raters. In this section, we will describe three coefficients of rater agreement: Cohens κ, Brennan and Predigers κ_n and an extension of the kappa coefficient by Kraemer, κ_k. The contents and definitions of this section are based on Cohen (1960), Brennan, Prediger (1981), Kraemer (1980), and Eye (2006).

Results of rater agreement of two raters can be represented in a table as in Table 2.3. In Table 2.3, rater agreement of two raters A and B for I categories is listed.

To describe Cohens κ, we use all cells that indicate rater agreement, $p_{ii}, i = 1, \ldots, I, I \in \mathbb{N}$, to calculate the proportion of instances in which the two raters agree exactly:

$$\theta_1 = \sum_{i=1}^{I} p_{ii}.$$

I is the number of categories used by both raters. If we assume independence of the two raters, we can calculate the proportion of cases in which the two raters coincidentally match:

$$\theta_2 = \sum_{i=1}^{I} p_{i.}p_{.i},$$

where $i.$ indicates the ith row total and $.i$ indicates the ith column total. If θ_2 is subtracted from θ_1, a measure of the agreement between the raters beyond chance is obtained. If the difference is positive, the two raters agree more often than would be expected by chance. If this difference is negative, the raters agree less often than would be expected by chance. The largest possible discrepancy between θ_1 and θ_2 is $1 - \theta_2$. This is the case in which all observed ratings appear in the diagonal cells of the cross-classification of the judgments of the two raters. Weighting the difference by $1/(1 - \theta_2)$ gives

Definition 2.16. *Cohen's Measure of Rater Agreement*
Cohens measure of rater agreement is defined as

$$\kappa = \frac{\theta_1 - \theta_2}{1 - \theta_2},$$

with θ_1 being the proportion of instances in which the two raters agree exactly and θ_2 is the proportion of instances in which the two raters agree by chance.

Cohens κ thus indicates the proportion of cases in which the two raters A and B select the same categories to rate a series of objects and this proportion is corrected for chance. It has several characteristics (cf. Eye (2006)):

Table 2.3.: Table of rater agreement of two raters A and B among I categories used by both raters. Cell p_{12} states the probability of rater A choosing category 1 and rater B choosing category 2.

A/B	Cat_1	Cat_2	...	Cat_I
Cat_1	p_{11}	p_{12}	p_{1i}	p_{1I}
Cat_2	p_{21}	p_{22}	p_{2i}	p_{2I}
...	p_{i1}	p_{i2}	p_{ii}	p_{iI}
Cat_I	p_{I1}	p_{I2}	p_{Ii}	p_{II}

1. Coefficient κ has range $\frac{-\theta_2}{1-\theta_2} < \kappa \leq 1$. If κ is greater than zero, there is an agreement better than chance. A negative κ indicates agreement less than chance.

2. Coefficient κ is zero if the probability of agreement is as expected; this can be the case even if the raters' judgments are not independent.

3. Coefficient κ is eqaul to 1 only if the probability of disagreement is zero.

4. For κ to be defined, at least two categories must be used by both raters, that is, the probability p_{ij} must be greater than zero for at least two cells.

5. If the probability in at least one of the off-diagonal cells is non-zero, the maximum value of κ decreases as the marginals deviate from a uniform distribution.

6. When the probability of disagreement decreases and is smaller than the number of agreements, κ increases monotonically; when the probability of disagreement increases and is greater than the probability of agreement, κ does not decrease monotonically.

There exist several guidelines on how to interpret κ. Those guidelines differ and that is why it is always advisable to report the value of κ alongside the interpretation. One possible and common interpretation is from Landis, Koch (1977) (see Table 2.4).

Many features of Cohen's κ have been criticized, especially item five in the list above (Eye (2006)). For this reason, there are several alternatives to Cohen's κ. In the following, we will present the coefficient κ of Brennan, Prediger (1981), κ_n: To solve the problem that the maximum agreement score is reduced when raters use the categories at different rates, Brennan, Prediger (1981) proposed to replace the formulation used by Cohen with $\theta_2 = 1/I$, where I is the number of rating categories. This substitution results in

Table 2.4.: Interpretation of coefficient κ after Landis, Koch (1977).

κ	Rater Agreement
$\kappa < 0$	poor agreement
$0 \leq \kappa \leq 0.2$	slight agreement
$0.21 \leq \kappa \leq 0.4$	fair agreement
$0.41 \leq \kappa \leq 0.6$	moderate agreement
$0.61 \leq \kappa \leq 0.8$	substantial agreement
$0.81 \leq \kappa \leq 1$	(almost) perfect agreement

Definition 2.17. *Brennan and Prediger's Measure of Rater Agreement* *Brennan, Prediger (1981) measure of rater agreement* is defined as

$$\kappa_n = \frac{\theta_1 - \frac{1}{I}}{1 - \frac{1}{I}},$$

where I denotes the number of rating categories. Brennan and Prediger's coefficient κ_n is equivalent to Cohen's κ when the expected cell frequencies are estimated under a no-effect log-linear model, or when the marginal distributions are uniform for both raters.

Both Cohen's κ and Brennan and Prediger's κ_n may only be used if the following conditions are met (Cohen (1960)):

1. The units (material that is evaluated, e.g., psychological test protocols) are independent.

2. The categories of the nominal scale are independent, mutually exclusive, and exhaustive.

3. The raters operate independently.

Since assumption 2 is not fulfilled in the case of the replication study, since the categories are not disjunctive, we will introduce an extension of Cohen's κ that also holds for non-disjunctive categories, e.g., the case where the rater can choose several categories to rate an object. To understand the definition of the extension of the κ coefficient, we first need to understand Spearman's rank correlation.

Spearman's rank correlation examines the undirected linear correlation between two variables. "Undirected" means that there is no dependent and independent variable. Therefore, no causal statements can be made. The only requirement for calculating Spearman's rank correlation is ordinal scale data. The idea of rank correlation is based on the data ranking. That is, the measured values are represented by ranks and the actual test is performed on these ranks. The test is therefore based only on an order of the data, such as greater than or equal to. The absolute distances between the data are not considered. The ranks are constructed as follows (cf. Kraemer (1980)): We assume that there are K response categories. A single choice, A, imposes a rank of 1 on category A and $\frac{1}{2}(K+2)$, that is, the average of the ranks $2, 3, \ldots, K$, on the remaining $K-1$ response categories. A ranked response, AB, means that there is a rank of 1 on A, 2 on B, and $\frac{1}{2}(K+3)$ on the remaining $K-2$ responses. An equivocal response, A/B, imposes a rank of 1.5 on categories A and B and $\frac{1}{2}(K+3)$ on the remaining $K-2$ response categories. The following example explaines the calculation of ranks in more detail.

Example 2.18. *Calculation of ranks of Spearman's rank correlation*
We assume that two raters, S and T, evaluate a task by four categories: A, B, C and D. So the number of categories is $K = 4$. Rater S chooses categories A and B (equivocal), rater T choses categories B. The ranks for this sitation are calculated as follows. Rater S: rank of 1.5 on A and B, rank of $\frac{1}{2}(K+3) = \frac{1}{2}(4+3) = 3.5$ on categories C and D. Rater T: rank of 1 on B and rank $\frac{1}{2}(K+2) = \frac{1}{2}(4+2) = 3$ on categories A,C and D.

Now we can define Spearman's rank correlation coefficient:

Definition 2.19. *Spearman's Rank Correlation Coefficient*
Spearman's rank correlation coefficient ρ is defined as

$$\rho = 1 - \frac{6 \cdot \sum_{i=1} n(r_i - s_i)^2}{n^3 - n},$$

with r_i being the rank within variable X of the ith respondent, s_i being the rank within variable Y of the ith respondent and n denotes the number of respondents.

Note, that Definition 2.19 only holds if all n ranks are distinct integers. Otherwise, we have to use more complicated definitions.

On this basis, we define the extension of the κ-coefficient after Kraemer (1980):

Definition 2.20. *Extension of the κ-Coefficient after Kraemer (1980)*
Kraemer's extension of the κ-coefficient is defined as

$$\kappa_k = \frac{r_I - r_T}{1 - r_T},$$

where r_I is the average of $r_1, r_2, \ldots, r_N$ with $r_i, i = 1, 2 \ldots, N$ being the average Spearman rank correlation coefficient among the $\frac{1}{2}m_i(m_i - 1)$ pairs of observations of subject i. The number of observations of subject i is m_i. Then, r_I represents the unweighted average of the intrasubject correlation coefficients. The coefficient r_T is defined as the average Spearman rank correlation coefficient among all pairs of observations, in total $m(m-1)$ pairs, where $m = \sum_i m_i$.

The expected value of coefficient κ_0 under the null hypothesis is zero. It holds $\kappa_0 = 1$ if and only if there is absolute agreement among all observations of a single subject (i.e. $r_I = 1$) and also some heterogenity among subjects (i.e. $r_T \neq 1$). Notably, calculating the extended κ-coefficient, as defined in Definition 2.20, involves calculating the Spearman rank correlation coefficient for all possible combinations of scores, which requires significant computational effort.

In the case of ties (equal values), let R_{ij} be the average rank assigned to the response category j by the m_i observations of subject i. The tie correction is computed as follows: One computes $\sum(t^3 - t)$, where t is the number of categories with a tied rank and the sum is over all tied sets for each observation.

Example 2.21. *Calculation of Tie Correction*
We assume, that the observations (in rank order) of a subject were $(1.5, 1.5, 4, 4, 4, 6)$. Then we have two sets of ties, one of length $t = 2$ and one of length $t = 3$. The tie correction for this observation is $(2^3 - 2) + (3^2 - 3) = 30$.

We denote the average tie correction of the m_i observations on subject i by T_i. Now we define S_i^2 as the sampe variance of the quantities R_{ij}: $S_i^2 = \sum_i (R_{ij} - \bar{R})^2/(K-1)$, where $\bar{R} = \frac{1}{2}(K+1)$. With the coefficient of concordance for m_i rankings of subject i by Kendall (1962), $W_i = 12(K-1)S_i^2/(K^3 - K - T_i)$, we get

$$r_I = (m_i W_i - 1)/(m_i - 1).$$

Let $\bar{R}_{.i} = \sum_{i=1}^{N} R_{ij}/N$ be the overall average rank of category j and let $T_. = \sum_{i=1}^{N} m_i T_i / m_.$ be the weighted average of the T_i. Let S_T^2 be the sample variance of the quantities $\bar{R}_{.i}$. That is, $S_T^2 = \sum_i (\bar{R}_{.i} - \bar{R})^2/(K-1)$. The overall coefficient of concordance is then $W_T = 12(K-1)S_T^2/(K^3 - K - T)$, and consequently

$$r_T = (m_. W_T - 1)(m_. - 1).$$

In this case, we see the calculation of k_0 does not need the computation of a large number of rank correlation coefficients, which is a great advantage.

Note that unlike Cohen's κ or Brennan and Prediger's κ, Kraemer's extended κ-coefficient allows raters to assign not just one but multiple codes to an item. This is accomplished by considering ranks in the extended κ-coefficient. If rater 1 coded two codes, e.g., A and B, these two codes are assigned a rank of 1.5 each and the remaining codes are assigned the average of the other ranks. If rater 2 coded only A, it is assigned a rank of 1 and the remaining codes are assigned the average of the other ranks. The fact that coder 1 also said B can therefore still be displayed.

Now, we want to define the standard error for k_0:

Definition 2.22. *Standard Error for κ-Coefficient after Kraemer (1980)*
The *standard error for k_0* is estimated as

$$SE(\kappa_0) = S_r/\{\sqrt{N(1 - r_t)}\},$$

where S_r^2 is the variance of the r_i over subjects, $\sum_i (r_i - r_I)^2/(N-1)$.

Now we will introduce some quantitative statistical methods. Our goal is to investigate the extent to which students' response accuracy to linear function tasks depends on context, task type, representation, and other factors. We also want to investigate the extent to which differences in gaze data can be statistically measured when working on similar math and physics tasks.

Classical Test Theory

The goal of test and item analysis is to determine whether a test is reliable and discriminating. Reliable tests have similar results when administered twice or more (at different times). This means that the test conditions and participant performance are stable. A test is discriminating if the results clearly distinguish between those with robust knowledge and those without. We will now introduce the basic concepts of item and test analysis. This subchapter is based on Ding, Beichner (2009), Ceuppens et al. (2019), and Bortz, Schuster (2010).

Since we want to replicate results by Ceuppens et al. (2019), we will use the same measures of analysis to obtain our results. For item analysis, Ceuppens et al. (2019) uses item difficulty index P and discrimination index D.

Definition 2.23. *Item Difficulty Index*
The *item difficulty index* is defined as the proportion of correct responses:

$$P = \frac{N_1}{N},$$

where N_1 denotes the number of correct responses of an item and N is the total number of responses of this item.

Note, although this measure is called "item difficulty", it is a measure of the simplicity of an item. In practice, item difficulty levels in the middle range (0.3-0.9) are desirable after Doran (1980), since too high and too low values do not discriminate between students.

The item discrimination index measures how powerful an item is in distinguishing high-achieving students from lowachieving students:

Definition 2.24. *Item Discrimination Index*
The *item discrimination index* is defined as the difference in percentages of correct response to an item between the top quartile and the bottom quartile students:

$$D = \frac{N_H - N_L}{N/4},$$

with N_H, N_L being the numbers of students in the top (H) and bottom (L) quantile and N the total number of students.

A discriminative item has a high number of correct responses in the top and a low number in the bottom quartile. A item discrimination index of $D \geq 0.3$ is desirable (Doran (1980)). Higher values are better. Indicators of a low item discrimination index are a task that is too easy (high D) or too difficult (low D) or a task that shows comprehension problems with regard to the wording. Poorly worded questions may cause proficient students to overthink, negatively impacting their performance.

To evaluate an entire test, Ceuppens et al. (2019) use the measure of Ferguson's δ. This measure measures the discriminatory power of the entire test.

Definition 2.25. *Ferguson's δ*
Ferguson's Delta is defined as

$$\delta = \frac{(k+1)(n^2 - \sum_{i=1}^{k} f_i^2)}{kn^2}.$$

Here, k is the number of items in the test, N the number of students and f_i the frequency of a certain score $i, i = 1, \ldots, k$ (Ferguson (1949)). In our case, f_i is the number of people with i correct answers.

Ferguson's δ examines how broadly students' total scores are distributed across the possible range. A well discriminating test has a broad score distribution. Desirable is a Ferguson's δ greater than 0.90 (Kline (1986)).

Cronbach's α is determined to measure the reliability of a test. The α coefficient is used to estimate the reliability of the sum score formed from all test items. Here, all test items are considered as independent "tests" for the same trait; the reliability of the sum score (α) is calculated as the average parallel test reliability for all possible pairs of test items. To illustrate, think of the α-coefficient as follows: Suppose you want to measure a complex trait such as intelligence. You select ten variables from a population, all of which are suitable for measuring the characteristic. The α-coefficient asks about the generalizability of the linear combination formed from the ten variables, where all variables are weighted by 1. Cronbach's α is defined as follows:

Definition 2.26. *Cronbach's α*
Cronbach's α is defined as

$$\alpha = \frac{p}{p-1} \cdot \left(1 - \frac{\sum_i s_i^2}{s_{tot}^2}\right),$$

where p denotes the number of variables (items), s_i^2 is the variance in variable i and s_{tot}^2 is the variance of the observed total test score (linear combination).

Desirable values for α are values grater than 0.9 (excellent), greater than 0.8 (fairly high) or grater than 0.7 (acceptable) after Engelhardt (2009) and Mallery, George (2000).

We have now defined all the measures that are important for the analysis of the test and the test items. In the following, we will examine GEE, an analysis method for modeling longitudinal or cluster data, as we also find in Ceuppens et al. (2019).

Generalized Estimating Equations

Longitudinal studies examine changes over time, such as changes in growth, ageing, time profiles or the effects of covariates (Peter, Song (2007)). They attempt to describe the dependence of the outcome on the predictor variables as accurately as possible (Liang, Zeger (1986)). The data from longitudinal studies are repeated measurements, as measurements are made on one and the same person in several time periods. In addition, they are often correlated because measurements are made on one and the same person (Burton et al. (1998)). However, most statistical models and tests assume that the observations are independent (e. g. linear regression, t-test, Wilcoxon test). In longitudinal studies, it is essential to consider the within-subject correlation. If this is neglected, the results can be misleading (Arge (2016) and Burton et al. (1998)). Performing a statistical analysis for repeated measures that assumes all observations are independent, even though there is a positive intraclass correlation, leads to estimated standard errors that are too small (Burton et al. (1998)).

We use GEE to obtain information about the dependency of the response correctness of students on covariates such as context (math, kinematics), slope sign (+,-) or question type.

There are numerous statistical analysis methods for normally distributed variables. If the outcome variable is not normally distributed, there is a lack of joint multivariate distributions. In these cases, likelihood methods are not available or only available to a limited extent (Liang, Zeger (1986)).

With the help of generalized estimating methods and the quasi-likelihood approach, linear models for normally distributed data can be applied to non-normally distributed data. However, this application is only valid for single observations of a subject and not for repeated measurements. This extension is possible because the quasi-likelihood approach does not require a specific distribution. Only two assumptions must be met: linear relationship between marginal expected value and covariates and variance is a known function of the expected value (Arge (2016)).

In addition to the two assumptions mentioned above, it is also necessary to establish a correlation structure between the repeated measures for a subject. This correlation structure ensures consistent estimators of the coefficients of their variances. (Liang, Zeger (1986)).

We want to give a short overview of Quasi-likelihood estimator, before introducing GEE, since this estimator is used in GEE.

Quasi-likelihood estimator: The methodology of regression quasi-likelihood assumes two relationships: First, there is a relationship between mean response and covariates, and second, there is also a relationship between mean response and variance. This section is based on Arge (2016).

Let Y_i be the response variable for each subject $i = 1, \ldots, N$ and X_i be a $p \times 1$ vector of covariats. Let β be $p \times 1$ vector of regression parameters to be estimated. Now we define $\mu_i = E(Y_i|X_i)$ as the conditional expectation of Y_i and a function of covariats and regression parameters, so that $\mu_i = h(X_i^T\beta)$. The inverse of h is the link function. This function relates the mean response to the linear predictor $(X_i^T\beta)$. For quasi-likelihood, the variance of each Y_i (denoted as v_i) is a function of expectiation μ_i, so that $v_i = f(\mu_i)\cdot\phi$. Here, the scale parameter ϕ is a nuisance parameter. The quasi-likelihood estimator is the solution of the equations:

$$S_k(\beta) = \sum_{i=1}^{N} \frac{\mathrm{d}\mu_i}{\mathrm{d}\beta_k} v_i^{-1}/Y_i - \mu_i) = 0, k = 1, \ldots, p.$$

Estimators of regression parameters $\hat{\beta}$, are obtained by iteratively reweighted least squares method.

Marginal Models: This section is based on Fitzmaurice et al. (2012). GEE use a marginal model. Marginal models are the extension of generalized linear models to longitudinal data. They are used if one wants to make inference on population, not individuals. For this reason, within-subject correlation is regarded as a nuisance characteristic and regression parameters and within-subject correlation is modelled separately. This model is called *marginal* model since the model for the mean response Y depends only on the covariates of interest. It does not depend on subject specific random effects or previous responses of the subject.

Let's set up some notation for repeated measuremnts. We assume that N subjects are measured repeatedly. Let Y_{ij} denote the response variable for the ith subject on the jth measurement occasion. At time t_{ij}, Y_{ij} can be continuous, binary, ordinal or a count. We assume that Y_{ij} is observed at time t_{ij}. We assume the number of repeated measurements can be different for subjects and we assume that there are n_i repeated measurements for the ith subject.

The response variables for the ith subject can be grouped into an $n_i \times 1$ vector:

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \ i = 1, \ldots, N.$$

The vectors of responses, Y_i, assume to be independent of one another. Nevertheless, the repeated measures on the same subject are not assumed to be independent. Associated with each response at a given time point j, there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \; i = 1, \ldots, N, \; j = 1, \ldots, n_i.$$

Now we can group the vectors of covariats into an $n_i \times p$ matrix of covariates:

$$X_i = \begin{pmatrix} X_{i1}^T \\ X_{i2}^T \\ \vdots \\ X_{in_i}^T \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \ldots & X_{i1p} \\ X_{i21} & X_{i22} & \ldots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \ldots & X_{in_ip} \end{pmatrix}, \; i = 1, \ldots, N.$$

Here, the rows of X_i correspond to the covariates associated with the responses at the n_i different measurement occasions, and the columns of X_i correspond to the p distinct covariates.

To obtain GEE, we need to set three specifications for a marginal model: the mean response at each occasion, the variance of the response at each occasion, and the pairwise within-subject association among the responses. We will do this for binary responses since we have those in our replication study.

We assume that Y_{ij} is a binary response with values of 0 (denoting "failure" or "uncorrect answer") or 1 (denoting "success" or "correct answer"). We want to relate changes in $E(Y_{ij}|X_{ij}) = Pr(Y_{ij} = 1|X_{ij})$ to the covariates (context, concept, slope value, ...). With a binary response, we have a Bernoulli-distribution of Y_{ij}. We model the probability of success using a logit link function. For a Bernoulli random variable, the variance is a known function of the mean. So we get the following illustration of a marginal model for Y_{ij}:

1. The probability of success (mean of Y_{ij}) is related to the covariats by a logit link function,
$$log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \eta_{ij} = X_{ij}^T \beta.$$

2. The variance of each Y_{ij}, given the effects of the covariats, depends on the mean response. The variance function is a known function of the mean. The scale parameter θ is 1 in this case.

3. The within-subject association among the vector of repeated responses is assumed to have an unstructured pairwise log odds ratio pattern,
$$logOR(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha_{jk},$$

where

$$OR(Y_i, Y_k) = \frac{Pr(Y_j = 1, Y_k = 1)Pr(Y_j = 0, Y_k = 1)}{Pr(Y_j = 1, Y_k = 0)Pr(Y_j = 0, Y_k = 1)}.$$

The set of parameters α characterize the within-subject correlation and need to be estimated. The "working" covariance matrix for Y_i is given by

$$V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} / \theta.$$

The correlation matrices $R_i(\alpha)$ can be different for subjects. They are fully specified by α and α is the same for all subjects. The matrix A_i is a $n_i \times n_i$ diagonal matrix with $log(\frac{\mu_{ij}}{1-\mu_{ij}})$ on the diagonal. "Working" covariance matrix here means that we do not know the true correlation structure between repeated measurements and that we are not assuming we are specifying it correctly. We would like to get consistent estimates of regression parameters regardless of the chosen structure.

Correlation Structures: Correlation structures should describe dependencies in the process of determining the solution. Note that correlations of the measurement process are strictly separated from explanatory variables: For example, we want to explain correlations between item types. This is therefore an explanatory variable and not a correlation that should be considered in the measurement proces. The following common types of correlation structures are used to create the covariance matrix $R(\alpha)$ (cf., Liang, Zeger (1986) and Ford (2022)). We have added a short comment on the suitability for our study in each case:

- *Independence:* The independent correlation structure is simply the matrix with zeros on the off-diagonal positions. This structure assumes that all pairwise correlation coefficients are zero and there are ones on the diagonal:

$$Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1, \text{ if } j = k \\ 0, if j \neq k. \end{cases}$$

Within the framework of this correlation structure, it is assumed that all observations on the 24 items are independent. But this is not true, since observations/measurements for 24 items are not independent. Each student solves all 24 items and therefore observations on these items are dependent because they come from the same person. Furthermore, the performance of the students is not considered. Furthermore, the tasks are also not independent. Answers to tasks of a similar type correlate.

- *Unstructured:* The unstructured correlation structure assumes that all pairwise correlation coefficients are different:

$$Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1, & \text{if } j = k \\ \alpha_{jk}, & \text{if } j \neq k. \end{cases}$$

  This means that correlations between responses to all items are estimated. In our case, this is true, as each student is individual and responds individually, but tendencies such as differentiation between good and bad students are not possible here. An important point here is: If we were to use this correlation structure, we would already be including correlations that we actually want to explain, such as relationships between task correctness and task type, as correlations in the measurement process. We would be incorrectly including an explanatory variable as a correlation of the measurement process. This correlation structure is therefore inappropriate.

- *Exchangeable:* The exchangeable correlation structure assumes that all pairwise correlation coefficients are equal and therefore exchangeable:

$$Cor(Y_{ij}, Y_{ik}) = \begin{cases} 1, & \text{if } j = k \\ \alpha, & \text{if } j \neq k. \end{cases}$$

  In our case, this means that a high-performing student is more likely to get the answers right, and a low-performing student is more likely to get them wrong. However, this is not quite correct. Interchangeable means that every answer is the same as every other answer, so you could dice them. However, this is doubtful because the problems are grouped together: A slope type problem cannot simply be exchanged for an intercept type problem. The relationship would not remain the same. However, since we want to explain group dependencies, they do not belong as a correlation in the measurement process. This structure is therefore appropriate. It provides a baseline of “good” and “bad” students: we expect that if we have a high-performing student, that student will give good answers. If we have a poor performing student, they should also give incorrect answers. Both baselines are equally correlated.

- *Autoregressive* The autoregressive correlation structure of order 1 assumes that the correlation coefficients decrease exponentially over time and that the responses are ordered in time and correlate more strongly when they are closer together in time than when they are further apart:

$$Cor(Y_{ij}, Y_{i,j+t}) = \alpha^t \text{ for } t = 0, \ldots, n_i - j.$$

In our case, this means that the further apart the answers are, the less correlated they are. This does not apply in our case, as questions were randomized. We have no chronological order in our data.

- *User-defined:* The user-defined correlation structure is a correlation structure provided by the user.

Model inference and interpretation of parameters: This part is based on Arge (2016). The link function influences the interpretation of regression coefficients. Parameter estimates are always interpreted in a population-average manner. Statistical significance tests and parameter estimates' confidence intervals are based upon Wald test. Here, regression coefficient estimates are divided by their robust standard error and the results are treated as normal standardized deviates (Z statistic) (see Burton et al. (1998)). It can underperform when the sample size is small.

Quasi-likelihood Information Criterion (QIC): The Quassi-likelihood information Criterion is used to chose the right model with appropriate correlation structure. For GEE models, QIC was implemented by Pan (2001). It has the form

$$QIC(R) = -2Q(\widehat{\beta}(R), I, \Delta) + tr(\widehat{\sum}^{-1}_{M(IN)} \widehat{\sum}_{S(R)}),$$

where $Q(\widehat{\beta}(R), I, \Delta)$ is the quasi-likelihood under the assumption of working correlation matrix R, $\widehat{\sum}^{-1}_{M(IN)}$ is the model-based covariance estimator under independence working correlation structure R and tr is the notation for matrix trace (Jang (2011)). The model with smaller QIC value is preferred.

In a next step, we want to statistically compare the mean values of test accuracies depending on influencing factors as slope sign, context, or concept.

One-Factorial Analysis of Variance

This part is based on Schwarz (2023). The one-factorial ANOVA tests whether the mean values of several independent groups (or samples) defined by a categorical independent variable differ. This categorical independent variable is referred to as a "factor" in the context of the analysis of variance. Accordingly, the characteristics of the independent variable are called "factor levels." An ANOVA is referred to as "single-factorial" if it uses only one factor, i.e., a grouping variable. We use the one-factorial ANOVA to compare the mean values of test accuracies depending on influencing factors as slope sign, context, or concept. The principle of variance analysis is the decomposition of the variance of the dependent variable. The total variance is made up of the "variance within the groups" and the "variance between the groups." These two proportions are compared with each other in an ANOVA.

The following conditions must be met for the application of a one-way ANOVA:

- The dependent variable is interval-scaled.
- The independent variable (factor) is categorical (nominal or ordinal scaled).
- The groups formed by the factor are independent.
- The dependent variable is normally distributed within each of the groups (from > 25 subjects per group, violations are usually unproblematic).
- Homogeneity of variances: The groups come from populations with approximately identical variances of the dependent variable (cf. Levene test).

If a first look at the group mean values shows differences, the ANOVA checks whether these differences are also significant. The basic idea is that each measured value can be broken down into three parts:

- Basic level: This is the mean value of the entire sample (regardless of group affiliation), the so-called "overall mean value."
- Effect of group affiliation: This proportion represents the group affiliation and corresponds to the difference between the overall mean value and the respective group mean value.
- Error value: This proportion includes measurement errors and all individual deviations from the overall mean value that cannot be explained by group membership.

An ANOVA now attempts to explain the deviations of the individual values from the overall mean. All individual deviations from the overall mean are squared and added up as a measure of the deviations to be explained for all individuals. This results in the total sum of squares SS_{total}:

$$SS_{total} = \sum_{g=1}^{G} \sum_{k=1}^{K_G} (y_{gk} - \bar{y})^2,$$

where g denotes the running index of the groups from 1 to G and k denotes the running index of individuals within a group from 1 to K_G. This sum of squares is broken down into two parts as part of the variance analysis: That which can be explained by group membership and that which cannot. The former is the sum of squares between the groups, $SS_{between}$, and is based on the difference between the group means and the overall mean:

$$SS_{between} = \sum_{g=1}^{G} K_g (\bar{y}_g - \bar{y})^2.$$

The second proportion is the sum of squares within the groups SS_{within} and is based on the individual deviations from the respective group mean:

$$SS_{within} = \sum_{g=1}^{G}\sum_{k=1}^{K_g}(y_{gk} - \bar{y}_g)^2.$$

The following relationship exists between these sums of squares:

$$SS_{total} = SS_{between} + SS_{within}.$$

The more variation can be explained by the group membership, i.e., by the model, the greater the variation between the groups ($SS_{between}$) and the smaller the variation within the groups (SS_{within}).

To calculate the test statistic F, the mean sums of squares $MS_{between}$ and MS_{within} are required. To do this, the sums of squares are divided by their respective degrees of freedom:

$$MS_{between} = \frac{SS_{between}}{df_{between}},\ df_{between} = G - 1,$$

$$MS_{within} = \frac{SS_{within}}{df_{within}},\ df_{within} = K_{total} - G,$$

where K_{total} denotes the sample size across all groups and G is the number of factor levels (groups). This results in the test statistic F_{model}:

$$F_{model} = \frac{MS_{between}}{MS_{within}}.$$

The more variation is explained by the group membership, the higher the F-value (as $MS_{between}$ is a measure of the explained variance, while MS_{within} is a measure of the residual variance of the model). This F-value is compared with the critical value on an F-distribution characterized by the degrees of freedom $df_{between}$ and df_{within}. If the F-value is higher than the critical value, the test is significant.

Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is a statistical test. In data-analysis, statistical tests are used to determine whether the results obtained are significant (cf., Bortz, Schuster (2010)). The Wilcoxon signed rank test is a nonparametric test, meaning that unlike parametric tests, the data do not need to be normally distributed to be reliably analyzed (Fritz et al. (2012) and Fahrmeir et al. (2016)). This is because this test uses the median as a measure of central tendency. So, to perform this test, at least, ordinally scaled data are required (cf., Fahrmeir et al. (2016)).

Table 2.5.: Example of how data is ranked in Wilcoxon signed rank test.

Testperson	Measurement A	Measurement B	Difference	Rank	Signed Rank
P1	4	5	1	1	(+)1
P2	7	4	-3	4	(−)4
P3	5	7	2	2 or 3	(+)2.5
P4	3	7	4	5	(+)5
P5	4	4	0	-	-
P6	3	5	2	2 or 3	(+)2.5

The Wilcoxon signed rank test is used to test whether two measurement series differ significantly with respect to their median of a characteristic (Witte, Witte (2019)). It is used for the comparison of paired measurement series, e.g., if two measurements are performed on the same test persons (cf., Bortz, Schuster (2010)), which is the case in our ET-studies, in which we want to investigate how gaze behavior measured by ET metrics such as gaze duration or gaze change differs between mathematics and physics tasks.

Assume that data are collected from a group of test persons at different points in time A and B, with the aim of comparing the medians (M_A and M_B) of the two measurement series. The two hypotheses are: Either there is no difference between the medians (null hypothesis H_0) or there is a difference (alternative hypothesis H_1) (Handl, Kuhlenkasper (2018)):

$$H_0 : M_A = M_B, vs.H_1 : M_A \neq M_B.$$

The operation of the Wilcoxon signed rank test can best be illustrated by an example, see Table 2.5. The two measurement series are compared pairwise. For each pair, the difference between the two measured values is determined. The absolute amounts of the difference are then ranked. The difference with the highest absolute value has the highest rank (Bortz, Schuster (2010)). If differences of different measurement pairs are identical, average ranks are assigned to them (Handl, Kuhlenkasper (2018)).

In the following, the sum of all ranks with negative and positive sign is calculated, in the present example: $\text{RankSum}_{-} = 4$ and $\text{RankSum}_{+} = 11$. Measurement pairs where the difference is zero, i.e., cases where no sign can be determined, are not considered. The larger the difference between the two sums RankSum_{-} and RankSum_{+}, the stronger the result argues against the null hypothesis (Bortz and Schuster, 2010).

More specifically, the Wilcoxon signed rank test is a linear rank statistic. After assigning ranks, indicated by i, the test statistic T for linear rank statistics can be computed by $T = \sum_{i=1}^{n} g(i) * Z_i$, where Z_i is an indicator variable indicating whether the associated pair difference is positive or negative, $g(i)$ is a weight function and n is the sample size. Expected value and variance for linear rank statistics are generally given by

$$E(T) = \frac{1}{2}\sum_{i=1}^{n} g(i) \text{ and } V(T) = \frac{1}{4}\sum_{i=1}^{n}(g(i))^2.$$

For the Wilcoxon signed rank test with the weight function $g(i) = i$, this results in an expected value and variance of

$$E(T_{Wilcoxon}) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4} \text{ and } V(T_{Wilcoxon}) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}.$$

Since the number of all possible permutations is 2^n, the probability of receiving a given value k is

$$P(T_{Wilcoxon} = k) = \frac{a(k)}{2^n},$$

where $a(k)$ is the number of all possible combinations of n-tupels $(z_1, \ldots, z_n)$ with value k. For samples $n > 20$, the distribution can be approximated by a normal distribution (Duller, 2019).

The Mann-Whitney U test is not the same as the Wilcoxon signed rank test, although both are nonparametric and involve the summation of ranks. The Mann-Whitney U test is applied to independent samples. The Wilcoxon signed rank test is applied to paired or dependent samples.

The resulting p-value indicates whether the results are statistically significant. The p-value is defined as the probability of obtaining the observed value or a value closer to the alternative hypothesis, assuming that the null hypothesis is true. If the p-value is very small, the observed values are unlikely to occur under the null hypothesis.

A significance level is defined usually at $\alpha = 0.05$ or $\alpha = 0.01$. If the p-value is smaller than this significance level, the null hypothesis is rejected, otherwise the null hypothesis is retained (Fahrmeir et al. (2016)).

The false discovery rate (FDR) is a method to conceptualize the rate of false rejections of the null hypothesis when conducting multiple comparisons. To avoid falsely rejecting the null hypothesis in multiple testing, the FDR can be controlled. The FDR is defined as the expected value of the quotient Q of the number of falsely rejected null hypotheses V and the number of hypotheses rejected R:

$$E(Q) = E\left(\frac{V}{R}\right).$$

FDR control procedures are used to control the FDR. The Benjamini-Hochberg method can be used to control FDR in multiple testing (Benjamini, Hochberg (1995)). It is used in this thesis to avoid falsely rejecting the null-hypothesis.

Based on the previous theoretical background, research questions, study design, and statistical methods, in the next chapter, we examine the learning processes of students solving similar linear function problems.

3. Better Understanding Students' Challenges on Linear Function Problems in Mathematics and Physics

The goal of this chapter is to gain deeper knowledge about the learning processes of students solving similar linear function problems in mathematics and physics. Similar means that these tasks require the same solution strategy to be solved. Furthermore, we transfer the results to teacher training in order to sensitise teachers to the latest scientific findings and thus improve learning in schools in a very concrete way. To achieve this goal, we first replicate the study by Ceuppens et al. (2019) and extend the sample size to higher grades. We show that the students' challenges identified by Ceuppens et al. (2019) for Belgian 9th graders persist across two countries and grade levels. But we also identify interesting differences. In a second step, we analyze the results of a large ET study ($N = 131$) to see if students' gaze differs depending on the context of the task. In particular, we examine tasks that show graphs or formulas. Here, we show that students' visual attention is context dependent. In a final step, we transfer the results to the school as part of a teacher training course.

3.1. Replication of a Study by Ceuppens et al. (2019) and Extension to High School Students

Validation

In this section, we report the internal consistency, the item difficulty index, and the discriminative power to analyze the validity of the testinstrument by Ceuppens et al. (2019). For details on the test instrument see Section 2.2. We report the results for German 9th graders ($N = 249$) and German 11th graders ($N = 298$) separately.

Internal Consistency: Internal consistency is measured by Cronbach's α (see Definition 2.26), which is $\alpha_9 = 0.846$ for 9th graders and $\alpha_{11} = 0.861$

Table 3.1.: Validation results 9th graders sorted by *context* ("K" for kinematics and "M" for mathematics in question number Q), by *representation* (Graph or Formula), by *task* (Compare, Determine), by *slope sign* ("+" for positive and "-" for negative) and by *concept* (y-intercept and slope). Statistical parameters are item difficulty index P, its standard deviation SD, Cronbach's α when the item in question would be deleted and item discrimination index D. Items marked with * require the student to provide an explanation for their answer.

Q	Representation	Task	Slope Sign	Concept	P	SD	α when deleted	D
M1	Graph	Compare	+	y intercept	0.32	0.47	0.81	0.74
M2*	Graph	Compare	+	Slope	0.76	0.43	0.82	0.39
M3	Graph	Compare	-	y intercept	0.39	0.49	0.82	0.61
M4*	Graph	Compare	-	Slope	0.16	0.37	0.82	0.17
M5	Graph	Determine	+	y intercept	0.24	0.43	0.81	0.65
M6*	Graph	Determine	+	Slope	0.26	0.44	0.81	0.48
M7	Graph	Determine	-	y intercept	0.21	0.41	0.81	0.70
M8*	Graph	Determine	-	Slope	0.16	0.37	0.81	0.48
M9	Formula	Determine	+	y intercept	0.24	0.43	0.81	0.74
M10*	Formula	Determine	+	Slope	0.26	0.44	0.81	0.48
M11*	Formula	Determine	-	y intercept	0.26	0.44	0.81	0.83
M12*	Formula	Determine	-	Slope	0.21	0.41	0.81	0.39
K1	Graph	Compare	+	y intercept	0.54	0.50	0.82	0.57
K2*	Graph	Compare	+	Slope	0.65	0.48	0.82	0.48
K3	Graph	Compare	-	y intercept	0.66	0.48	0.81	0.43
K4*	Graph	Compare	-	Slope	0.08	0.27	0.83	0.00
K5	Graph	Determine	+	y intercept	0.31	0.46	0.82	0.52
K6*	Graph	Determine	+	Slope	0.21	0.41	0.82	0.39
K7	Graph	Determine	-	y intercept	0.16	0.37	0.81	0.52
K8*	Graph	Determine	-	Slope	0.03	0.18	0.82	0.13
K9	Formula	Determine	+	y intercept	0.05	0.21	0.82	0.13
K10*	Formula	Determine	+	Slope	0.06	0.23	0.82	0.22
K11*	Formula	Determine	-	y intercept	0.06	0.23	0.82	0.17
K12*	Formula	Determine	-	Slope	0.01	0.08	0.82	0.04

for 11th graders. For both groups, we got a value greater than 0.8, which is a fairly high internal consistency according to Engelhardt (2009). Ceuppens et al. (2019) reports a slightly lower value with $\alpha = 0.791$ for Belgian 9th graders.

Cronbach's α was also calculated for each case of omitting an item from the test, and the results are shown in Table 3.1 for German 9th graders and Table 3.2 for German 11th graders. We see for both age groups that omitting a test item would not increase α in most cases (except for an increase of 0.1 for item K4 in the group of 11th graders). Ceuppens et al. (2019) reports a maximum increase of 0.4 in the case of item M4 for Belgian 9th graders.

Item Difficulty: Since our coding for the accuracy is dichotomous, the average score (accuracy) of each item directly translates into the item difficulty index P (cf. Definition 2.23). Ceuppens et al. (2019) measured item difficulty values ranging between 0.05 (item K12) and 0.92 (item K2) with and average item difficulty of 0.51 ± 0.26. We obtain the following results for German 9th graders: item difficulty ranging between 0.01 (item K12) and 0.76 (item M2), average:

Table 3.2.: Validation results 11th graders sorted by *context* ("K" for kinematics and "M" for mathematics in question number Q), by *representation* (Graph or Formula), by *task* (Compare, Determine), by *slope sign* ("+" for positive and "-" for negative) and by *concept* (y-intercept and slope). Statistical parameters are item difficulty index P, its standard deviation SD, Cronbach's α when the item in question would be deleted and item discrimination index D. Items marked with * require the respondent to provide an explanation for their answer.

Q	Representation	Task	Slope Sign	Concept	P	SD	α when deleted	D
M1	Graph	Compare	+	y intercept	0.71	0.45	0.85	0.52
M2*	Graph	Compare	+	Slope	0.87	0.34	0.86	0.14
M3	Graph	Compare	-	y intercept	0.64	0.48	0.86	0.57
M4*	Graph	Compare	-	Slope	0.35	0.48	0.86	0.41
M5	Graph	Determine	+	y intercept	0.59	0.49	0.86	0.59
M6*	Graph	Determine	+	Slope	0.57	0.50	0.86	0.59
M7	Graph	Determine	-	y intercept	0.59	0.49	0.86	0.66
M8*	Graph	Determine	-	Slope	0.44	0.50	0.86	0.43
M9	Formula	Determine	+	y intercept	0.61	0.49	0.86	0.70
M10*	Formula	Determine	+	Slope	0.70	0.46	0.86	0.43
M11*	Formula	Determine	-	y intercept	0.54	0.50	0.86	0.57
M12*	Formula	Determine	-	Slope	0.66	0.47	0.86	0.39
K1	Graph	Compare	+	y intercept	0.77	0.42	0.86	0.32
K2*	Graph	Compare	+	Slope	0.78	0.42	0.86	0.27
K3	Graph	Compare	-	y intercept	0.81	0.39	0.86	0.32
K4*	Graph	Compare	-	Slope	0.10	0.30	0.87	0.18
K5	Graph	Determine	+	y intercept	0.47	0.50	0.86	0.59
K6*	Graph	Determine	+	Slope	0.36	0.48	0.86	0.55
K7	Graph	Determine	-	y intercept	0.33	0.47	0.86	0.50
K8*	Graph	Determine	-	Slope	0.18	0.38	0.86	0.41
K9	Formula	Determine	+	y intercept	0.17	0.37	0.86	0.30
K10*	Formula	Determine	+	Slope	0.23	0.42	0.86	0.55
K11*	Formula	Determine	-	y intercept	0.19	0.40	0.86	0.48
K12*	Formula	Determine	-	Slope	0.10	0.30	0.86	0.20

0.26 ± 0.20; for German 11th graders: item difficulty ranging between 0.10 (item K12) and 0.87 (item M2), average: 0.49 ± 0.24. It is striking that the values for average item difficulty vary widely among the analyzed groups. Compared to Belgian 9th graders, German 9th graders perform nearly 30% worse on average! Compared to German 11th graders, German 9th graders perform 23% worse on average. In the end, German 11th graders' performance remains worse on average than that of Belgian 9th graders.

Results of the item difficulty analysis are shown in Table 3.3. We see that for German 9th and 11th graders, we obtain similar results than Ceuppens et al. (2019) for Belgian 9th graders. In summary, across countries and years, students face the least difficulty with "compare given a graph" question types and positive slope signs. However, they encounter the most challenges with kinematics problems, especially those involving the concept of velocity and mostly negative signs. In general, the most difficult and the easiest items are mostly kinematics items. The math items lack such extreme item difficulty indices. Ceuppens et al. (2019)

Table 3.3.: Item difficulty analysis.

	Highest four average scores	Lowest four average scores	note
Ceuppens et al.	K1,K2,K3,M2	K4,K8,K10,K12	
German 9th graders	K1,K2,K3,M2	K8, K9, K10, K11, K12	$P_{K4} = 0.08$
German 11th graders	K1,K2,K3,M2	K4, K8, K9, K12	$P_{K10} = 0.23$

did not ask for an explanation for the items in which the y-axis intercept had to be compared (item pairs no. 1 and no. 3), as they assumed that these items were easy to solve and that explanations were therefore not of interest. Like Ceuppens et al. (2019), we note here that this applies more to the kinematics than for mathematics items. This is an indication that the students disregard the similarities between the contexts.

Ceuppens et al. (2019) also observed that in kinematics, it is easier to determine the initial position (K-items with an odd number) than the velocity (items with an even number). In mathematics, it is easier to determine the slope (items with an even number) than the y-intercept (M-items with an odd number). We cannot confirm this pattern in our analysis. We obtain results the other way round for both 9 graders (cf. K1/2, K9/10) and items M3/4, M7/8 for 9th and 11th graders.

Discriminatory Power: The discriminatory power of the full test is $\delta_9 = 0.955$ (9th graders) and $\delta_{11} = 0.989$ (11th graders) (see Definition 2.25). For both age groups, we obtain a good discriminatory power higher than 0.9 which is desirable according to Kline (1986). Ceuppens et al. (2019) report a value of $\delta = 0.971$ for Belgian 9th graders that is only slightly higher than our value for German 9th graders.

On a per-item basis, the results for German 9th graders are similar to those for Belgian 9th graders: The discrimination index per item for German 9th graders is between 0.00 (item K4) and 0.83 (item M11) with an average of 0.43 ± 0.23 (Belgian 9th graders: value between 0.07 (item K12) and 0.82 (item M7), average: 0.44 ± 0.20). As with Ceuppens et al. (2019), we obtain very low values, less than the desired 0.3, for items K4, K8, K10, and K12. Furthermore, for the German 9th graders, we also get these smaller values for the items M4, K9, and K12. All items exept item M4 also have a very low item difficulty index smaller than 0.3. This shows that these items are very difficult for the students. Very few students can answer these items correctly, so they do not discriminate well between high and low performing students. For 11th graders, we obtain generally similar values than for 9th graders. Here, discriminatory power ranges between 0.14 and 0.70 with an average of 0.44 ± 0.15. Very low values (≤ 0.3) are obtained for items M2, K2, K4, K9, and K12. For K4 and K12, we also have low item difficulty indices of 0.10. For items M2 and K2, we obtain very low item discimination indices smaller than 0.3 and high item difficulty indices greater than 0.77. This

Table 3.4.: Sample data for GEE analysis (9th graders).

id	gen	item	ac	context	Qtype	concept	slope_sign
1	f	K1	1	K	compare_graph	y-intercept	+
1	f	K2	1	K	compare_graph	slope	+
1	f	K3	1	K	compare_graph	y-intercept	-
1	f	K4	0	K	compare_graph	slope	-
1	f	K5	0	K	determine_graph	y-intercept	+
1	f	K6	0	K	determine_graph	slope	+

means that those items are very easy for most of the students and that its score distribution is very narrow.

In a next step, we closely examine the effect of grouping concerning question task, concept, concept or slope sign on student's response accuracy.

Main Effects and Interaction Effects

In this section, we will model our clustered data using GEEs (see Section 2.2). Our goal is to gain a deeper understanding of the effects of the clusters in our data for 9th and 11th graders, such as context, gender, concept, slope sign, or question type, on students' response accuracy.

For the GEE analysis, our data is structured like in Table 3.4 for 9th and 11th graders, seperately. The variables are:

- *id:* numeric variable identifying the person (values 1-249, for 9th graders, 1-298 for 11th graders)
- *gen:* gender of each person (m: male, f: female, d: divers)
- *item:* name of the item (for kinematics items: K1, K2, . . . , K12 and mathematics items: M1, M2, ... , M12)
- *ac:* answer correctness for each item (1: correct, 0: not correct)
- *context:* context of the item (K: kinematics, M: mathematics)
- *Qtype:* question type of the item (compare_graph: compare two graphs, determine_graph: determine given a graph, *determine_formula:* determine given a formula)
- *concept:* concept of the item (y-intercept: tasks about y-intercept in mathematics or initial position in kinematics,
- *slope:* tasks about slope in mathematics or velocity in kinematics

Table 3.5.: Coefficients section of GEE analysis with basic model and correlation structure *exchangeable* for 9th graders.

Coefficients	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-0.59852	0.48083	1.549	0.21322
genf	-0.60594	0.49779	1.482	0.22350
genm	-0.26649	0.49592	0.289	0.59102
contextm	0.25075	0.08807	8.106	<0.01
concepty-intercept	0.27685	0.08656	10.229	<0.01
slope_sign+	0.70863	0.05180	187.155	<0.001
Qtypedetermine_formula	-1.84959	0.11518	257.889	<0.001
Qtypedetermine_graph	-1.38108	0.09529	210.075	<0.001

- *slope_sign:* sign of the slope of the linear function presented in the task ("−": negative slope sign, "+": positive slope sign)

After some data cleaning, setting NA values in the *ac* column to zero and then deleting all columns with NA values (for details see Knippertz (2024a)), we obtained a data set containing 215 of the original 249 participants (9th graders) and 267 of the original 298 participants (11th graders). Based on the considerations regarding the individual correlation structures (cf. Section 2.2), we will limit ourselves to the three structures *exchangeable*, *unstructured*, and *user-defined* to compare only those three correlation structures. For this, we will use a basic model without any second- or higher order correlations. The formula of the model is *ac* ∼ *gen* + *context* + *concept* + *slope_sign* + *Qtype*. This means that the answer accuracy (*ac*) is modelled by a linear predictor consisting of several variables (gender, context, concept, slope sign and question type). The "+" between the variables means, that we include the variables and their interaction in the model (Chambers, Hastie (1992)). For GEE analysis, we use the *geeglm*-function in R with binomial distribution and logit-link-function.

If we now use the basic model together with the *exchangeable*-correlation structure, we obtain coefficients in Table 3.5 for 9th graders. Results for 11th graders look similar. Here, the term *genm* is significant with $p < 0.01$ (cf. Table A.1 in Appendix). Note that character-values in the coefficients section are treated as numeric values in the analysis. The coefficient in the Coefficients column is always named with the characteristic that was set to 1. The remaining characteristic is set to 0. For example, *contextm* means that the context math was set to 1 and the other context kinematics was set to 0 in the analysis.

Table 3.5 shows the estimated marginal model for 9th graders. The coefficients are on the logit scale. We interpret these coefficients the same way we would any other binomial logistic regression model. The estimate for *contextm* coefficient (mathematics context) is 0.25. If we exponentiate, we get 1.28. This means the

Table 3.6.: Comparison of GEE-models with different correlation structures using QIC. Compared structures ara *exchangeable* (ex), *unstructured* (unstr), both with and without gender (Gen) and *fixed* (user-defined).

GEE model	QIC_9	QIC_{11}
gee.ex	4.79e+03	4.79e+03
gee.unstr	1.80e+32	1.80e+32
gee.fixed	4.85e+03	4.85e+03
gee.unstr_withoutGen	5.45e+03	5.45e+03
gee.ex_withoutGen	4.79e+03	4.79e+03

odds of solving a task correctly (accuracy=1) increase by about 28% for tasks with mathematical context. We see that for gender (*gen*), results are not significant and the associated standard errors are big. All other coefficients (except *intercept*) are significant and much bigger than their associated standard errors.

We will also analyze the GEE model with the *exchangeable* correlation structure without the variable *gender (gen)*, since it is not significant in model with *exchangeable* correlation structure. In this model, all remaining variables remain significant (cf. Table A.2 in Appendix A.2). Also for 11th graders, remaining variables remain significant (cf. Table A.3 in Appendix).

To decide which correlation structure is best suited for our purposes, we first visualized the correlation structures by using colored correlation matrices and then compared the structures using the QIC, cf. Section 2.2). Visualizations for 9th graders data are shown in Figure 3.1. The values in the *user-defined* structure are between 0.804 and -0.094 (diagonal is not considered). Mean value of all matrix elements without diagonal is 0.172 (close to *exchangeable* structure with $\alpha = 0.16$). We see clear blocks in the visualizations *unstructured* and *user-defined*. This means that there are correlations between the individual variables. The *unstructured* correlation matrix shows that y-intercept tasks tend to have a weaker correlation than slope tasks. (Visualizations for 11th graders look similar, see Knippertz (2024a).) Models with different correlation structures can be compared using QIC. We compared correlation structures *unstructured* (with and without gender), *exchangeable* (with and without gender) and *user-defined*. Results for 9th and 11th graders are shown in Table 3.6.

While the exponent of the power of ten of the QIC-value for the *exchangeable* structure is 3, the exponent of the power of ten of the QIC-value of the *unstructured* structure with 32 is much larger. For the *user-defined* correlation structure, we obtain QIC-value in the same order of magnitude as model with *exchangeable* structure. QIC-value of the *unstructured* structure gets similar to *exchangeable* structure when ignoring not significant gender. Models with *exchangebale* structure (with and without gender) have identical QIC-values. Models with

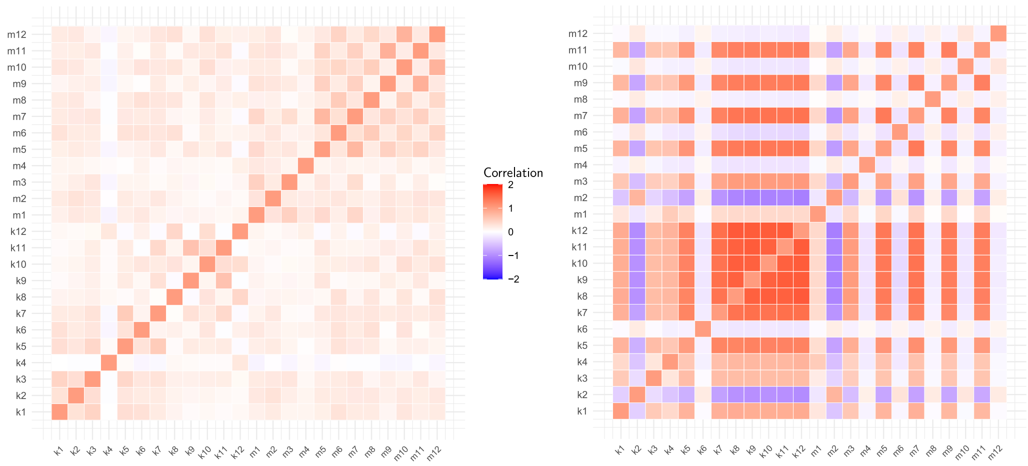


Figure 3.1.: Correlation matrices *user-defined* (left) and *unstructured* (right) (9th graders data).

the smaller QIC-values are preferred (cf. Pan, (2001)). Here, the model with *exchangeable* structure (with or without gender) is preferred!

Therefore, for further analysis, we use the correlation structure *exchangeable* together with the basic model without gender ($ac \sim context + concept + slope\_sign + Qtype$).

Now, we analyze main and interaction effects using GEE. We want to get information about the effect of *context, concept, slope sign, question type* and their interactions on the answer accuracy of the students. Note that we use the basic model without *gender*, since *gender* is not significant for GEE-analysis. Results for the main effect *concept* are obtained using ANOVA to test on significant differences between the basic model without *gender* and with the basic model without *gender* and without *concept*. We proceed in the same way for the other (interaction) effects. Details can be found in Knippertz (2024a).

To test whether the mean values of the response correctness of the individual characteristics of the variables ($context, concept, \dots$) are different, we apply the Wilcoxon test. We're not allowed to perform the t-test here because the response accuracy is not normally distributed (Shapiro-Wilk test: $p = 2e - 16 < 0.05$ (9th and 11th graders)). We obtain the following results (mean, Confidence interval (CI) and p-value of Wilcoxon-test, with Benjamini-Hochberg-Correction Benjamini, Hochberg (1995)) for the answer accuracy dependend on different variables:

- Concept: 0.302 $[0.264; 0.338]$ (y-intercept), 0.258 $[0.232; 0.289]$ (slope), $p < 0.001$ (9th graders); 0.482 $[0.446; 0.516]$ (y-intercept), 0.349 $[0.349; 0.406]$ (slope), $p < 0.001$ (11th graders)

Table 3.7.: Significant one-way, two-way and three-way effects with Wald $\mathcal{X}^2$, degrees of freedom (DOF) and significance (p-value).

Term	$\mathcal{X}_9^2$	DOF_9	p_9	$\mathcal{X}_{11}^2$	DOF_{11}	p_{11}
Concept	10	1	0.001	48	1	<0.001
Context	8	1	0.004	207	1	<0.001
Slope sign	187	1	<0.001	278	1	<0.001
QType	313	2	<0.001	350	2	<0.001
Context*concept	27	1	<0.001	43	1	<0.001
Context*Qtype	70	2	<0.001	210	2	<0.001
Context*slope sign	132	2	<0.001	307	2	<0.001
Qtype*slope sign	63	2	<0.001	141	2	<0.001
Concept*slope sign	185	1	<0.001	220	1	<0.001
Context*Qtype*slope sign	132	2	<0.001	307	2	<0.001
Context*concept*slope sign	176	1	<0.001	230	1	<0.001
Qtype*concept*slope sign	272	2	<0.001	320	2	<0.001

- Context: 0.258 [0.229; 0.285] (K), 0.312 [0.278; 0.346] (M), $p = 0.001$ (9th graders); 0.377 [0.346; 0.410] (K), 0.610 [0.567; 0.647] (M), $p < 0.001$ (11th graders)
- Question type: 0.258 [0.231; 0.288] (compare graphs), 0.056 [0.045; 0.070] (determine formula), 0.084 [0.069; 0.103] (determine graph), for all pairwise comparisons $p < 0.001$; 0.377 [0.346; 0.405] (compare graphs), 0.167 [0.146; 0.190] (determine formula), 0.197 [0.176; 0.221] (determine graph), for comparison determine on graph and determine on formula $p < 0.01$ for all other pairwise comparisons $p < 0.001$ (11th graders)
- Slope sign: 0.258 [0.230; 0.287] (−), 0.414 [0.374; 0.451] (+), $p < 0.001$ (9th graders); 0.377 [0.348; 0.407] (−), 0.540 [0.506; 0.573] (+), $p < 0.001$ (11th graders)

The differences within the values of a variable are always significant. This corresponds to the results of Ceuppens et al. (2019), although they used other statistical models.

Table 3.7 shows all significant one-way, two-way and three-way effects for 9th and 11th graders. Ceuppens et al. (2019) observed the same significant effects using another GEE model, which contains all main, two-way and three way effects, including the term gender. Comparisons of our results with GEE results of Ceuppens et al. (2019) are not actually permitted because of comparing two different GEE models.

Figure 3.2 shows the results for the term *context*concept* for both grade levels. We see that, concerning the slope, students perform significantly better in mathe-

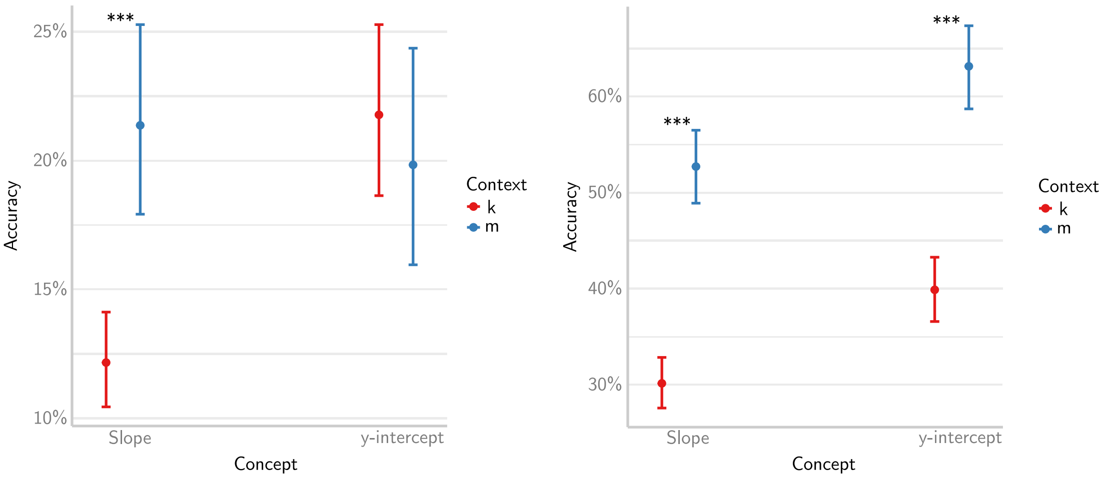


Figure 3.2.: GEE results for the terms *context*concept* for 9th (left) and 11th graders (right). The p-value of the comparisons is indicated by ** for $p < 0.01$ and by *** for $p < 0.001$.

matics than in kinematics. Concerning the y-intercept, students tend to perform better in kinematics; however, contrary to Ceuppens et al. (2019), this effect is insignificant. These effects could not yet be observed through item analysis alone (see previous section).

The results for the term *context*Qtype* are shown in Figure 3.3. In all question types, students perform better in mathematics than in kinematics. For the question types *compare graphs* and *determine on formula*, the context-specific differences concerning the answer accuracy are significant. Ceuppens et al. (2019) observed here significant difference the other way round for the question type *compare graphs* (like our results for 11th graders) and similar results otherwise.

For the term *context*slope sign* results are shown in Figure 3.4. Here, we can see that context-independent, students solve tasks with a negative slope sign significantly less successfully than tasks with a positive slope sign. Both types of task are generally solved correctly more often in mathematics, and the difference between mathematics and kineamtics is even significant for tasks with a negative slope sign. Here, Ceuppens et al. (2019) slightly observed different results concerning positive slope sign.

In general, accuracy values of German 9th graders are lower than values observed by Ceuppens et al. (2019).

Comparing the two-way-effects of German students in grades 9 and 11, grade 11 students perform better than grade 9 students. But the contextual difference in performance also increases significantly from grade 9 to 11. This means that performance in physics increases only slightly, while performance in mathematics increases enormously.

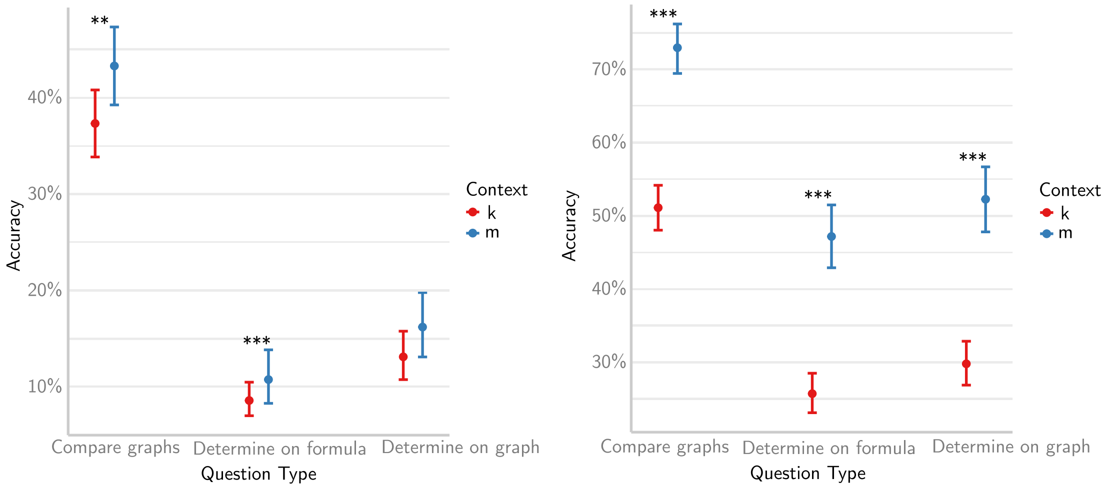


Figure 3.3.: GEE results for the terms *context*Qtype* for 9th (left) and 11th graders (right). The p-value of the comparisons is indicated by ** for $p < 0.01$ and by *** for $p < 0.001$.

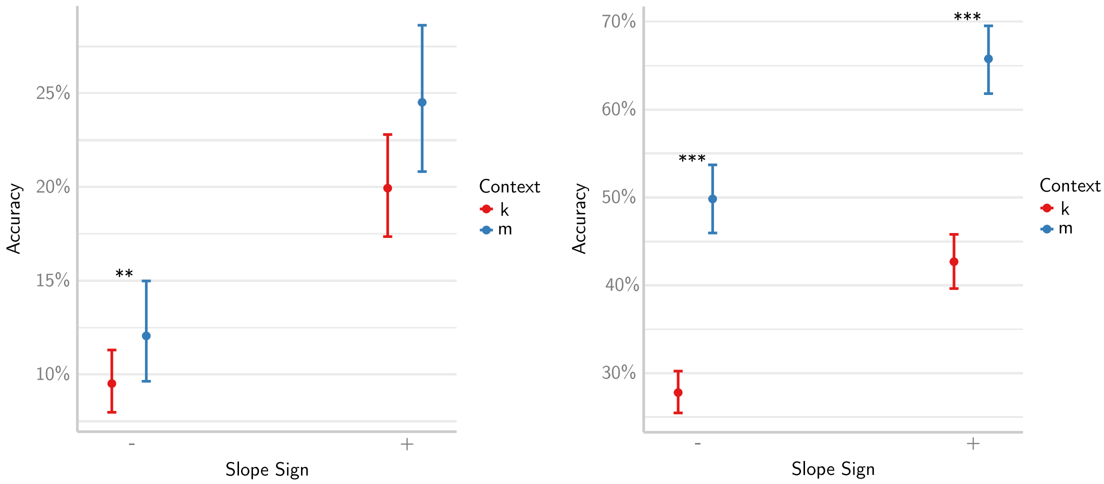


Figure 3.4.: GEE results for the term *context*slope sign* for 9th (left) and 11th graders (right). The p-value of the comparisons is indicated by ** for $p < 0.01$.

Results support the results from the validation. Students have the most difficulties in kinematics, concerning negative velocity, slope and all three question types (items K4, K8 and K12).

Qualitative Results

For the qualitative analysis, we use the coding guide based on the study by Ceuppens et al. (2019). An english version of the coding guide copied from Ceuppens et al. (2019) can be found in Section A.2 in the Appendix. In contrast to Ceuppens et al. (2019), we use an extension of the κ-coefficient to validate the classification scheme. The reason for this is that Cohen's κ as used by Ceuppens et al. (2019) is not appropriate here because the categories of the classification scheme are not disjoint (cf. Section 2.2). For example, a graph can be constructed (code I4) and then the x-intercept can be marked on this graph (code I2). For validation, two independent coders for each grade level (9 and 11) coded the entire material ($N_9 = 249$ and $N_{11} = 298$) exactly once, based on a translated version of the classification scheme in Section A.2 in the Appendix. The extended kappa coefficient for the 9th grade material is $\alpha_9 = 0.2036 \pm 0.0153$ and for 11th graders' material $\alpha_{11} = 0.0933 \pm 0.0166$. This only corresponds to a slight agreement accoring to Landis, Koch (1977).

When comparing the data from two coders, it is noticeable that there are relatively many tasks that were coded by one coder and not the other. One reason for this is that the coding guide allowed for both, i.e., one rater considered (none of) the solution strategies to be correct and the other rater considered a different solution strategy to be correct. On the other hand, due to the time-shifted coding and the enormous amount of material, the raters had no opportunity to discuss their coding and thus refine the coding guide. In order to maintain statistical accuracy, our aim was to have the coders code the entire material, not just N = 40 subjects, as one researcher in Ceuppens et al. (2019) did.

The results of the coding are shown in Table 3.8 for the slope tasks and Table 3.9 for the intercept tasks. The relative frequencies for the coded categories are shown in each case. Once overall for all slope or y-intercept tasks and once separated by context.

Note that all coders coded math tasks more often than physics tasks. This may be related to the fact that the category *Other (S99/I99)* was also coded more often in physics. Many explanations in physics seem not to be assigned to any solution strategy or are missing. It is also noticeable that the strategies *Identification of the x intercept (I2)*, *Construct an equation (I3)* and *Calculate a specific function value (I7)* dominate in mathematics, whereas the strategy *Intersection with vertical axis (I6)* dominates in physics. For the slope tasks, we see that the strategies *Location in an equation (S1)*, *Construct an equation (S3)* and *Drawing a triangle on graph (S7)* dominate in mathematics tasks and *Ratio of difference (S8)* dominates in physics tasks.

Table 3.8.: Relative Frequencies for the slope questions of 9th graders and 11th graders. f denotes the relative frequency across both contexts, f_K the rel. frequency in kinematic context and f_M in mathematics context. All frequencies are each calculated relative to the total number of observations N in the corresponding category (both context, M and K).

	9th graders						11th graders					
	Coder 1			Coder 2			Coder 3			Coder 4		
Category	f	f_K	f_M	f	f_K	f_M	f	f_K	f_M	f	f_K	f_M
S1: Location in an equation	0,18	0,06	0,27	0,12	0,04	0,17	0,17	0,07	0,27	0,17	0,06	0,25
S2: ID of x intercept	0,04	0,03	0,05	0,07	0,06	0,08	0,02	0,02	0,02	0,03	0,03	0,03
S3: Construct an equation	0,03	0,02	0,04	0,08	0,04	0,11	0,02	0,00	0,03	0,04	0,02	0,05
S4: Construct a graph	0,01	0,01	0,02	0,01	0,01	0,02	0,00	0,00	0,00	0,00	0,00	0,00
S5: Construct a table	0,01	0,01	0,01	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
S6: Reasoning with steepness	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
S7: Drawing triangle on graph	0,09	0,05	0,11	0,09	0,05	0,11	0,15	0,11	0,18	0,19	0,16	0,22
S8: Ratio of differences	0,30	0,34	0,26	0,29	0,35	0,25	0,16	0,16	0,15	0,28	0,30	0,27
S9: Ratio of coordiantes	0,00	0,01	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00
S99: Other	0,31	0,44	0,21	0,24	0,37	0,16	0,44	0,57	0,32	0,22	0,31	0,15
Sum	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
N	1164	501	663	1271	500	771	2183	1056	1127	1738	745	993

Table 3.9.: Relative Frequencies for the y intercept questions of 9th graders and 11th graders. f denotes the relative frequency across both contexts, f_K the rel. frequency in kinematic context and f_M in mathematics context. All frequencies are each calculated relative to the total number of observations N in the corresponding category (both context, M and K).

	9th graders						11th graders					
	Coder 1			Coder 2			Coder 3			Coder 4		
Category	f	f_K	f_M	f	f_K	f_M	f	f_K	f_M	f	f_K	f_M
I1: Location in an equation	0,07	0,13	0,02	0,04	0,05	0,03	0,02	0,03	0,01	0,02	0,03	0,01
I2: ID of the x intercept	0,17	0,05	0,26	0,26	0,15	0,32	0,12	0,04	0,21	0,14	0,06	0,21
I3: Construct an equation	0,09	0,03	0,14	0,17	0,09	0,21	0,02	0,01	0,04	0,07	0,03	0,11
I4: Construct a graph	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
I5: Construct a table	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
I6: Intersection with vertical axis	0,21	0,33	0,12	0,16	0,30	0,09	0,21	0,23	0,19	0,25	0,30	0,21
I7: Calc. a spec. function value	0,20	0,06	0,30	0,23	0,09	0,31	0,29	0,22	0,36	0,39	0,31	0,46
I99: Other	0,26	0,40	0,16	0,12	0,31	0,03	0,33	0,48	0,18	0,13	0,27	0,01
Sum	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
N	567	248	319	825	276	549	1952	971	981	1571	694	877

Regardless of the slope and y-intercept tasks, graphs and tables are almost never drawn.

Those results are all consistent with the results of Ceuppens et al. (2019). In contrast to Ceuppens et al. (2019), strategies *Reasoning with steepness (S6)* and *Construct a graph (I4)* do almost never appear in German students' material.

In summary, in this section, we identified students' difficulties in kinematics, especially with negative velocity across school years. For this, we performed an item analysis of the test instrument by Ceuppens et al. (2019) for 9th and 11th graders, examined cluster-effects with GEE analysis and performed a qualitative analysis on the student's task's explanations. In the next section, we will use ET data to examine contextual differences in student's visual attention when solving tasks of the test instrument by Ceuppens et al. (2019).

3.2. Visual Attention as an Indicator of Context-Dependent Solution Strategies

The goal of this section is to deeper understand cognitive processes of students when solving graph- or formula tasks in mathematics and kinematics. For this, we will analyze results of a large ET-school study of $N = 113$ German upper secondary students.

Graphs

Linear functions are an essential part of school and university education. Nevertheless, this topic is challenging for many students—especially in STEM topics. In this section, we contribute toward the understanding of learning difficulties concerning graphs regarding mathematical and physical problems. Here, we present the results of an ET study on visual attention during the interpretation of linear graphs using RTA method. Gaze data of $N = 131$ and interview data of $N = 60$ students were recorded while solving items of a validated test instrument from Ceuppens et al. (2019). We show that students' ET behavior differs significantly between mathematical and kinematic contexts. Triangulation of ET data with RTA data reveals context-dependent solution strategies—transfer from mathematics to physics is a major problem for students.

In this section, we show the analysis results of the ET data for IP 8, considering only subjects who completed both tasks M8 and K8 (N=105). Finally, results of the associated gaze-cued RTA interviews are provided ($N = 60$). For the descriptive results of the performance data for the entire test ($N = 131$) see Section A.2 in the Appendix.

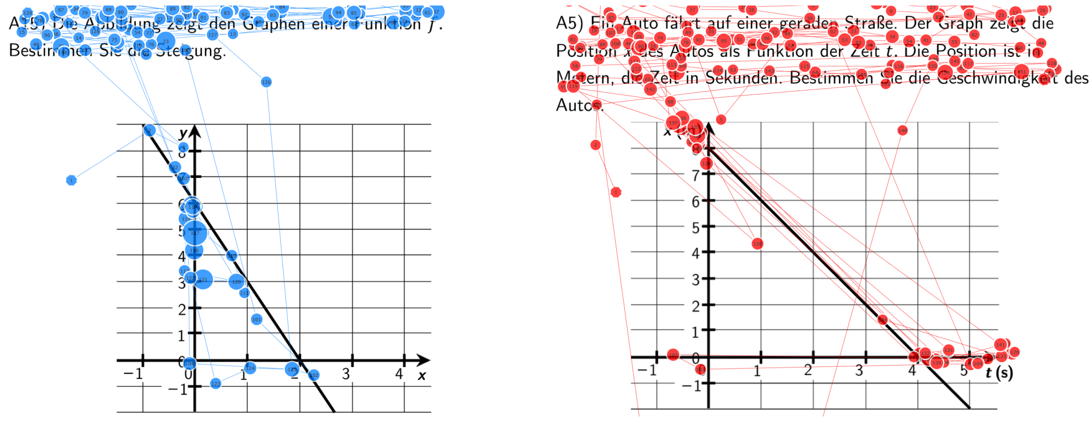


Figure 3.5.: Exemplary gaze plots of one and the same student for the mathematical item (M8, left) and the kinematic item (K8, right). The circle with numbers (1, 2, 3, 4...) indicates the order of observation, i.e., where the student looked first (1), subsequently (2), and so on, until the position where the student looked last.

Analysis of Eye-Tracking Data

Gaze plots: To get a first impression of the gaze movements for the two items of the similar IP 8, it is worth looking at the gaze plots first. In this way, the eye movements of one and the same subject can be directly compared between mathematical and kinematic context (see Figure 3.5). In the kinematic context, an accumulation of fixations (and thus of visual attention) on the axis segments is clearly visible, whereas in the mathematical context, the gaze is rather along the axes. However, this methodology only allows an exemplary comparison of gaze movements on individual levels. To determine whether eye movements differ between contexts across all subjects, the following sections report the results of statistical analyses on the ET metrics *TVD* and *transitions* are reported in the next sections.

TVD: Table 3.10 shows the means and standard errors of the TVD in seconds with respect to the related AOI for the kinematic and mathematical item of IP 8, as well as the p-values of the Wilcoxon tests. In case of a significant difference, the effect size was calculated with the corresponding confidence interval. Significant differences could be found for two AOIs (highlighted in Figure 3.6): the TVD for the kinematic item was significantly larger for the y-axis label (AOI 1), whereas the TVD for the similar mathematical item was significantly larger for the graph (AOI 8).

Transitions: The number of transitions between AOIs in both items of IP 8 were analyzed. For this purpose, the temporally resolved AOI hits of each subject were used to create transition matrices for item M8 and item K8, respectively, which contain all gaze changes between AOIs of an item.

Table 3.10.: Average TVD (Mean) in seconds per AOI with standard error (SE) for K8 and M8. The adjusted p-values of the Wilcoxon Test and effect sizes r with 95% confidence intervals are given for all Wilcoxon tests with significant results.

AOI	Mean_M	SE_M	Mean_K	SE_K	p (adj.)	r	CI
1	0.17	0.05	1.97	0.18	< 0.001	0.816	[0.783; 0.862]
2	3.45	0.41	3.48	0.32	0.541	-	-
3	2.43	0.25	2.19	0.24	0.626	-	-
4	0.43	0.08	0.41	1.06	0.778	-	-
5	0.58	0.07	0.82	1.09	0.229	-	-
6	1.82	0.30	2.44	0.31	0.156	-	-
7	0.48	0.07	0.68	0.10	0.370	-	-
8	3.79	0.35	2.63	0.37	0.008	-0.297	[-0.474 ; -0.133]
9	25.81	2.09	27.53	1.96	0.258	-	-

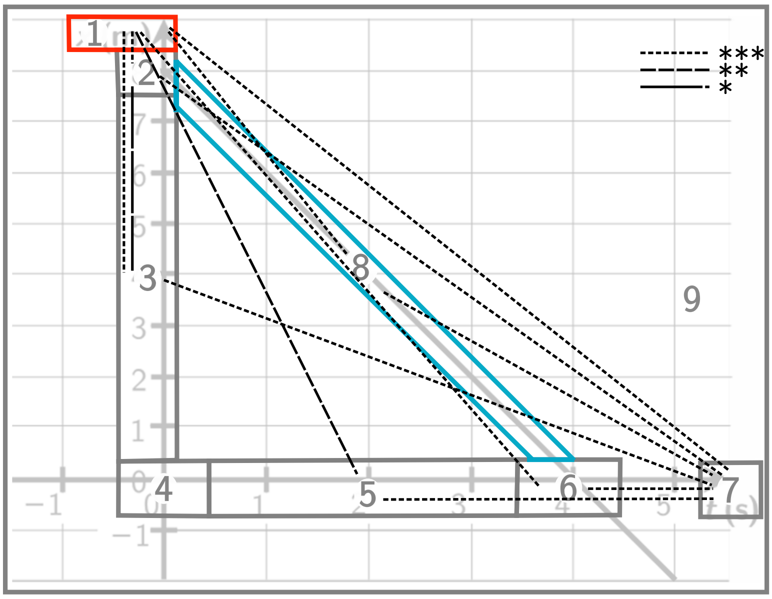


Figure 3.6.: Diagram area of item K8 divided into AOIs 1-9. AOIs with significantly longer TVDs in item M8 than item K8 are marked in blue. AOIs with significantly longer TVDs in item K8 than item M8 are marked in red. Dotted lines indicate transitions between AOIs that appear significantly more often in item K8 than in the similar counterpart.($p < 0.001$***, $p < 0.01$**, $p < 0.05$*.)

Table 3.11.: Average number of transitions (Mean) per transition between two AOIs with standard error (SE) for K8 and M8. The adjusted p-values of the Wilcoxon Test and effect sizes r with 95% confidence intervals are given for all Wilcoxon tests with significant results.

AOIs	Mean_M	SE_M	Mean_K	SE_K	p (adj.)	r	CI
1 | 2	0.18	0.05	4.00	0.37	< 0.001	0.797	[0.763;0.831]
1 | 3	0.10	0.04	0.81	0.12	< 0.001	0.570	[0.497;0.645]
1 | 5	0	0	0.12	0.04	0.003	0.328	[0.242;0.417]
1 | 6	0.01	0.01	0.30	0.07	< 0.001	0.393	[0.320;0.477]
1 | 7	0.04	0.02	0.30	0.06	< 0.001	0.366	[0.238;0.529]
1 | 8	0.04	0.02	0.65	0.11	< 0.001	0.500	[0.425;0.587]
2 | 3	2.10	0.26	2.71	0.29	0.048	0.228	[0.064;0.396]
2 | 7	0.01	0.01	0.29	0.06	< 0.001	0.425	[0.348;0.508]
3 | 7	0.02	0.01	0.20	0.05	< 0.001	0.343	[0.243;0.473]
5 | 7	0.01	0.01	0.37	0.08	< 0.001	0.444	[0.360;0.527]
6 | 7	0.16	0.05	1.61	0.21	< 0.001	0.658	[0.598;0.722]
7 | 8	0	0	0.30	0.05	< 0.001	0.480	[0.405;0.564]

To perform the Wilcoxon test, the transition matrices of all subjects were summed up. Table 3.11 shows the results of the Wilcoxon test (p values with effect sizes r for significant results) together with the average number of transitions between two AOIs. All significant results refer to more frequent transitions in item K8 than in item M8 and are visualized in Figure 3.6 differentiated by significance levels. Transitions along the axes that link information between axis labels or axis labels and other axis content occur significantly more often in item K8 than in item M8. Other gaze shifts between the axis intersection points and the graph also occur significantly more often in item K8 than in item M8. Four transitions that link information from the two axes also occur more frequently in the kinematic context than in the mathematics context. Except for one, all calculated effect sizes report medium ($> .3$) to large effects ($> .5$) after Cohen (1988).

RTA Analysis

Figure 3.7 shows the absolute frequencies of the codes *slope triangle*, *quotient of axis intercepts* (quot axis int), *other*, and *no strategy* (no strtg) for the RTA protocols of item M8 and item K8. The solution strategy *slope triangle* was coded more frequently for the mathematics item in the interviews, while the solution strategy *quotient of axis intercepts*, on the other hand, was coded more often for item K8.

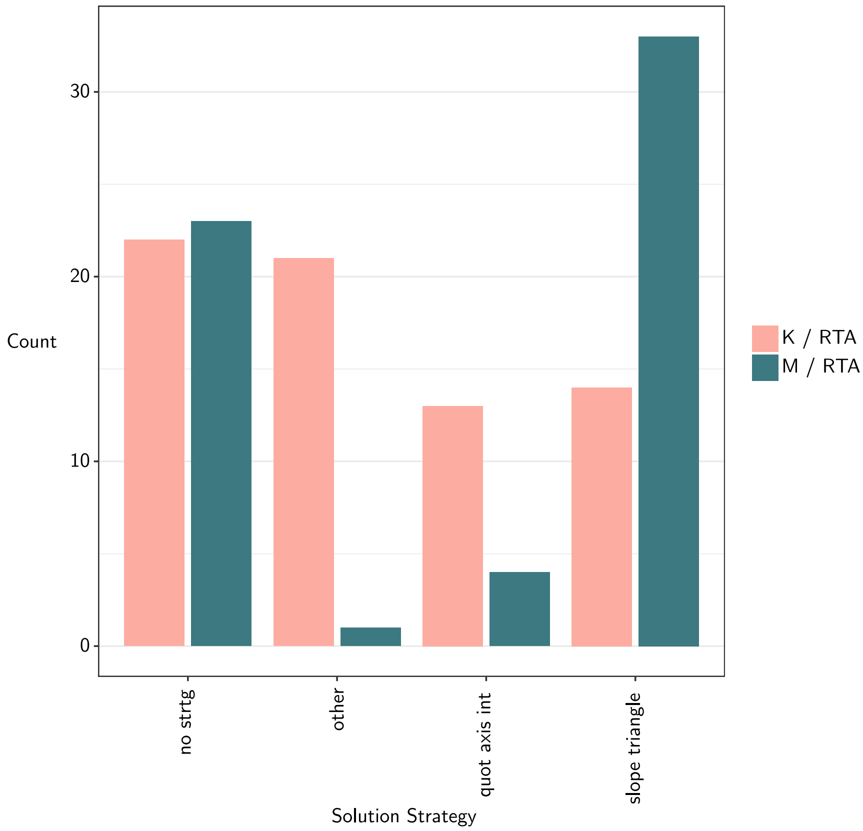


Figure 3.7.: Frequency of solution strategies in the mathematics (M) and kinematics (K) context for IP 8. The strategies *no strategy* (no strtg), *quot axis intercept* (quot axis int), *slope triangle*, and *other* were coded in the participants' interview data.

The measure of intercoder agreement was determined after Brennan, Prediger (1981) to $\kappa = 0,766$ and according to Landis, Koch (1977) corresponds to a substantial agreement. In a next step, we analyze ET and RTA data of IP 12—an IP with formulas.

Formulas

Mathematics is more than a crucial tool for scientists. It is a language to describe nature. Equations in particular are part of this language and they are ubiquitous in almost every scientific text and in STEM education. Numerous problems in understanding equations lead to difficulties in knowledge acquisition, transfer problems between different disciplines, understanding statistical data, communication problems and errors with potentially far-reaching consequences.
Here, we demonstrate for the first time that elements of equations are cogni-

tively processed in a non-linear sequence, which is very different to written text and sensitive to the context. These findings reveal the underlying difficulties in transfer between different disciplines. Our results provide a basis for improving learning processes and communication in interdisciplinary teams.

Equations, especially formulas, are an ubiquitous part of mathematical models to describe relationships of nature. Despite their omnipresence, cognitive processes involved in understanding formulas, influencing humans from their early days in school, are widely unclear. In this work, we use ET to reveal how high school students solve formula tasks on linear relationships which pairwise differed only in their context (math and physics respectively). Both tasks were similar, i.e., they required the same solution strategy. For the analysis, we used a triangulation of students' ET data, RTA protocols and written solutions. We found significant differences in the distribution of visual attention in the equation section in both contexts. More specifically, we measured a higher visual attention on the slope and the variable in the math formula as well as considerably more transitions in this area than in the physics formula. Combined with interview data and written solutions, we resolved context-sensitive cognitive processes. The analysis of gaze changes in the equation area showed that equations are not read in left-to-right reading direction like text. The impact of our results is present on two levels: On a small scale, transfer difficulties between math and physics in linear relationships are discovered; on a larger scale, problems concerning learning processes and interdisciplinary communication are uncovered by visual data. Our results show for the fist time, that similar equations in different contexts are read differently and also different to textual reading process in general.

In this section, we first present the descriptive results of the performance data for IP 12. For this purpose, the students' written solutions to IP 12 are analyzed. Then, the results of the ET analysis for IP 12 and results of the associated gaze-cued RTA interviews are shown, considering only participants who completed both tasks M12 and K12 ($N_{ET} = 112$, $N_{RTA} = 49$).

Descriptive Results

First, the item difficulty of the students for the tasks M12 and K12 is calculated. We found that the item difficulty for item M12 is 0.52 and for item K12 is 0.04 (cf. Section A.2 in Appendix). Thus, the physics item was solved correctly by very few students, whereas the math item was solved correctly by more than half of the students.

A more detailed view on the written student information is shown in Figure 3.8. Here, the relative frequencies of solution statements of IP 12 are presented. All frequencies are calculated relative to the total number of students, who answered items M12 and K12 ($N = 91$). For this purpose, the written solutions to items M12 and K12 were categorized by two experts into 6 categories: *correct*, *misstatement with slope value*, *slope-height-confusion*, *function equation*,

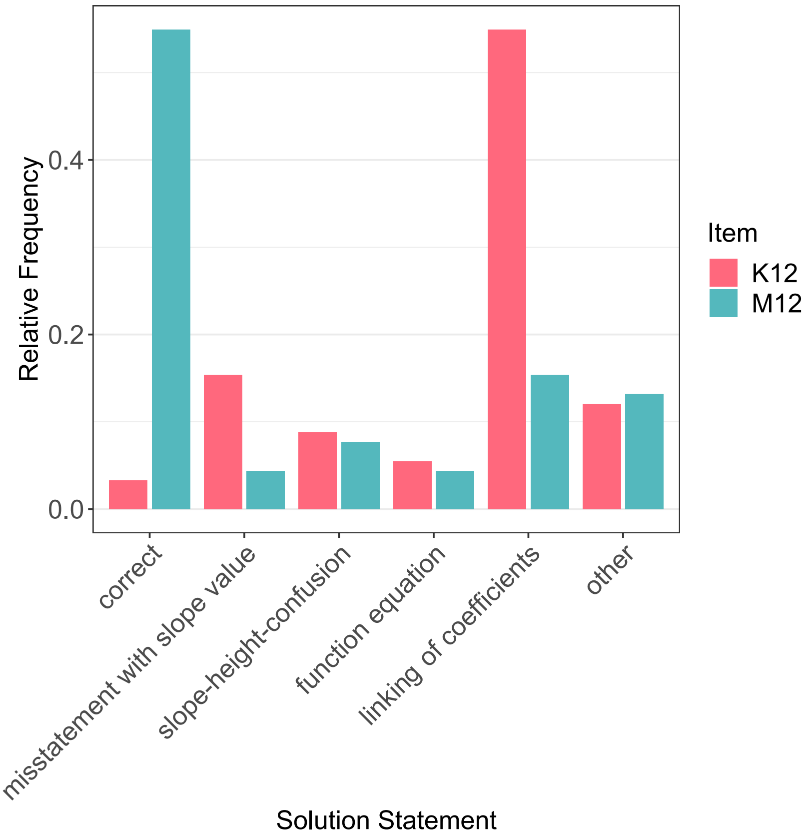


Figure 3.8.: Frequency of written statements, categorized in *correct*, *misstatement with slope value*, *slope-height-confusion*, *function equation*, *linking of coefficients* and *other*, relative to the number of statements on M12 and K12 ($N = 91$).

linking of coefficients and *other*. Clearly, contextual differences are recognizable: In the mathematical context, the correct solution strategy dominates, whereas in the kinematic context, the coefficients are often linked, i.e., the sum, difference, product or quotient of the coefficients was calculated.

Analysis of Eye-Tracking Data

TVD in content-specific AOIs: Table 3.12 shows the means and standard errors of the relative TVD in seconds with respect to the related AOI for the kinematic and mathematical item of IP 12, as well as the p-values of the Wilcoxon tests. The relative TVD is calculated in relation to the TVD of the complete formula area (AOI 1-7). In case of a significant difference, the effect size is calculated with the corresponding confidence interval. Significant differences could be found for seven AOIs (highlighted in Figure 3.9):

Figure 3.9.: Formula area of items K12 and M12 divided into AOIs 1-7. AOIs with significantly longer TVDs than in the other item are marked in red. Transitions between AOIs that appear significantly more often in one item than in the similar counterpart are shown as arrows.

Table 3.12.: Average relative TVD (mean) in seconds per AOI with standard error (SEM) for K12 and M12. The total visit duration was calculated relative to the total visit duration of the whole formula area (AOI 1-7). The adjusted p values of the Wilcoxon test and effect sizes r with 95% confidence intervals are given for all Wilcoxon tests with significant results.

AOI	Mean_M	SEM_M	Mean_K	SEM_K	p (adj.)	r	CI
1	0.13	0.01	0.06	0.01	0.000002	-0.548	[-0.734; -0.409]
2	0.04	0.01	0.06	0.01	0.000043	0.451	[0.262; 0.644]
3	0.12	0.01	0.06	0.01	0.000123	-0.417	[-0.601; -0.244]
4	0.10	0.01	0.21	0.01	<0.000001	0.564	[0.424; 0.721]
5	0.08	0.01	0.20	0.02	<0.000001	0.694	[0.574; 0.832]
6	0.23	0.02	0.34	0.02	0.000053	0.437	[0.271; 0.622]
7	0.29	0.02	0.06	0.01	<0.000001	-0.815	[-0.898; -0.768]

The relative TVD for the kinematic item is significantly larger for the equal sign (AOI 2), both coefficients (AOIs 4 and 6) and the sign of the linear coefficient (AOI 5) than in the mathematical counterpart, whereas the relative TVD for the similar mathematical item is significantly larger for the function name (AOI 1), the x variable (AOI 7) and the sign of the constant coefficient (AOI 3) as in the kinematics item.

TVD in position-specific AOIs: Figure 3.10 shows the mean TVD as a function of position in the formula area. For this purpose, the formula area of both items M12 and K12 was divided into 16 equidistant AOIs (positions). There are clearly different courses of the TVD values in both items: In item M12, the longest visual attention is on the slope value (AOI 10), whereas the course of the visual attention in item K12 shows two peaks that cover the two coefficients (AOI 8 and AOI 14/15). In item M12, there is a slightly stronger focus on the variables than in item K12.

Transitions: The number of transitions between AOIs in both items of IP 12 were analyzed using content-specific AOI positioning (cf. Figure 3.9).

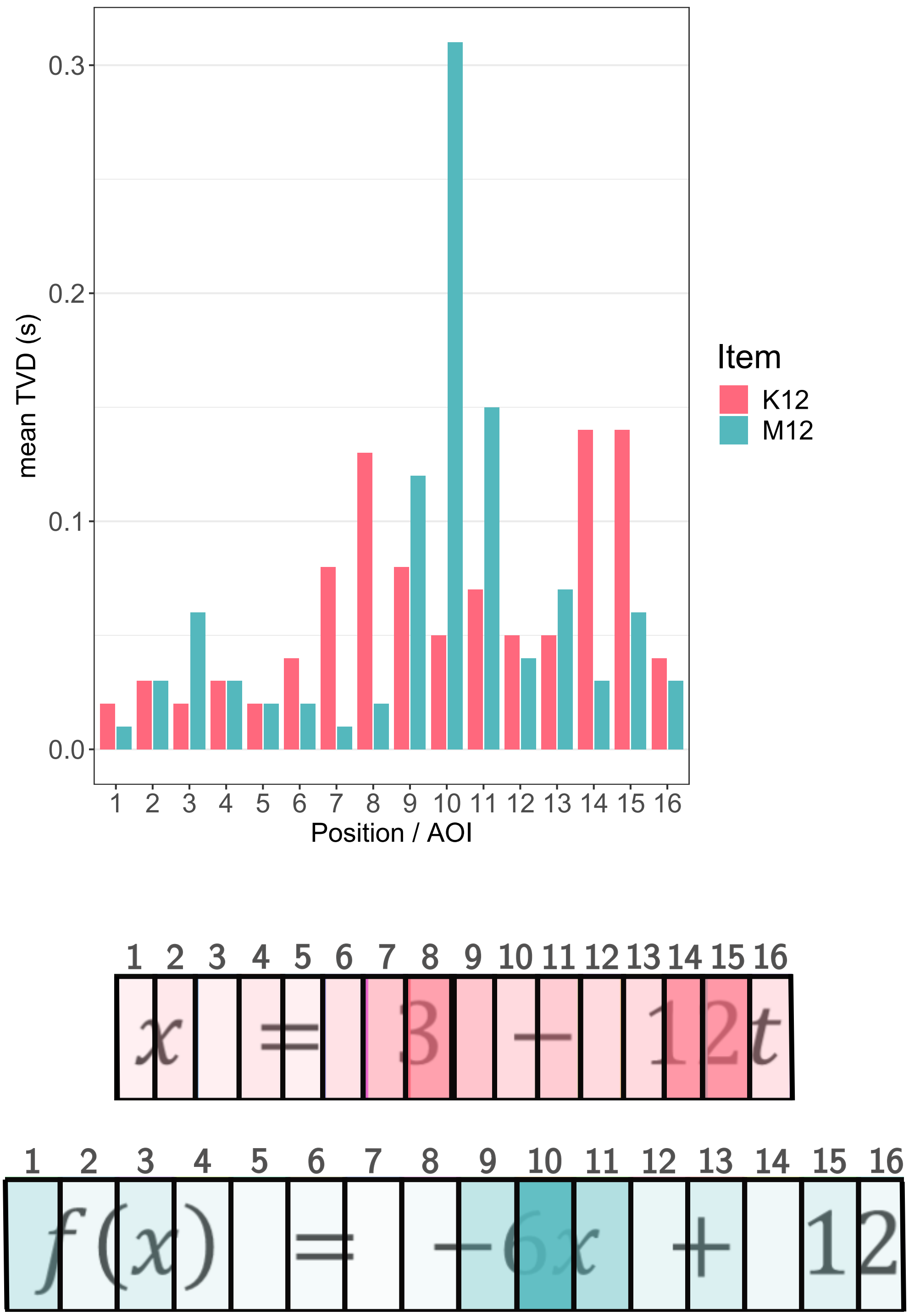


Figure 3.10.: Mean total visit duration (in seconds) depending on the position within the formula range for M12 and K12. The positions for each item M12 (in red) and K12 (in blue) are shown below the diagram. The colour intensity of the AOI range of each position 1-16 is scaled according to the mean TVD value of this AOI in the diagram, i.e., the darker the colouring, the higher the mean TVD value.

Table 3.13.: Average number of transitions (mean) per transition between two AOIs with standard error (SEM) for items K12 and M12. The adjusted p values of the Wilcoxon Test and effect sizes r with 95% confidence intervals are given for all Wilcoxon tests with significant results. Table contains results with and without considering reading direction.

AOIs	Mean_M	SEM_M	Mean_K	SEM_K	p (adj.)	r	CI*
1|2	0.68	0.16	0.90	0.19	0.497	-	-
1|3	0.47	0.09	0.81	0.24	0.497	-	-
1|4	0.23	0.06	1.20	0.24	<0.001	0.690	[0.567;0.875]
1|5	0.60	0.10	0.30	0.07	0.015	-0.412	[-0.738;-0.161]
1|6	1.44	0.24	0.33	0.07	<0.001	-0.653	[-0.819;-0.505]
1|7	1.69	0.27	0.08	0.03	<0.001	-0.850	[-0.896;-0.822]
2|3	0.22	0.07	0.45	0.12	0.209	-	-
2|4	0.05	0.02	1.16	0.24	<0.001	0.802	[0.734;0.901]
2|5	0.17	0.05	0.55	0.10	0.003	0.521	[0.288;0.796]
2|6	0.35	0.07	0.33	0.08	0.736	-	-
2|7	0.57	0.12	0.05	0.03	<0.001	-0.780	[-1.010;-0.637]
3|4	2.51	0.35	1.23	0.32	<0.001	-0.503	[-0.728;-0.330]
3|5	0.35	0.08	0.59	0.12	0.118	-	-
3|6	1.53	0.24	0.88	0.21	0.004	-0.385	[-0.619;-0.158]
3|7	2.42	0.32	0.11	0.04	<0.001	-0.824	[-0.897;-0.781]
4|5	0.18	0.05	2.68	0.31	<0.001	0.872	[0.868;0.872]
4|6	1.20	0.22	4.73	0.58	<0.001	0.714	[0.613;0.858]
4|7	2.55	0.40	0.73	0.16	<0.001	-0.627	[-0.801;-0.478]
5|6	1.49	0.39	4.84	0.49	<0.001	0.668	[0.542;0.800]
5|7	1.09	0.14	0.81	0.14	0.152	-	-
6|7	3.99	0.56	1.78	0.30	<0.001	-0.429	[-0.631;-0.234]

*Results considering reading direction differ slightly due to bootstrap method.

For this purpose, transition matrices containing all gaze changes between AOIs of item M12 and item K12, respectively, were created from the temporally resolved AOI hits of each subject. The transition matrices of all participants were summed up to perform the Wilcoxon test. Table 3.13 shows the results of the Wilcoxon test (p-values with effect sizes r for significant results) together with the average number of transitions between two AOIs.

There are more transitions in item M12 than in item K12, that appear significantly more often in the mathematics item than in the kinematic counterpart (cf. Figure 3.9). In M12, the three transitions with the highest effect sizes (> 0.78) start from the linear term (AOI 7). In contrast, in K12, the three transitions with

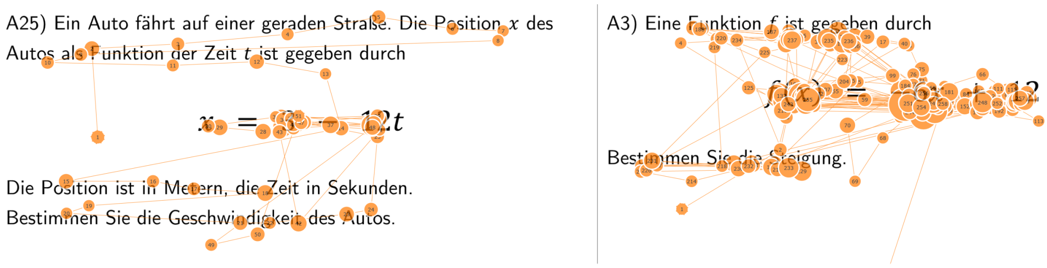


Figure 3.11.: Exemplary gaze plots of one and the same student solving items K12 (left) and M12 (right) illustrating the different visual processing of formulas and text.

the highest effect sizes (> 0.71) start from the constant term (AOI 4) or even combine the constant term with the linear term (transition 4–6). Apart from three values, all calculated effect sizes report large effects ($> .5$) after Cohen (1988).

Influence of reading direction: To examine the influence of the reading order on formulas, the transition analysis is carried out without transitions in the reading direction. Transitions in reading directions are transitions 1-2, 1-3, 1-4, ..., 1-7, 2-3, 2-4, ..., 2-7, ... for K12 and 1-2, 1-5, 1-6, 1-7, 1-3, 1-4, 2-5, 2-6, ... for M12. Transitions that occur first in the reading direction and then in the opposite direction are counted. Surprisingly, all results from Table 3.13 remain unchanged. The small deviations within the limits of the confidence intervals occur because in our non-parametric setting all confidence intervals were estimated using the Bootstrap method (Efron (1987)). These results show: Reading effects do not play a role in the processing of equations. Exemplary gaze plots illustrating these results are shown in Figure 3.11. While the text is usually read from left to right, information in equations are linked to each other and to the text. This is illustrated by numerous transitions between individual formal elements and between text and equation in Figure 3.11.

RTA Analysis

Figure 3.12 shows the absolute frequencies of the codes *read from formula*, *equation arrangement*, *linking of coefficients*, *slope-height-confusion*, *other* and *no strategy* for the mathematics item M12 and the kinematics item K12 in the interviews. Contextual differences can be seen with respect to the solution strategies in the diagram. The solution strategy *linking of coefficients* is dominant in the kinematics item, whereas, in mathematics, the only correct solution strategy, *read from formula*, was chosen most often. It is striking that, in both contexts (M and K), often *no strategy* was chosen. This is either an indication of the students' difficulties in explaining orally or an indication of many guessed answers. The measure of intercoder agreement was determined according to Brennan, Prediger

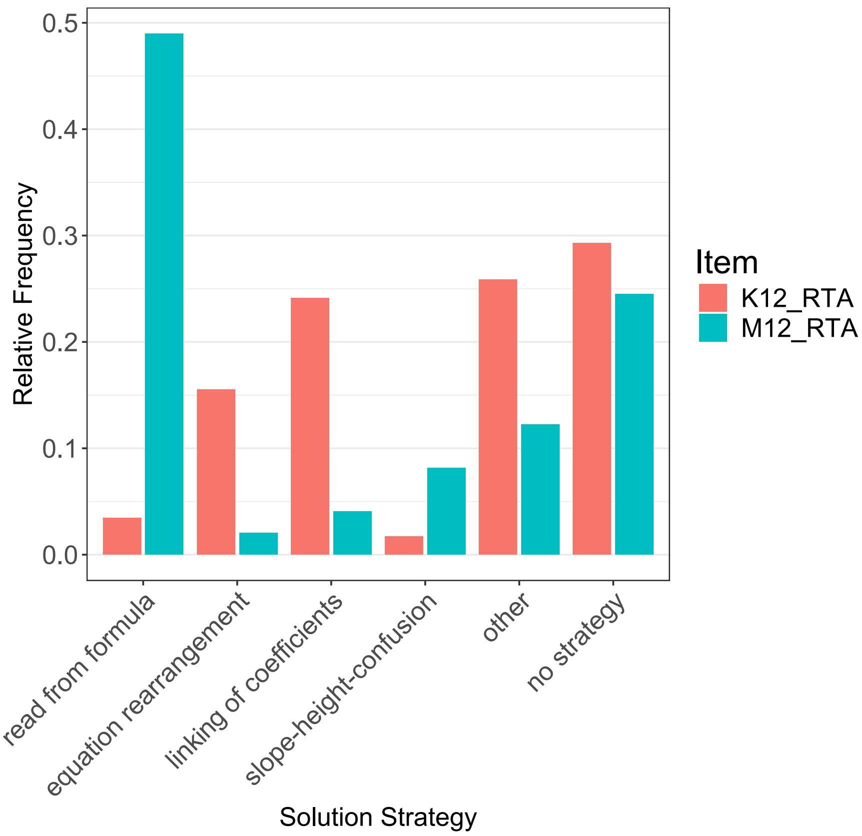


Figure 3.12.: Frequencies of solution strategies in RTA data of mathematical item M12 and kinematic item K12 relative to the total number of interview data on IP 12 ($N = 49$). The strategies *read from formula*, *equation rearrangement*, *linking of coefficients*, *slope-height-confusion*, *other* and *no strategy* were coded in the participants' interview data.

(1981) to $\kappa = 0.836$. Following Landis, Koch (1977), this corresponds to a almost perfect agreement (> 0.81).

Do learners in the interview choose the same strategy to solve similar tasks in different contexts? Only RTA pairs from IP 12 ($N = 49$) were considered. For each student, it was checked whether the strategy chosen in their interview for items M12 and K12 matched or not. A match was present if, for example, the student chose the same solution strategy *linking of coefficients* in both items M12 and K12. There was a mismatch if the student's solution strategies did not match, e. g. *read from formula* was selected for M12 and *no strategy* for K12. All strategies were considered, whereby there were no matches for the category *other*. There were 12 matches and 37 mismatches. When looking at all cases of agreement, it is noticeable that the strategy *no strategy* was coded in 50% of all cases and is therefore the most common strategy. In the non-matches, there were 23 of 37 cases in which the only correct solution strategy *read from formula* was chosen for item M12 and the strategy *other* was chosen in item K12. This means that the correct strategy in M12 is lost in 62% of all cases with a change of strategy when transferring to K12.

3.3. Teacher Training Course on Students' Challenges with Linear Functions

This section deals with one way of transferring the previous empirical results into practice in the form of a teacher training course, especially from mathematics to physics. This part is not empirical, but merely an explorative attempt to sensitize teachers to transfer problems in the classroom and to provide concrete help on how these problems can be solved.

The decision to hold a teacher training course on transfer difficulties from mathematics to physics was first and foremost a personal one. Research in the school sector requires concrete feedback from the school in order to have an impact and should not only take place at university level - we see this as a necessity. The intention of our research is essentially to improve learning and teaching, so it thrives on a transfer to the school. The word transfer comes from the Latin "transferre," which means "to carry over". In relation to schools, the transfer of scientific knowledge is intended to provide practical solutions to school issues and problems. However, this transfer does not work in a linear fashion, i.e. in the sense that scientific knowledge can be communicated to schools on a one-to-one basis. The findings must be adapted to the school's needs and conditions. In our case, this means that the scientific findings must not only be communicated, but also transferred – or rather transformed – into concrete teaching examples in order to improve teaching and learning (Bieber et al. (2018)). Therefore, teacher training focuses on providing concrete practical examples that can be used immediately in the classroom. It is important here that the needs and problems of teachers regarding challenges in transferring mathematical topics to physics are considered in particular, so that the training can be specifically linked to this and thus represent as much added value as possible for the participants. A physics and mathematics teacher who also works at the Kaiserslautern study seminar, was interviewed for this purpose. She sees two major subject areas in particular where transfer difficulties occur on a massive scale: Functional relationships and vectors.

One main intention of the training was to sensitize teachers to the fact that it is not the lack of mathematical knowledge of their students that is the predominant problem, but rather the transfer of this knowledge into new (non-mathematical) contexts. Once (physics) teachers would be aware of this issue, they could place a greater focus in their own teaching on the use of mathematical solution procedures in solving physics problems and make their students aware of this in various contexts. A targeted linking of mathematics lessons and physics lessons could here on the one hand help to promote the transfer of mathematical procedures for solving physical problems, but on the other hand also bring an application connection into mathematics lessons. Through this linking, students develop a repertoire of mathematical solution procedures for physical problems, which they can then access in different contexts to successfully solve subject problems. The resulting mental flexibility of students to select the appropriate mathematical procedure for a given problem could help students to solve problems not only in physics specifically, but in the natural sciences in general.

The training course presented below was attended by four teachers from secondary schools in Rhineland-Palatinate and took place at the RPTU in Kaiserslautern on June 28, 2024. The structure and procedure of the training course are presented first. We then go into the content elements of the training and the evaluation of the training by the teachers. Finally, there will be a brief summary of the participant survey for the course.

The teacher training addresses primarly challenges transferring mathematical topics to physics. The teachers participating in the training course teach mathematics and physics (2), mathematics and geography (1) and mathematics and chemistry/biology (1). For this reason, challenges in transferring mathematics to other subjects besides physics were also considered.

Structure and Material

The training is divided into two parts: In the morning, primarily theoretical basics were covered. The focus was on communicating scientific results. In addition to the theoretical and technical knowledge imparted, this first part also included numerous practical exercises to apply and consolidate what had been learned. First, the theoretical basics of problem-solving processes were covered. This was followed by the methodological principles of the empirical studies on transfer difficulties. Finally, the main results of the studies on transfer difficulties with linear functions were presented and implications for teaching were discussed.

The second part of the training occurred in the afternoon and included numerous tasks to raise awareness of further transfer difficulties and how to deal with them in the classroom. This was followed by a small final round with a short evaluation.

A presentation was used to convey the content, the tasks were partly paper-based and partly required the use of HyperDocs (Fitting, Hornung (2021)) or GeoGebra (International GeoGebra Institute Wolfauerstr 90 (2021)). The entire original material can be found on Knippertz (2024c).

Content

Functional relationships have enormous everyday relevance. For example, the train journey from Munich to Hamburg can be approximated as a linear place-time function: The following assumptions must be made for this. Only the average speed of the train over a time interval on approved routes, in which the ICE1 travels at an average speed of 250 km/h, is considered. We assume that the train is point-shaped and moves uniformly on a straight line. Frictional forces are neglected. Other examples are climate diagrams. These diagrams show the temperature and the amount of precipitation as a function of time (usually one year) at a location. This information can also be available in tabular form and also requires adequate interpretation. Functional correlations can be seen in newspapers, magazines and web content.

To better understand problem-solving processes, CTML can be helpful. This theory identifies three processes regarding problem solving and learning with multimedia: selection, organization, and integration (see Section 2.1). The teachers applied the theory to a given task on linear functions and other, specially constructed tasks from their subject areas. Modeling cycles in mathematics, chemistry, and physics were then dealt with. The aim here was to have a tool for analyzing problems in these subject areas. Similarities and differences between the cycles were discussed. In addition, student difficulties in physics were assigned to individual areas of the modeling cycle. The teachers found that most problems in physics, such as the interpretation of borderline cases, graph-as-picture-error, etc., can be located in the area of the physical-mathematical model. The connection between mathematical structures and physical meaning is therefore often missing (cf. Section 2.1). The work with modeling circuits was deepened in subsequent exercises on mathematical, chemical and physical modeling cycles.

The next part of the training focused on the methodological basics of ET and RTA. The part on ET dealt with definitions of ET and how an eye tracker works, theoretical principles such as the eye-mind hypothesis, various visualization technics such as gazeplot and heatmap and analysis methods for measuring ET measures in areas of interest. Finally, it was shown how ET measures can be interpreted cognitively using CTML (see Section 2.1) and the opportunities and risks of ET were discussed. The RTA method was explained in terms of how it works.

In the last part before the lunch break, the main results of the replication study and the ET study were discussed. There was also a brief excursion into network analysis and the detection of slope triangles (see Chapters 2 and 4). This part was accompanied by practical sections on linear functions in mathematics and physics, some of which also required the interpretation of real student ET and interview data.

The second training session in the afternoon was divided into three parts. First, mathematical concepts in physics, chemistry, and geography were covered. In addition to a task to raise awareness and promote understanding of the concept of rates of change in physics and everyday life, which the teachers worked on independently, difficulties in understanding other mathematical concepts such as ratio or product in other contexts were also dealt with here. The second section dealt with teaching concepts for formulas and diagrams. A student task by Geyer, Pospiech (2016) served as a template. Among other things, physical diagrams have to be assigned to physical formulas in a playful memory style. In the area of formulas, a model for decoding formulas according to Jansen et al. (2007) was dealt with first. Afterwards, the teachers tried out a playful exercise in snip-snap style from the same authors of the model and reflect on its use in their lessons. Like the memory task, this can also be easily transformed into subject areas other than physics.

The last section before the concluding discussion and participant survey dealt with vectors. The focus here was initially on different approaches and ideas of vectors in mathematics and physics (cf. Dilling (2019)) and the associated sensitization to stimulate these differences in math and physics lessons. This was followed by (negative) examples from math and physics books. After a short excursus on units in math and physics, the teachers finally had the opportunity to create or work on their own HyperDoc on exponential functions or, alternatively, a modeling task on math and sports. The results of the participant survey will be presented in the next section. The entire material including exercises in German and English can be found on Knippertz (2024c).

Results

A total of four teachers took part in the training. All four teachers teach at secondary schools. They rated the usefulness and enjoyment of the event very highly (usefulness: 4/4, enjoyment: 4.8/5). There was enough time to develop own ideas, mathematical and didactic input, examples and discussions (scale 1 to 5, 1: too little, 5 too much, average rating: 3). The training was rated three times as very good and one time as good. The teachers stated that the event helped them to deepen their mathematical and didactic knowledge. They also took away new ideas for their lessons and got to know other teachers. In the free text responses to the survey, the teachers specifically indicated a benefit with regard to ET as help and support, raising awareness of the topic, insights into new tools and careful consideration of transfer difficulties, making more connections. The exchange with colleagues, new insights into procedures/evaluations, exchange within the group, the pace of the speakers, the selection of examples, the number of breaks, the new input ET, the relaxed atmosphere (informative and varied) were considered particularly good. No points were mentioned under “not good.”

3.4. Discussion and Summary

Discussion

RQ1 and RQ1.1: Difficulties diagnosed by Ceuppens et al. (2019) were problems in kinematics especially with negative velocity and transfer difficulties applying mathematical knowledge in physics. All those challenges are replicable for German students. As Ceuppens et al. (2019), we found that students have more difficulties solving tasks in kinematics than mathematics and severe problems with tasks with negative velocity (RC1.1). In contrast to Belgian 9th graders, German 9th graders performed much worser in general with an average item difficulty index nearly 30% lower than the index of the Belgian 9th graders. Reasons for such a large difference in performing could be that we conducted the study during the last period of the 2022 coronavirus pandemic. During this year, many schools became familiar with homeschooling and digitization, but the quality of teaching and performance may have deteriorated since the beginning of the pandemic and learning gaps may still not be filled. This may have contributed to a decline in the quality of instruction and thus performance. Another reason could be the differences between the German and Belgian school systems and curricula. However, a look at the results of the 2022 and 2018 PISA studies shows that the differences in performance between the two countries only diverge in mathematics (cf. OECD (2019) and OECD (2023)): Here, Belgium (489 points (2022), 508 points (2018)) performs over the years significantly better than Germany (475 points (2022), 500 points (2018)). In science, the two countries are comparable (Belgium 491 (2022) and 499 (2018) points, Germany 492 (2022) and 503 (2018) points). Interesting is, whether the PISA results reflect the context-specific differences in performance. The average item difficulty index for mathemtics tasks is with $0.29 \pm 0.16(SD)$ for German 9th graders lower than the average index of $0.55 \pm 0.18(SD)$ for Belgian 9th graders. Regarding kinematics, Belgian students perform significantly better with an average item difficulty index of $0.47 \pm 0.33(SD)$ than German 9th graders (0.23 ± 0.24(SD)). Note that both measures are subject to large fluctuations. The PISA results reflect only performance in mathematics for this test instrument, in physics, Belgian students perform much better than German students.

With regard to solution strategies, the results of the qualitative analysis show that both the German and Belgian 9th graders preferred the strategies *Identification of the x intercept (I2)*, *Construct an equation (I3)* and *Calculate a specific function value (I7)* dominate in mathematics, whereas the strategy *Intersection with vertical axis (I6)* dominates in physics for y-intercept tasks. Here, we observe, that in mathematics, algebraic expressions are used more frequently than in kineamtics. This is in line with studies by Acevedo Nistal et al. (2012).The fact that the x-intercept was often incorrectly calculated instead of the y-intercept in mathematics may be due to the fact that only 81,5% of students were familiar with the term function value at 0 (cf. Subsection 2.2).

For the slope tasks, we see that the strategies *Location in an equation (S1)*, *Construct an equation (S3)* and *Drawing a triangle on graph (S7)* dominate in mathematics tasks and *Ratio of difference (S8)* dominates in physics tasks. This results show that students are able to read the slope-value from a function equation in mathematics but they do not see the parallels to physics. In kinematics, they use this strategy only rare (frequencies below 0.07). What is alarming here is that the most common strategy in kinematics does not lead to the correct solution. When calculating the ratio of difference, the sign was often ignored in kinematics. In contrast to Ceuppens et al. (2019), we did not observe that students reasoned with steepness in slope-tasks or used a graph to explain the answer. In a next step, we integrate results of German 11th graders.

RQ2 and RQ2.1: First, we confirm that challenges with kinematics and with negative slope remain across school years. This is alarming because it shows that school learning is not sustainable. Students learn the material in class, but cannot remember it when tested a few weeks or years later. On the other hand, the results show that many difficulties persist throughout the school years. Therefore, it is not possible to close the achievement gap across the school years.

GEE analysis revealed interesting results showing that the contextual performance-gap increases drastically among school years. Encouragingly, performance in mathematics improved by 110% from $0.29 \pm 0.16(SD)$ in grade 9 to $0.61 \pm 0.13(SD)$ in grade 11. In kinematics, on the other hand, there was only a slight improvement (of 61%) from $0.23 \pm 0.24(SD)$ in grade 9 to $0.37 \pm 0.27(SD)$ in grade 11.

Concerning solution strategies, we did see same patterns for 11th than for 9th graders.

Finally, another reason should be mentioned that could have led to problems, particularly in connection with the kinematics context. Ceuppens' test instrument has several insignificant weaknesses in this respect:

- Text: In the task description of the physics items, the definition of the variables t and x are missing.
- Formulas: Units are missing in all formulas in the physics context.
- Content critique on IP 4: Solution of this task is unclear and depends on the selected reference system: Both objects can be located either in the same or in different reference systems. In the case where both position curves are in the same reference system, the magnitudes of the velocities of the two objects are compared. For example two cars with velocities -4m/s (car A) and -6m/s (car B) are in the same reference system. Since both cars drive in the same direction, the sign can be neglected. Then, car B moves backward faster than car A. For the case that both locus curves of

the objects are located in different reference systems, the velocities of the two objects are compared independently of each other. Here, the signs of the velocities are relevant and and must be considered when comparing. In this case, $-4\text{m/s} > -6\text{m/s}$, and therefore, car A is faster.

- Ceuppens et al. (2019) describe the item pairs of the test instrument as isomorphic. However, there is no definition or well-founded characterization of the term *isomorphic*. In this work, we use the somewhat weaker term *similarity* instead of the strong term *isomorphism* to make clear that there are mainly two differences within the item pairs that are worth discussing: First, the item pairs differ in their surface features, e.g., in the functional equation, the linear term forms the first summand in mathematics and the second summand in physics. Second, the necessary interpretation in the physical context always adds an additional difficulty to the physical problems, which has no counterpart in the mathematical context.

RQ3 and 4: The significant differences in TVD demonstrate that the position-axis label receives more visual attention for the kinematic item. In contrast to that, for the mathematical item, it is the graph section. This indicates that information for interpreting the diagram is increasingly extracted from the function graph in the mathematical context, whereas in the kinematics context information is more likely to be extracted from the axis labeling of the position axis. This is also supported by the results of the transition analysis: In the kinematic item, students perform significantly more transitions between the axis labels and other axis content, which indicates that for the interpretation of the diagram, an attempt is made to integrate information between the axis sections. However, it should be noted, that the items are similar but not congruent, so that they differ minimally in individual graphic design elements such as the different spacing of the lines of the coordinate grid. Thus, it cannot be excluded that such graphical design elements automatically attract a subject's gaze focus and contribute to differences in gaze behavior (see e. g., Madsen et al. (2012)). To elucidate the causes of differences in gaze behavior, the ET analyses are now interpreted with the aid of the findings from the RTA analysis.

Remarkably more students state that they used the quotient of the axis intercepts to solve the kinematics item, whereas a slope triangle was used notably more often to solve the mathematics item. Synergizing the results of the ET analysis and the RTA analysis, the following picture emerges: When faced with the task of determining the slope of a linear function in a purely mathematical context, students are more likely to use a slope triangle as a solution strategy compared to the kinematics context, which is indicated in the gaze data by higher visual attention time to the graph. For forming a slope triangle, it is necessary to look at the graph at least twice. In the similar kinematics item, on the other hand, students more often try to determine the slope of the function graph using the quotient of the axis intercepts (position and time). The results of the gaze data analysis show that for the kinematics item, learners more often link information along the position and time axes, as well as between the two axes and the graph. This indicates that the information to form the quotient from the axis intercepts is not done by systematically changing gaze between the axis intercepts. Learners tend to first extract information about position and time and then, based on that, form the quotient.

RQ5: Cognitive processes in the interpretation of functional equations can be clarified with the help of gaze data. The cognitive processes of extracting, selecting and integrating information associated with the different solution strategies used are reflected in the ET metrics examined. The distribution of visual attention across the range of formulas reflects the most frequently selected solution strategies given by the students in text and audio form. Solution strategies can thus be differentiated from one another on the basis of objective, quantitative data.

RQ6: Considering the contextual differences in gaze data from the same formal representations in different contexts, we showed that solution strategies for slope determination in linear functional equations in mathematics and physics are context-dependent. In mathematics, the slope value is read from the functional equation, whereas in physics, the coefficients of the linear functional equation are linked by addition, subtraction, multiplication or division. A comparison of the verbally expressed solution strategies shows that the correct solution strategy in mathematics is often lost when transferred to kinematics.

These results are consistent with those of Planinic et al. Green et al. (2007), Van den Eynde et al. Van den Eynde et al. (2019), Ivanjek et al. Ivanjek et al. (2016), and Carli et al. (2020), who also identified context-dependent strategies in mathematics and physics by analyzing students' results. (Context-dependend) strategies were observed using ET data (Susac, Bubic, Kaponja, et al. (2014)). Transfer difficulties in mathematics and physics were observed by Carli et al. (2020), Becker, Knippertz, et al. (2023) and Ivanjek et al. (2016). Considering that the physics item was solved only by few students, whereas more than half of the students solved the mathematics item correctly, we are in line with studies showing that math problems are more often solved correctly than physics problems Planinic et al. (2012), Van den Eynde et al. (2019), and Carli et al. (2020). Also, this is according to results by Hegarty et al. (1995). They showed that unsuccessful students focus on numbers and relational terms, while successful students have a higher number of fixations on variables, suggesting the use of a problem modeling strategy. Merkley, Ansari (2010) showed that participants had more fixations on the correct number than on the incorrect number in mathematics. This is consent with results of mathematics item M12, showing a peak of attention on the correct slope value of the equation.

The students tend to see the physical equation as a mathematical problem. This is reflected in our results by the fact that linking the coefficients as a sum, product, etc. is the most frequently chosen solution strategy in the interviews and written statements of the students. These outcomes show that the learners apply any operators to the task and thus somehow mathematize, but lose sight of the task goal. Item K12 requires not only adequate mathematization steps (the physical relationship between position and time corresponds mathematically to a linear functional equation, reading the slope value from this equation) but also interpretation steps: The slope must be recognized as velocity and the negative sign of the slope must be interpreted physically. We see, that both the mathematization and interpretation steps are unsuccessful. One reason for this could be that students make a strict distinction between mathematical calculations and physical meaning. As Carli et al. (2020) have already noted, skills in mathematization need to be practiced in order to avoid these errors. It is worrying that these problems are already present at school. Presumably, they also endanger areas outside of school. A lack of understanding of mathematical relationships weakens the ability to recognize false information and impairs the reliability of scientific studies. It can compromise trust in science.

Regarding the triangulation of ET with interview data and written responses, we confirmed that this method is suitable to analyze context-dependent solution strategies.

RQ7: With regard to the investigation of reading effects in equations (RQ7), we were able to show that equations are not read like text with the help of a gaze transition analysis between the elements of linear functional equations in mathematics and physics. The analysis of gaze changes with and without consideration of the reading direction provides identical results. Exemplarely gazeplots showed that text is usually read from left to right, but term structures in equations are linked to each other and to the text. Fixations are denser distributed in equations than in text, i.e., equations have a higher information density than text. This is according to research by Kohlhase et al. (2018), who found first that equations in mathematical textbooks are much denser in content than text and second, that backjumps in gazeplots from equation to text appear frequently. Malone et al. (2020) showed that mean visit duration was lower for text than for equations when comparing different kinds of representations in mathematics. In contrast to our results, research from Schneider, Maruyama, et al. (2012) showed that equations are mainly read like text. Similarly, research from Jansen et al. (2007) disagree with our results, since they showed that equations are read like text.

Concerning the replication study, we confirmed challenges in kinematics and with negative slopes diagnosed by Ceuppens et al. (2019) for Belgian 9th graders for German 9th graders. Thus, although the mathematical skills are present, as evidenced by the higher probabilities of solving the math items, the students fail to transfer the existing mathematical solution strategies to the kinematics context. This is contrary to a widely held belief among physics teachers that one of the main difficulties students have in interpreting diagrams in a physics context is due to their lack of mathematical knowledge and skills. Difficulties in interpreting kinematic diagrams were thereby demonstrated for both Belgian and German students, so it seems to be a fundamental problem that goes beyond the borders of national school systems. It was also shown that these difficulties persist across the grades. It is striking that the German 9th graders performed significantly worse than the Belgian 9th graders. This difference shows that the Ceuppens study cannot simply be transferred to other countries. The basic problems remain, but are more serious in Germany. This result is alarming and shows that kinematics teaching needs to be improved in the sense that the content, especially the negative velocities, must be taught in a more sustainable way. The teacher training we have developed provides approaches for implementation. The German school system, with its canon of subjects and the fact that many subjects are taught every 45 minutes, certainly represents a hurdle here. The extent to which interdisciplinary, project-based teaching over longer periods of time is more beneficial could be investigated in the future.

Especially tasks dealing with linear graphs with negative slope in the kinematic context are challenging for 9th and 11th grade students. This was the case both when graphs were to be compared with each other and when the slope was to be determined from a given graph or from a given formula. This indicates that the students find it particularly difficult to correctly interpret the meaning of slope in a kinematic context and, accordingly, to apply the necessary mathematical solution procedure in this context.

It should be noted, however, that the study results refer only to a specific physical context, kinematics, and a specific type of function, linear functions. It is worth investigating whether the failure to transfer mathematical solution procedures also turns out to be a major cause of error in other contexts and forms of representation. Beyond physics, this could also include the other natural sciences, since in biology or chemistry the degree of mathematization is comparable to physics. Furthermore, it would be interesting in this context to investigate empirically to what extent teacher training has an influence on the use of mathematical solution procedures; for example, do teachers who teach both mathematics and physics succeed better in linking mathematics and physics to solve physical problems?

A further discussion concerns the statistical analysis of the paper-based study. In their study of linear graphs, Ceuppens et al. (2019) choose Cohen's kappa as the statistical measure to measure the agreement between two raters on a category system. However, the categories in this category system are not disjoint. This violates a condition for calculating Cohen's kappa. We were able to correct this by using Kraemer's extended kappa coefficient as an appropriate measure of agreement. The qualitative analysis in the study of Ceuppens et al. (2019) is statistically imprecise because only the material from 40 students was coded and not the entire material. We coded all the material according to Mayring's QIA, but this was not optimal: several rounds of coding and a refinement of the coding guide had to be omitted due to the amount of work involved. In retrospect, it is clear that refining the coding guide would have been a sensible option. This would have avoided many inconsistencies in the coding. Therefore, it would have been better to select the maximum amount of data to analyse without sacrificing multiple coding runs and refinements in scope and timeframe.

Since the test-based data (correct/incorrect) disallowed us to draw conclusions about the solution strategies used, we investigated the gaze patterns during the solution process using ET. For this purpose, we selected the similar IP in which the slope of the graph of a linear function with negative slope should be determined. On the one hand, the mathematics item was solved correctly more often than the kinematics item, and on the other hand, the different solution strategies in this task format of quantitative slope determination should be reflected in the gaze patterns (for example, the formation of a slope triangle).

Using gaze plots, it was already clear on an individual level that the gaze movements between mathematical and kinematic items differed significantly. Thus, in the mathematical context, the gaze was increasingly along the axes, whereas in the kinematic context, more attention was paid to the axes intercepts. A significantly different gaze behavior could also be demonstrated for the entirety of the subjects by analyzing the ET metrics TVD and transistions. It could be shown that depending on the context, different diagram areas are more strongly focused and visually linked with each other. This suggests different strategies for interpreting graphs in mathematical and kinematic contexts, which is also a possible explanation for the much lower solution probability for the kinematic item.

To better interpret these findings, we performed a supplemental triangulation of the ET data with gaze cued RTA interviews. The analysis indicates that, in the mathematical context, the solution strategy of the slope triangle is often applied, whereas, in the kinematic context, the attempt is mostly made to form the quotient of position and time. In the kinematic context, this leads indicates that the negative sign of the velocity is not considered and the task is thus solved incorrectly. The reason for this could be the physics lessons, in which an algorithmic procedure for calculating the velocity from the position and time data is taught, so that the application of mathematical procedures is no longer considered.

To prevent this problem, when introducing velocity, teachers could define it both as the quotient of distance traveled and time elapsed and as the slope of the graph in the time-position diagram, and also train them to flexibly fall back on the appropriate definition and thus solution strategy depending on the problem. This multi-perspective approach to velocity could also promote an understanding of negative velocity, which is common in physics but rather unfamiliar from the students' lifeworld.

Concerning our results on graph- and formula-tasks, our results contain several suggestions for future instructions in physics and other areas that use domain-specific function equations and graphs. For instance, to avoid difficulties in transfer from math to other disciplines, we suggest to explicitly highlight the correspondences and relations between formulas and graphs in math and in other disciplines at the moment they are domain-specific introduced. For instance, this could be facilitated via contrasting cases, as it allows to directly compare commonalities and differences between equations/graphs in different domains and may facilitate the transfer solution strategies (Schwartz et al. (2011) and Kuo, Wieman (2016)). Apart from that, we showed that the total dwell time on AOIs that contain the slope in the math context as well as transition from and to these AOIs have been most pronounced in the math context. To encourage students to perform similar eye movements when learning or solving problems with formulas and graphs in other domains, it could be helpful to make efficient visual processing of formulas and graphs explicit. This can be achieved, for example, by presenting eye-movement modeling examples, which show the eye movements of experts demonstrating efficient formula/graph processing (Gegenfurtner et al. (2017)). Similarly, it may be helpful to explain efficient cognitive procedures to process and evaluate information from equations and graphs to students.

On the other hand, there are some limitations of our work. One limitation is the choice of items from a validated test instrument that was not explicitly developed for ET research. As a result, the items are unevenly structured in mathematics and physics in terms of their surface features. For example, in the functional equation, the linear term forms the first summand in mathematics and the second summand in physics, and in the physics item, the functional relationship between place and time was not called $x(t)$, but only x. This means that the functional relationship between position x and time t is not clear. In addition, the equation in physics is shown completely without units. Unfortunately, there is no comparable validated test instrument specifically for ET research. In the future, attention should be paid to a comparable surface structure when creating test instruments so that they can also be used adequately in ET research. Existing test instruments should be revised and revalidated accordingly.

A teacher training course on transfer difficulties was offered. The focus was on research-related transfer difficulties from mathematics to physics. Four secondary school teachers from Rhineland-Palatinate took part in the training course. The aim of the training was to raise awareness of the topic of transfer difficulties and practical applications for their own lessons, combined with concrete tips on how to adequately address transfer difficulties experienced by learners. In addition to theoretical foundations, the content of the training included research findings and numerous exercises on transfer difficulties with functional contexts, concepts and units. The feedback from the participant survey regarding the benefits, enjoyment and added value of the training course was extremely positive (three out of four participants rated the training course as very good overall, one as good).

The training represents a first attempt to successfully transfer scientific findings to schools in a practical way. The training was well received by the teachers. A limiting factor was the small number of participants—only four people. This limits the significance of the quality of the training enormously. It is therefore planned to test the training again for a larger target group and evaluate it on this basis. Depending on the size of the target group, a well-founded empirical evaluation can then also be considered. As all the training material is available online in German and English (cf. Knippertz (2024c)), the course can also be used by other people in the future.

4. Basic Research Into Potential Applications in Adaptive Learning Systems

This chapter presents two possible features for ALS supporting students solving tasks on linear functions in mathematics and physics using ET data and network analysis methods. In the first section, we examine how geometric structures representing solution strategies can be recognized from ET data. For this, we use special kind of networks—temporal netwoks. The idea of modeling ET data with temporal networks and the recognition of geometric structures from ET data is completely new, which is why there is no literature on this topic. In the second section, we deal with network metrics distinguishing correct from incorrect solvers. In principle, there is very little literature on the analysis of ET data using mathematical networks in education.

Zhu, Feng (2015) use networks to visualize and analyze gaze data for tasks with linear functions. The networks represent transitions from ET data. Network properties and three-node structures are used for the analysis. On the one hand, their results show that transitions between AOIs are not random. Second, that students tend to revisit previously visited AOIs.

In a study by Schneider, Pea (2014), collaborative learning is examined with the help of ET and network analysis. Here, networks are represented in such a way that nodes represent fixations and edges represent saccades. In addition, network features and machine learning algorithms are used to predict the quality of collaboration.

Outside of education, network analysis in connection with ET data is also used in other areas, for example in authismn and consumer research (Guillon et al. (2015) and Maslowska et al. (2020)).

4.1. Algorithm for Detection of Geometric Structures From Gaze Data Using Network Analysis Methods

We believe that the integration of sequential information in network models generated from ET data and the detection of geometric structures in gaze data have not been explored. In this section, we present a method to model ET data as a temporal network. The model and its static representations, specifically weighted time-aggregated and kth order time-aggregated networks, are investigated. In addition, a procedure for detecting geometric structures in ET data is developed. The procedure combines the use of kth order time-aggregated networks and geometrical criteria. It is tested on ET recordings of students solving a mathematical task. The results show that the detection of geometric structures, namely slope triangles, is feasible. The results provide a basis for later integration into ALS to enhance its diagnostic capabilities. The Python code for this section can be found in Knippertz (2024b).

How to Construct a Temporal Network Based on ET-Data

This section descibes the construction of temporal networks by the use of ET data. Let us assume that we already have ET data clustered in AOIs. In a first step, all AOIs are identified with the natural numbers $0, \dots, n-1$. We assume that AOIs do not overlap. Then, we intoduce a node i in the temporal network for each AOI i. For fixations located outside of the AOIs, the node -1 is added. The weight function $\omega_1^V : V \mapsto \mathbb{N}$ assigns to each node the total duration of fixations in the corresponding AOI in milliseconds. Moreover, a second weight function $\omega_2^V : V \mapsto \mathbb{R}_{>0}$ is added that assigns to each node the total number of fixations in the corresponding AOI. An arc $(i, j, t) \in V \times V \times \{1, \dots, T\}, i \neq j$ is added if the fixation was located inside AOI j immediately after it was located in AOI i and the transition from i to j is the t-th time two different AOIs were fixated consecutively. Therefore, it is possible to reconstruct the order in which the student looked at the AOIs from the set E^T. A temporal network that is created as described above is called an *ET network*.

Example 4.1. Table 4.1 consists of sample ET data, that gets transformed into an ET network. The first column consists of indices of eleven fixations. Thus, each row describes one fixation. The second column contains the duration of each fixation. There is one column for each AOI and each color represents an AOI. If the ith fixation is located inside the AOI j, the corresponding entry in row i and the column of AOI j is one, otherwise the entry is set to be zero. For each of the four AOIs, a node is introduced. In row three of Table 4.1, a different AOI than AOI 0 is fixated for the first time. Therefore, the edge $(0, 2, 1)$ is added to the set E^T. In the fourth row, we see that the next fixation is located inside AOI 1. Thus, the arc $(2, 1, 2)$ is added. One can proceed analogously to obtain the temporal network $G^T = (V, E^T)$ with $V = \{0, 1, 2, 3\}, T = 7$ and

$$E^T = \{(0, 2, 1), (2, 1, 2), (1, 3, 3), (3, 2, 4), (2, 1, 5), (1, 2, 6), (2, 3, 7)\}.$$

The first node weight ω_1^V of node 0 is calculated by summing up the two durations 117 and 138. We obtain $\omega_1^V(0) = 255$. Since there are two ones in the column

Table 4.1.: Sample ET data showing fixation index, fixation duration in ms and AOI hits from AOI 0 to 3.

Fixation index	Fixation duration in ms	AOI 0	AOI 1	AOI 2	AOI 3
1	117	1	0	0	0
2	138	1	0	0	0
3	212	0	0	1	0
4	142	0	1	0	0
5	233	0	1	0	0
6	75	0	0	0	1
7	224	0	0	1	0
8	97	0	1	0	0
9	172	0	1	0	0
10	82	0	0	1	0
11	122	0	0	0	1

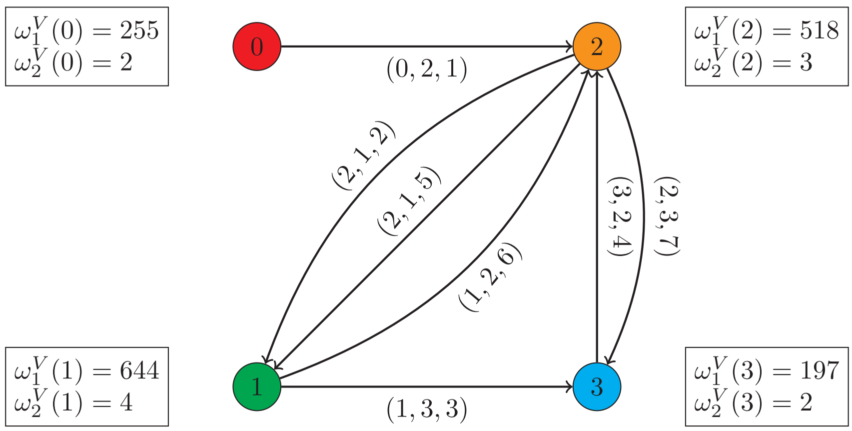


Figure 4.1.: Visualization of the temporal network based on sample data in Table 4.1. An alternative visualization of this network can be found in Section A.3 in the Appendix.

AOI 0, the node weight ω_2^V of 0 is set to be two. The further node weights and a visualization of G^T are displayed in Figure 4.1.

The implementation of the procedure described in Example 4.1 is described as peseudocode presented in Algorithm 1. The input is a list L of dictionaries that is structured as in Table 4.1. That means, every row in Table 4.1 is represented by a dictionary that contains the keys fixation index, fixation duration and AOI i for $i = 0, \ldots, n-1$.

We now proof that Algorithm 1 has runtime of $\mathcal{O}(h \cdot n)$, where h denotes the length of the input list.

Algorithm 1: BUILD ET-NETWORK

Input: List L of dictionaries, such that each dictionary contains the keys fixation index, fixation duration and AOI i for $i = 0, \dots, n-1$
Output: ET-network $G^T = (V, E^T)$
Data: Dataset structured as in Table 4.1

```
/* Initialization                                                */
1  V := ∅
2  E^T := ∅
3  previous node := None
4  edge count := 1
5  ω_1^V(i) = 0 for i = 0, ..., n − 1
6  ω_2^V(i) = 0 for i = 0, ..., n − 1
7  foreach dictionary d in L do
8      found fixation = FALSE
9      for i = 0, ... n − 1 do
10         if d[AOIi] = 1 then
11             found fixation = TRUE
12             V = V ∪ {i}
13             ω_2^V(i) = ω_2^V(i) + 1
14             ω_1^V(i) = ω_1^V(i) + d[fixationduration]
15             if previous node ≠ i then
16                 if previous node ≠ None then
17                     E^T = E^T ∪ {previous node, i, edge count}
18                     edge count = edge count +1
19                 previous node = i
       /* Fixation is outside of the AOI-allocation                  */
20     if found fixation = FALSE then
21         if previous node ≠None then
22             E^T = E^T∪(previous node,−1,edge count)
23             edge count = edge count +1
24         previous node = −1
25 return G^T = (V, E^T)
```

Theorem 4.2. *Runtime of Algorithm 1*
Let h be the length of the input list and n be the number of AOIs. Algorithm 1 builds an ET-network in $\mathcal{O}(h \cdot n)$.

Proof. The initialization of the algorithm is done in $\mathcal{O}(n)$, because of the initialization of the node weights. Since h is the length of the input list, the for-loop in line 7 is executed h times. Furthermore, the for-loop in line 9 is executed

n times. Adding the nodes to the set V, increasing the values of the weight functions, updating the previous node variable, increasing the edge counter and adding an edge to the set E^T can be done in constant time. In conclusion, the ET-network is built in $\mathcal{O}(h \cdot n)$. □

A static representation is a graph that is derived from the temporal network with the aim of preserving the additional information provided by the temporal network. The advantage of static representations is that common graph theoretic procedures can easily be applied while obtaining some of the temporal information.

We now want to introduce two static representations of a temporal network: weighted time-aggregated networks and kth order time aggregated networks. In addition, we consider upper bounds on the number of nodes and edges in a kth order time-aggregated network and learn how to transform a temporal network into a kth order time-aggregated network. The runtimes of the algorithms presented are also covered.

The first static representation is the weighted time-aggregated network. Here, the arcs between two nodes u and v are aggregated into one edge (u, v). The number of arcs that connect the nodes u and v is specified in the weight function. The definition is inspired by Scholtes et al. (2016).

Definition 4.3. *(Weighted) Time-Aggregated Network*
Let $G^T = (V, E^T)$ be a temporal network. The tuple $G = (V, E)$ is called the *time-aggregated network* of G^T, if $(v, w) \in E$ if and only if $(v, w, t) \in E^T$ for some $t \in \{1, \ldots, T\}$. The network $G = (V, E)$ is called *weighted time-aggregated* if there is a weight funtion $\omega^E : E \mapsto \mathbb{N}$ and

$$\omega^E((v, w)) = |\{t \in \{1, \ldots, T\} : (v, w, t) \in E^T\}|.$$

Example 4.4. Figure 4.2 shows the weighted time-aggregated network based on the ET-network in Example 4.1. The time-aggregated representation has the same set of nodes as the corresponding temporal network. Since the transition from 2 to 1 occurs twice in the temporal network, the weight of the edge $(2, 1)$ in the weighted time-aggregated network is set to be two.

Even though it is known how many times a transition was made, all information about the ordering of the edges is missing. Consequently, the weighted time-aggregated network is a rather simple static representation, in which the temporal aspect is almost completely lost.

To preserve some information about the ordering of the edges in the time-aggregated network, the definition of a time-respecting path is needed. The following definition of a time-respecting path is a specified version of the definition by Scholtes et al. (2016).

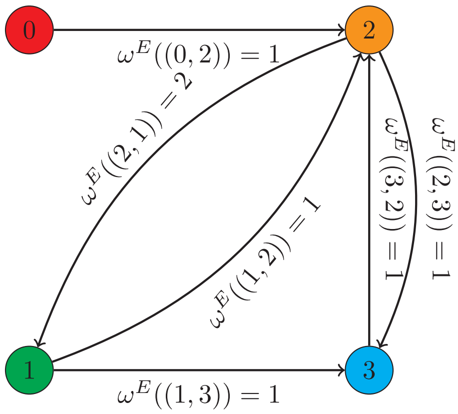


Figure 4.2.: Weighted time-aggregated network based on the temporal network in Example 4.1. The labels of the arcs are their ω^E-values.

Definition 4.5. *Time-Respecting Path*
Let $G^T = (V, E^T)$ be a temporal network. A sequence of arcs

$$(v_1, v_2, t_1), (v_2, v_3, t+1) \ldots, (v_k, v_{k+1}, t+k-1),$$

is called a *time-respecting path* of length k. The node v_1 is called the source node and v_{k+1} is called the target node. A time-respecting path is called *time-respecting circle* if the source node is equal to the target node.

Lemma 4.6. *Let $G^T = (V, E^T)$ be a temporal network and $G = (V, E)$ be the derived time-aggregated network. If the temporal network contains the time-respecting path*

$$(v_1, v_2, t_1), \ldots, (v_k, v_{k+1}, t_k),$$

the path $v_1, (v_1, v_2), \ldots, (v_k, v_{k+1})v_{k+1}$ is contained in G.

Proof. By definition of a time-aggregated network, it holds that $(v_i, v_{i+1}) \in E$ for $i = 1, \ldots, k$. Hence, the path $v_1, (v_1, v_2), \ldots, (v_k, v_{k+1})v_{k+1}$ is contained in G. □

The following example shows that a path $v_1, (v_1, v_2), \ldots, (v_k, v_{k+1})v_{k+1}$ in the time-aggregated network does not necessarily imply a time-respecting path with source node v_1 and target node v_{k+1}.

Example 4.7. Observe the temporal network G^T with $V = \{0, 1, 2\}$ and $E^T = \{(0, 1, 1), (1, 2, 2), (2, 0, 3)\}$ and the corresponding weighted time-aggregated network G with $V = \{0, 1, 2\}$ and $E = \{(0, 1), (1, 2), (2, 0)\}$. The time-aggregated network contains the path $2, (2, 0), 0, (0, 1), 1$, whereas G^T does not contain a time-respecting path with source node 2 and target node 1.

To obtain more information about the order of the edges than in a time-aggregated network, an extension to kth order time-aggregated networks is required. The definition is based on Scholtes et al. (2016) and extended by the condition marked by $(\star)$.

Definition 4.8. *kth Order Time-Aggregated Network*
Let $G^T = (V, E^T)$ be a temporal network. For $k \geq 2$, a *kth order time-aggregated network* is a tuple $G^{(k)} = (V^{(k)}, E^{(k)})$ with $V^{(k)} \subseteq V^k := V \times V \cdots \times V$ and $E^{(k)} \subseteq V^{(k)} \times V^{(k)}$, where $v = (v_1, \ldots, v_k) \in V^{(k)}$ is an element of $V^{(k)}$ if there is a time-respecting path

$$(v_1, v_2, t_1), \ldots, (v_{k-1}, v_k, t_{k-1})$$

of length $k-1$ in G^T. Two nodes $u = (u_1, \ldots, u_k)$ and $v = (v_1, \ldots, v_k)$ are connected by an edge $(u, v) \in E^{(k)}$ if the following two conditions hold:

- $v_i = u_{i+1}$ for $i = 1, \ldots, k-1$ and
- $\exists t_1, \ldots, t_k$ such that $(u_1, u_2 = v_1, t_1), \ldots, (u_k = v_{k-1}, v_k, t_k)$ is a time-respecting path of length k in G^T.$(\star)$

Note that a node in the kth order time-aggregated network represents a time-respecting path of length $k-1$ and an edge in the kth order time-aggregated network represents a time-respecting path of length k.

Example 4.9. Figure 4.3 shows the second order time-aggregated network derived from the temporal network in Example 4.1.

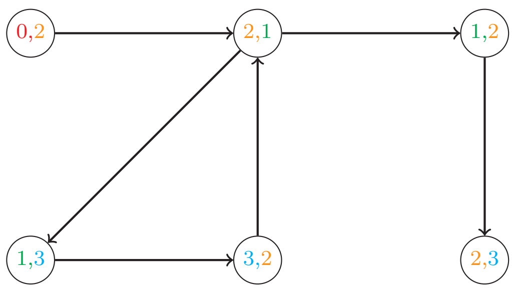


Figure 4.3.: Second order time-aggregated network based on the temporal network in Example 4.1.

Although more information on the numbering of the edges is obtained, it is not always possible to reconstruct the set E^T based on a kth order time-aggregated network.

Example 4.10. The edge $((2,3),(3,2))$ is added to the second order time-aggregated network displayed in Figure 4.3. Observe the node $(2,1)$ in the second order time-aggregated network displayed in Figure 4.4. Based on this network alone, one does not know if the time-respecting path $(0, 2, t_1), (2, 1, t_1 + 1), (1, 2, t_1 + 2)$ exists.

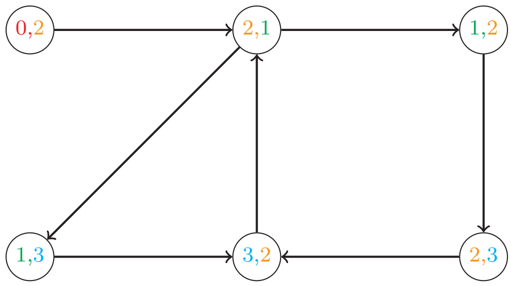


Figure 4.4.: Second order time-aggregated network based on the temporal network in Example 4.1 with the additional edge $((2,3),(3,2))$

In the following, we derive upper bounds on the number of nodes and edges in a kth order time-aggregated network based on the number of nodes and edges in the corresponding temporal network.

Theorem 4.11. *Let $G^T = (V, E^T)$ be a temporal network with $|V| = n$. Then, $n \cdot (n-1)^{k-1}$ is an upper bound on the number of nodes and $n \cdot (n-1)^k$ is an upper bound on the number of edges in the corresponding kth order time-aggregated network.*

Proof. Let $v = (v_1, \dots v_k)$ be a node in the kth order time-aggregated network. There are n possibilities for the first entry of v. As the temporal network has no loops, it holds that $v_i \neq v_{i+1}$ for $i = 1, \dots, k-1$, and therefore, there are $n-1$ possibilities for each of the remaining $k-1$ entries. Thus, in total, there are $n \cdot (n-1)^{k-1}$ possibilities to obtain a k-tuple with $v_i \neq v_{i+1}$ for $i = 1, \dots, k-1$ out of a set of n elements. Hence, $n \cdot (n-1)^{k-1}$ is an upper bound on the number of nodes in the kth order time-aggregated network.
Let $u = (u_1, \dots u_k)$ be another node in the kth order time-aggregated network. If there is an edge (u, v) in the kth order time-aggregated network, it holds that $v_i = u_{i+1}$ for $i = 1, \dots, k-1$. As $v_k \neq v_{k-1}$, there are $n-1$ possibilities for the last entry of v, and thus, the node u can at most be connected to $n-1$ nodes. Consequently, since $n \cdot (n-1)^{k-1}$ is an upper bound on the number of nodes, $n \cdot (n-1)^{k-1} \cdot (n-1) = n \cdot (n-1)^k$ is an upper bound on the number of edges. □

Theorem 4.12. *Let $G^T = (V, E^T)$ be a temporal network. Then, max$\{T-k+2, 0\}$ is an upper bound on the number of nodes and $T-k+1$ is an upper bound on the number of edges in the kth order time-aggregated network.*

Proof. Observe that $k \leq T+1$, as the nodes in the kth order time-aggregated networks are time-respecting paths of length $k-1$ in the temporal network. Let $E^T = \{(v_1, v_2, 1), \dots, (v_T, v_{T+1}, T)\}$ be the set of arcs. Then, the sequence $(v_i, v_{i+1}, i), \dots, (v_{i+k-2}, v_{i+k-1}, , i+k-2)$ is a time-respecting path of length $k-1$ for $i = 1, \dots, T-k+2$. Thus, $T-k+2$ is an upper bound on the number of nodes in the kth order time-aggregated network.

Furthermore, the sequence $(v_i, v_{i+1}, i), \dots, (v_{i+k-1}, v_{i+k},, i + k - 1)$ is a time-respecting path of length k for $i = 1, \dots, T - k - 1$. Since there is an edge in the kth order time-aggregated network for each time-respecting path of length k in the temporal network, $T - k + 1$ is an upper bound on the number of edges in the kth order time-aggregated network. □

Corollary 4.13. *Let $G^T = (V, E^T)$ be a temporal network with $V = |n|$. then, $min\{n \cdot (n-1)^{k-1}, T - k + 2\}$ is an upper bound on the number of nodes and $min\{n \cdot (n-1)^k, T - k + 1\}$ is an upper bound on the number of edges in the kth order time-aggregated network.*

Proof. The claim follows directly from Theorem 4.11 and Theorem 4.12. □

The procedure of deriving a kth order time-aggregated network from a temporal network is described in Algorithm 2.

Algorithm 2: *DERIVE kTH ORDER TIME-AGGREGATED NETWORK*

Input: Temporal network $G^T = (V, E^T)$ and order k
Output: kth order time-aggregated network $G^{(k)} = (V^{(k)}, E^{(k)})$
Data: Dataset structured as in Table 4.1

/* Initialization */
$V^{(k)} := \emptyset$
$E^{(k)} := \emptyset$
previous node = None
Let $E^T = \{(v_1, v_2, 1), \dots, (v_T, v_{T+1}, T)\}$ be the set of arcs.
for $i = 1, \dots, T - k$ **do**
  $v = (v_i)$
  for $j = 1, \dots, k - 1$ **do**
    Add v_{i+j} to the tuple v
  if *previous node* $\neq$ *None* **then**
    $E^{(k)} = E^{(k)} \cup \{(\text{previous node}, (v_i, v_{i+1}, \dots, v_{i+k-1}))\}$
  previous node = v
return $G^{(k)} = (V^{(k)}, E^{(k)})$

Theorem 4.14. *Algorithm 2 derives a kth order time-aggregated network from a temporal network G^T in $\mathcal{O}((T - k) \cdot k)$.*

Proof. The outer for-loop is executed $T - k$ times and the inner for-loop is executed $k - 1$ times. Updating the previous node, adding the nodes to $V^{(k)}$ and the edges to $E^{(k)}$ is done in constant time. In conclusion, the kth order time-aggregated network is derived in $\mathcal{O}((T - k) \cdot k)$. □

In summary, this section has introduced temporal networks and two static representations, namely weighted time-aggregated networks and kth order time-aggregated networks. The weighted time-aggregated network is a rather simple static representation where the information about the numbering of the edges is lost. However, with respect to the ET network, the derived weighted time-aggregated network captures the number of transitions between two AOIs. This can be useful information for ET measures. Compared to the weighted time-aggregated network, the kth order time-aggregated network preserves more information about the numbering of the edges and is therefore more useful for the detection of geometrical structures in gaze data.

Eye-Tracking Measures and Networks

The goal of this section is to relate existing ET measures to network properties. In particular, the ET network and the weighting functions of the ET network will be considered. Since we are only looking at ET data that has been collected in an educational context, we will look at ET measures that have been used in this area. All measures mentioned below are collected in Strohmaier et al. (2020) and Lai et al. (2013).

As stated by Lai et al. (2013), ET measures can be divided into three categories regarding the scale of measurement. The first category focuses on temporal aspects and includes measures that assess the duration of fixations, saccades, or both. Consider the ET network presented in Section 4.1. By defining the node weight function ω_1^V in Section 4.1, the total fixation duration of AOI i is given by $\omega_1^V(i)$. If the whole object is covered by AOIs and no two AOIs overlap, then each time the eye tracker detects a fixation, it is in exactly one AOI. Therefore, the total fixation duration of the item can be calculated by adding the fixation durations of each AOI

$$\omega_1(V) := \sum_{i \in V} \omega_1^V(i).$$

Due to different AOI sizes or widely varying gaze durations, it may be useful to consider the fixation duration relative to the total fixation duration. The relative fixation duration can then be calculated by dividing the fixation duration of AOI i by the total fixation duration

$$\frac{\omega_1^V(i)}{\omega_1}.$$

The fixation duration relative to the total fixation duration in a subset $V' \subset V$ can be calculated analogously. This is done by dividing the fixation duration of the AOI by the sum $\omega_1(V')$. Thus, the three ET measures fixation duration, total fixation duration, and relative fixation duration can be determined using the first node weight function ω_1^V.

Table 4.2.: Temporal ET measures and their corresponding calculations.

Temporal	
fixation duration in AOI i	$\omega_1^V(i)$
total fixation duration in the item	$\sum_{j \in V} \omega_1^V(j)$ (suitable AOI-allocation)
relative fixation duration of AOI i	$\frac{\omega_1^V(i)}{\sum_{j \in V} \omega_1^V(j)}$
average fixation duration of AOI i	$\frac{\omega_1^V(i)}{\omega_2^V(i)}$
time to first fixation for AOI j	estimation $\min\{i \in V, t : (i,j,t) \in E^T\}$

Furthermore, the average fixation duration in AOI i can be calculated by dividing the fixation duration by the total number of fixations in AOI i. Since the second node weight function ω_2^V defined in Section 4.1 assigns to each node the total number of fixations in the corresponding AOI, the average fixation duration in AOI i can be calculated by dividing the value of the first weight function by the value of the second weight function.

$$\frac{\omega_1^V(i)}{\omega_2^V(i)}.$$

Another temporal measure of ET is the time to first fixation. This measure determines the time in seconds until the first fixation on the AOI. Because the ET network stores the order of the transitions, it is possible to determine after how many transitions the AOI j was fixated for the first time:

$$\min\{t \in \{1, \dots, T\}, i \in V : (i,j,t) \in E^T\}. \tag{4.1}$$

If $\{t \in \{1, \dots, T\}, i \in V : (i,j,t) \in E^T\} = \emptyset$, AOI j was either not fixated at all or AOI j was the first AOI that was fixated. Observe that (4.1) is an estimation of the time to first fixation and no exact time measurement. However, the estimate is sufficient to compare the time to first fixations of the AOIs among each other.

A summary of the temporal ET measures mentioned above and their corresponding calculations is depicted in Table 4.2.

The second category of ET measures consists of measures that count events such as fixations, saccades, or visits. By defining the second weight function ω_2^V in Subsection 4.1, the total number of fixations in AOI i is given by $\omega_2^V(i)$. Thus,

Table 4.3.: ET measures that count events and their corresponding calculations.

Count	
total number of fixations in AOI i	$\omega_2^V(i)$
total number of fixations in the item	$\sum_{i \in V} \omega_2^V(i)$
relative number of fixations in AOI i	$\frac{\omega_2^V(i)}{\sum_{j \in V} \omega_2^V(j)}$
number of visits in AOI j	$\sum_{i \in V} a_{ij} \cdot \omega^E((i,j))^\star$ (weighted time-aggregated network)
total number of transitions	T

* $\sum_{i \in V} a_{ji} \cdot \omega^E((j,i))$ if j is visited first

the total number of fixations in an item and the relative number of fixations can be computed analogously to the corresponding temporal measures.

The number of visits counts how often an AOI is entered. The weight of the edge (i,j) in the weighted time-aggregated network is defined as

$$\omega^E((i,j)) = |\{t \in \{1,\ldots,T\} : (i,j,t) \in E^T\}|$$

in Definition 4.3. Therefore, the number of visits in AOI j can be calculated by

$$\sum_{i \in V} a_{ij} \cdot \omega^E((i,j)),$$

where a_{ij} is the corresponding entry of the adjacency matrix and AOI j is not the first AOI visited. If AOI j is the first AOI visited, the number of visits can be calculated as

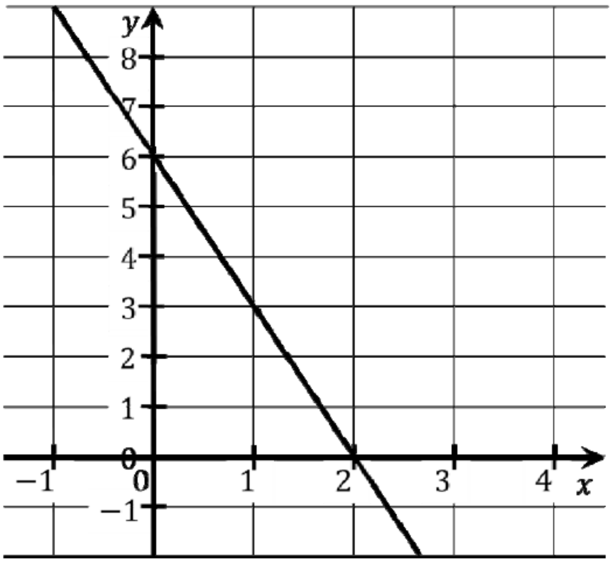
$$\sum_{i \in V} a_{ji} \cdot \omega^E((j,i)).$$

Since the transitions are numbered, the total number of transitions is T. The ET measures mentioned above and the corresponding calculations can be found in Table 4.3.

The third category consists of spatial ET measures. Since a node is denoted for each AOI and the position of the AOIs is known, fixations can be localized by the position of the AOIs.

In summary, this subsection has shown that the ET information can be transferred to the ET network in such a way that common fixation-related ET measures can be computed based on the ET network or the derived weighted time-aggregated network. The walk as a tuple might be the most compact variant, M8

The figure shows the graph a function .



Determine the slope. Explain your answer.

Figure 4.5.: Math item M8 from the validated test instrumetn by Ceuppens et al. (2019) where students had to determine the slope of the linear function.

with additional data saved separately. Thus, ET-networks can be considered as a compact and clearly structured alternative for saving ET data.

Detecting Geometrical Structures Based on Automated AOI-Allocation

In this section, we focus on the detection of geometric structures, more specifically slope triangles, in the gaze of subjects. Therefore, we present a procedure for the detection of slope triangles based on the combination of kth order time-aggregated networks and geometric criteria. We also test the algorithm on ET data and discuss the results. A slope triangle is a triangle structure in the coordinate system whose hypotenuse lies on the graph of the function and is used to determine the slope. We consider the item displayed in Figure 4.5, taken from Ceuppens et al. (2019), where subjects had to determine the slope of the depicted monotonically decreasing linear function.

Now, we will place AOIs for every subject individually. For this, we need to introduce some basic knowledge on cluster analysis based on Han et al. (2012).

The procedure of dividing sets of data elements into subsets is called *cluster analysis* or *clustering*. The resulting subsets are called *clusters* and the set of clusters is called *clustering*. Clusters consist of elements that are by some means similar to each other but dissimilar to elements from other clusters. Precisely how similarity or dissimilarity is defined depends on the type of data to be clustered and can be completely different. For example, if people are to be grouped into clusters, then similarity can be understood as character traits. Whereas, in our case, coordinates are to be grouped into clusters based on the euclidean distance.

Forming clusters is done by means of clustering algorithms. Different clustering algorithms result in different clusterings. Besides the common-known clustering algorithm *k-means*, there are several other clustering algorithms. Hence, it is needed to decide which algorithm is suitable for the given data set. The clustering algorithm based on the *mean shift procedure* presented by Fukunaga, Hostetler (1975) and extended by Comaniciu, Meer (2002) is shortly outlined in the remainder of this section. The description is based on Comaniciu, Meer (2002).

As described by Comaniciu, Meer (2002), the advantage of the mean shift clustering compared to other clustering algorithms, for example k-means, is that the number of clusters does not need to be known in advance. Since the total number of fixations to be clustered attains values from 30 to 100 and since the distribution of the fixation coordinates can vary widely, the same number of clusters is not appropriate for all subjects. Furthermore, in terms of real-time clustering and ALS, the mean shift procedure is beneficial, because, according to Comaniciu, Meer (2002), it is not computationally expensive.

Let $x_1, \ldots, x_n \in \mathbb{R}$ be the points to be clustered. The first step of the mean shift procedure consists of finding the modes of the underlying density function f of the input data. The modes of f can be found in the stationary points of f. A stationary point of a differentiable function is a point x for which it holds that $\nabla f(x) = 0$. Since the *mean shift vector*

$$m(x) = \frac{\sum_{i=1}^{n} x_i \cdot g(||\frac{x-x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x-x_i}{h}||^2)} - x \tag{4.2}$$

with *bandwidth* $h > 0$ and $G(x) = c \cdot g(||x||^2)$ for a *kernel* G is an ascent direction of f, each point x_i gets moved to the point $x_i + m(x_i)$. This is continued until convergence to a stationary point. The proof of convergence as well as further information concerning the kernel function, the bandwidth and a more detailed description can be found in Comaniciu, Meer (2002). After the mean shift procedure has been performed, only the local maxima among the stationary points are retained. For each mode, the set of all points that converge to that mode,

Table 4.4.: Excerpt of sample ET data.

Fixation index	x-coordinate	y-coordinate	AOI coordinate system
1	488	122	0
2	463	426	1
3	624	632	1
4	560	515	1

called the *basin of attraction*, gets evaluated. A cluster then consists of all points that are located in the basin of attraction of the same mode.

Certainly, the bandwidth h affects the returned clusters. As stated by Fukunaga, Hostetler (1975), the bandwidth influences how much the density is smoothed, and thus affects the exclusion of modes that are in close proximity to other modes. However, Fukunaga, Hostetler (1975) and Comaniciu, Meer (2002) agree that the selection of a suitable bandwidth h is a problem in itself.

Now, we use the clustering algorithm based on the mean shift procedure to divide the coordinates of the fixations into clusters, whereby each cluster defines an AOI.

An excerpt of sample ET data that is used in this subsection to determine the AOI-allocation is depicted in Table 4.4.

The first column consists of indices of 4 fixations. Hence, each row describes one fixation. The second and third column contain the x- and y-coordinate of the corresponding fixation in pixel, respectively. The fixations that are located outside must be eliminated, since only the fixations that are found in the coordinate system are of interest for detecting geometrical structures. Keeping these fixations would disturb the clusters in the coordinate system. Therefore, an AOI is placed on the whole coordinate system. Hence, the last column indicates whether the fixation was located inside the coordinate system or not. If the fixation is located inside the coordinate system, the entry in the corresponding row is one, otherwise the entry is zero.

The clustering algorithm based on the mean shift procedure is then applied to the coordinates (x_i, y_i) of the fixations located in the coordinate system. For this, the scikit implementation of the clustering algorithm by Pedregosa et al. (2011) is used. After each coordinate is assigned a cluster, the AOI-allocation is obtained as each cluster defines an AOI.

As mentioned above, selecting a suitable bandwidth h is a problem in itself and also depends on the further usage of the clusters. The smaller the bandwidth is chosen, the more clusters are built and therefore the more AOIs are defined.

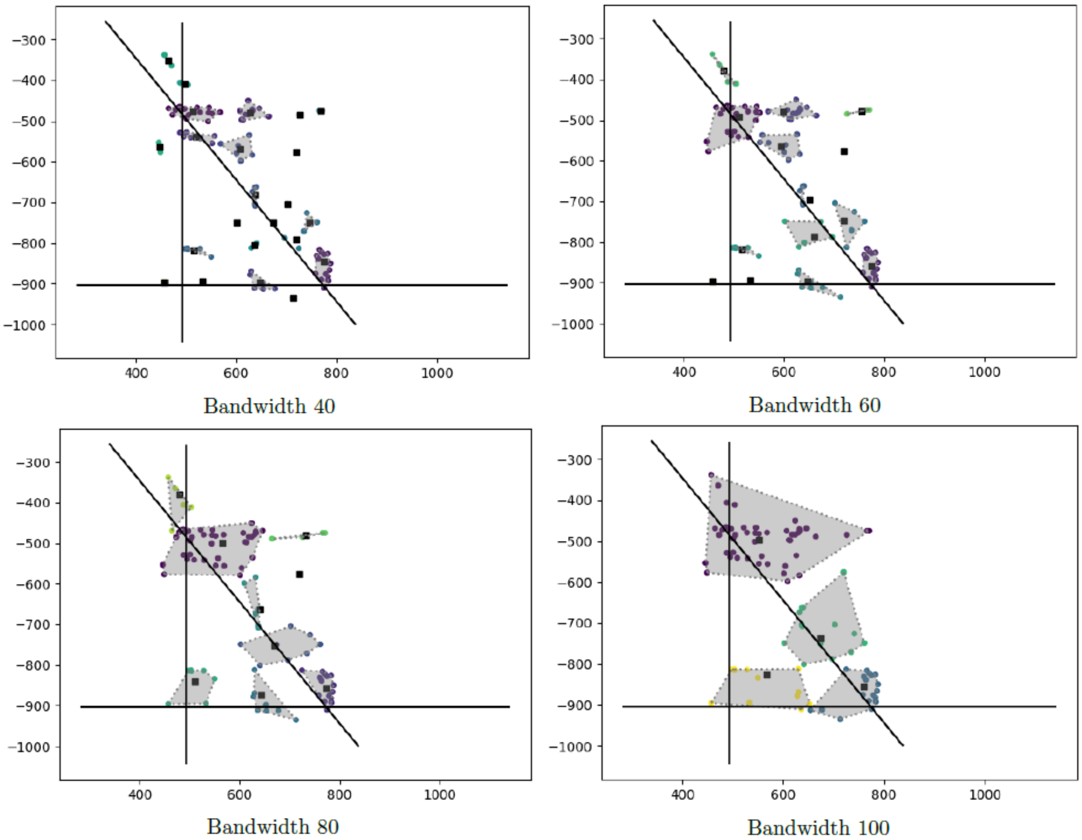


Figure 4.6.: Example clustering for four different bandwidths.

In Figure 4.6, an example clustering for four different bandwidths is depicted. The three black lines mark the x-axis, the y-axis and the linear function. Furthermore, each fixation is represented by a colored circular point and each grey polygon represents a cluster or rather an AOI. The black rectangles are the means of the points in the corresponding cluster. The obtained AOI-allocation can now serve as a basis for applying the procedure for detecting slope triangles.

In a next step, we connect the automated AOI-allocation based on clustering and the procedure for detecting slope triangles. Moreover, we present a new approach for excluding AOI sets based on geometrical criteria and test the combined method on ET data.

To connect the automated AOI-allocation with the detection of slope triangles, some adjustments are required. The pseudocode of the combined method is displayed in Algorithm 3. The parameter l, which describes the maximum order for which the kth order time-aggregated network is derived, is set to be 3, since the aim is to choose the bandwidth in such a way that the fixations that correspond to a vertex of a slope triangle are each contained in one AOI. Furthermore, the exclusion based on the geometrical arrangement performed by the function TEST

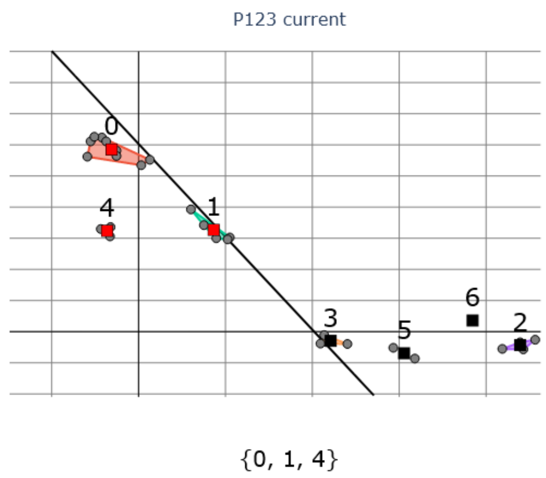


Figure 4.7.: Example of a returned plot of Algorithm 3.

POSITIONING CENTERS is changed due to the different conditions. In addition to the set of AOI-sets, a plot is returned. Figure 4.7 shows an example of such a plot. The AOIs that are contained in an output set are marked with a red center and the corresponding returned AOI-set is $\{0, 1, 4\}$.

As mentioned above, the automated AOI-allocation is strongly affected by the bandwidth. Therefore, a bandwidth that provides a suitable AOI-allocation for detecting slope triangles is needed. The fixations that belong to a vertex of the slope triangle should each be contained in one AOI. If the bandwidth is chosen too large, fixations corresponding to two or all vertices of the slope triangle are contained in one AOI, and thus, the slope triangle can not be detected. Nevertheless, the bandwidth should not be chosen too small to prevent that nearly each fixation forms a cluster. If the slope triangle is rewatched multiple times, every pass is returned as a slope triangle and multiple AOI-sets actually correspond to just one slope triangle. Hence, there is a trade off between too few triangles detected and too many AOI-sets returned due to too small or too large bandwidth.

To test the procedure of detecting slope triangles for several bandwidths, we test Algorithm 3 with a large set of different bandwiths. This makes it possible to deal with the full range problem as transparently as possible.

We now describe the function TEST POSITIONING CENTERS: TEST POSITIONING CENTERS can be generalized such that it can be applied to arbitrary items where students need to determine the slope of a monotonically decreasing linear function. For this purpose, two coordinates located on the linear function must be passed as an input parameter. To evaluate the position of the AOIs to each other and to the linear function, the mean of the fixation coordinates in the corresponding AOI is used. Hereafter, the term "center of the AOI" refers to the mean of the fixation coordinates in the AOI and the terms "all AOIs" or

Algorithm 3: DETECTING SLOPE TRIANGLES BASED ON AN AUTOMATED AOI-ALLOCATION

Input: Tsv file structured as in Table 4.4 and a list of bandwidths $\mathcal{H}$
Output: set $\mathcal{S}$ of sets of natural numbers that correspond to AOIs and a plot as in Figure 4.7, plus a numbering of the AOIs
Data: Tsv file structured as in Table 4.4

1 $\mathcal{S} = \emptyset$
2 Import coordinates of the fixations
3 $l = 0$
4 $\bar{h}$ =None
5 **for** $h \in \mathcal{H}$ **do**
6   Apply mean shift procedure with bandwidth h
7   BUILD ET-NETWORK $G^T = (V, E^T)$
8   DERIVE THIRD ORDER TIME-AGGREGATED NETWORK $G^{(3)} = (V^{(3)}, E^{(3)})$
9   **for** $u = (u_1, u_2, u_3) \in V^{(3)}$ **do**
10   counter(u)= 0
11   **for** $v = (v_1, v_2, v_3) \in V^{(3)}$ **do**
12   **if** $\{u_1, u_2, u_3\} = \{v_1, v_2, v_3\}$ **then**
13   counter(u)=counter(u)+1
14   **for** $u \in V^{(3)}$ **do**
15   **if** *counter(u)+*$\delta^-(u) > 2$ *and u has at least three different entries* **then**
            `/* Exclusion due to geometrical arrangement */`
16
17   **if** *TEST POSITIONING CENTERS(u)=TRUE* **then**
18   Add $\{u_1, u_2, u_3\}$ to $\mathcal{S}$
19   **if** $l \leq |\mathcal{S}|$ **then**
20         $l = |\mathcal{S}|$
21         $\bar{h} = h$
22 Repeat line **6** to **18** for bandwidth $\bar{h}$
23 **return** $\mathcal{S}$ and plot clustering

"AOI-sets" refer to the set of AOIs corresponding to the node that is passed to TEST POSITIONING CENTERS.

Let $u = (u_1, u_2, u_3) \in V^{(3)}$ be a node of a third order time-aggregated network that is passed to TEST POSITIONING CENTERS. By (x^{u_i}, y^{u_i}) the center of the AOI u_i is denoted. The aim of the function TEST POSITIONING CENTERS is to check whether the slope of the linear functions that each interpolate a pair of centers does deviate much from the slope of the catheti and the hypotenuse in a slope triangle. Therefore, each pair of points $(x^{u_i}, y^{u_i}), (x^{u_j}, y^{u_j})$ for $i \neq j, 1 \leq i, j \leq 3$ gets linearly interpolated. Hence, the following equations are obtained

$$\begin{aligned} y &= \frac{(y^{u_j} - y^{u_i})}{(x^{u_j} - x^{u_i})} \cdot (x - x^{u_i}) + y^{u_i} \\ \Leftrightarrow (y^{u_i} - y^{u_j}) \cdot x + (x^{u_j} - x^{u_i}) \cdot y &= y^{u_i} \cdot (x^{u_j} - x^{u_i}) + x^{u_i} \cdot (y^{u_i} - y^{u_j}) \end{aligned} \tag{4.3}$$

for $i \neq j, 1 \leq i, j \leq 3$ and $x, y \in \mathbb{R}$.

The node u is excluded if it is unlikely to contain a slope triangle due to the slope of the linear functions defined by Equation in (4.3). For this, the equation is divided by $m = \max\{|y^{u_i} - y^{u_j}|, |x^{u_j} - x^{u_i}|\}$ and the following equation is obtained

$$\frac{(y^{u_i} - y^{u_j})}{m} \cdot x + \frac{(x^{u_j} - x^{u_i})}{m} \cdot y = \frac{y^{u_i} \cdot (x^{u_j} - x^{u_i}) + x^{u_i} \cdot (y^{u_i} - y^{u_j})}{m} \tag{4.4}$$

Consequently, it holds that $|\frac{x^{u_j} - x^{u_i}}{m}| = 1$ or $|\frac{y^{u_i} - y^{u_j}}{m}| = 1$ or $|\frac{x^{u_j} - x^{u_i}}{m}| = |\frac{y^{u_i} - y^{u_j}}{m}| = 1$.

First, it is checked whether there are $i \neq j, 1 \leq i, j \leq 3$ that fulfill

$$|\frac{x^{u_j} - x^{u_i}}{m}| = 1 \text{ and } |\frac{y^{u_i} - y^{u_j}}{m}| \leq \eta \tag{4.5}$$

for some parameter $\eta < 1$. If this condition is not satisfied, none of the linear functions defined by the equations has an absolute slope less than $\eta < 1$. As the slope of the horizontal cathetus in a slope triangle is equal to zero, none of the linear functions corresponding to the Equation in (4.3) represents the horizontal cathetus to the extent specified by the parameter η. Thus, it is unlikely that the corresponding AOIs contain a slope triangle and FALSE is returned. Depending on the item, the parameter η can be adjusted. Furthermore, the equation corresponding to the linear function with the minimal absolute slope is marked.

Example 4.15. The three red points in Figure 4.8 mark the centers of the AOIs corresponding to the node that is passed to TEST POSITIONING CENTERS. Further, the dashed black lines represent the linear functions that interpolate the corresponding pair of points. The continuous black line represents the linear function whose slope the students have to determine. In this example, η is set to be 0.6. Other values for η would also be possible. The closer η is chosen to 1, the more the slope of the linear functions defined by the Equation (4.3) can deviate from the slope of the catheti. The linear function l_1 fulfills Equation (4.5) as the dashed gray lines in Figure 4.9 mark two linear functions with absolute slope 0.6 and y-axis intercept 3. Consequently, a linear function that is located in the gray

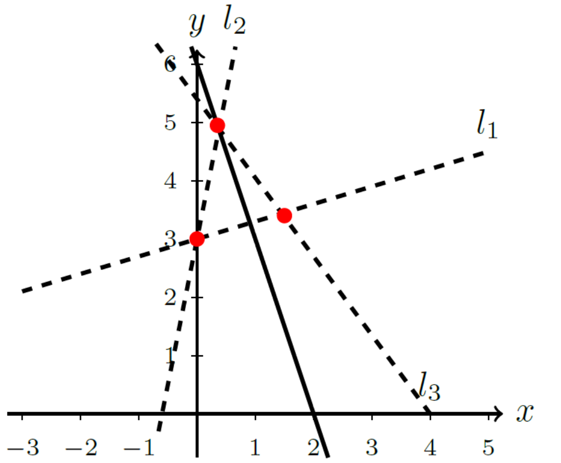


Figure 4.8.: Example of interpolated centers.

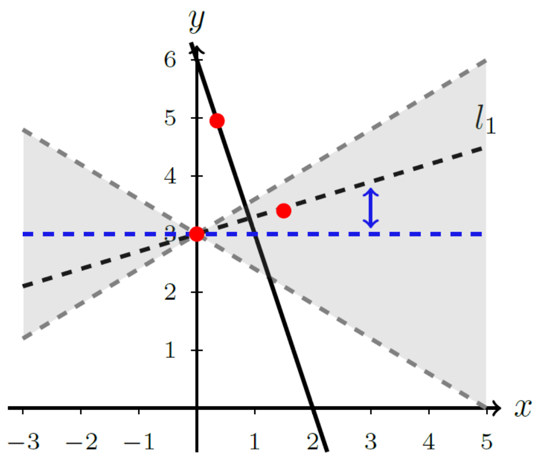


Figure 4.9.: Illustration of condition 4.5.

area has an absolute slope less than 0.6. Furthermore, the function l_1 is marked as it has the flattest slope.

Next, it is checked whether there are $i \neq j, 1 \leq i, j \leq 2$ that fulfill

$$|\frac{y^{u_i} - y^{u_j}}{m}| = 1 \text{ and } |\frac{x^{u_j} - x^{u_i}}{m}| \leq \eta. \tag{4.6}$$

If this condition is not fulfilled, none of the three linear functions corresponding to the Equation (4.3) has an absolute slope greater than $\frac{1}{\eta}$. Thus, none of them represents a vertical cathetus to the extent specified by the parameter $\frac{1}{\eta}$ and FALSE is returned. Analogously to above, the linear function corresponding to the equation with the maximal absolute slope is marked.

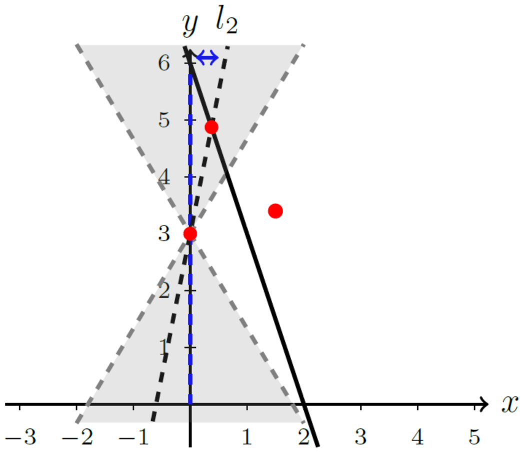


Figure 4.10.: Illustration of Condition 4.6.

Example 4.16. Analogously to Figure 4.9, a linear function located in the grey area in Figure 4.10 has a y-axis intercept of 3 and a slope greater than $\frac{1}{\eta} = \frac{10}{6}$. Thus, the slope of in Figure 4.10 is greater than $\frac{1}{\eta} = \frac{10}{6}$. Moreover, the function l_2 is marked, as it has the steepest slope.

If both of the Conditions (4.5) and (4.6) are fulfilled, the equation that is not yet marked is examined. Since the linear function corresponding to the unmarked equation neither has the steepest nor the flattest slope, it is assumed to be the supposed hypotenuse. As the linear function, for which the students are to determine the slope, is monotonically decreasing, it is checked whether the slope of the linear function that corresponds to the unmarked equation is negative. If this is not the case, FALSE is returned. Additionally, the following convex combination $(\bar{x}, \bar{y}) = \frac{1}{2} \cdot (x^{u_i}, y^{u_i}) + \frac{1}{2} \cdot (x^{u_j}, y^{u_j})$ of the points $(x^{u_i}, y^{u_i}), (x^{u_j}, y^{u_j})$ that correspond to the unmarked equation is calculated. If the distance $|f(\bar{x}) - \bar{y}|$, where f is the linear function of which the students need to determine the slope, is greater than a constant d_h, FALSE is returned. This condition is to ensure that the supposed hypotenuse is located near the function f.

The following condition is added to exclude cases in which the x-axis intercept of f and the y-axis intercept of f are fixated consecutively. An example gaze path for this is displayed in Figure 4.11.

Therefore, it is checked whether the term $\sum_{i=1}^{3} \sum_{j=i+1}^{3} ||(x^{u_j}, y^{u_j}) - (x^{u_i}, y^{u_i})||$, where $||\cdot||$ denotes the euclidean norm, is less than a constant d_{total}. If the term is greater than or equal d_{total}, FALSE is returned. Although, gaze paths as in Figure 4.11 need to be excluded, a disadvantage of this condition is that large slope triangles get excluded. Thus, an alternative way of excluding gaze paths as in Figure 4.11 is desirable. One possibility might be to observe the ratio of the length of the catheti. In slope triangles, the ratio of the length of the catheti is constant and equal to the absolute slope of the linear function. In the case when

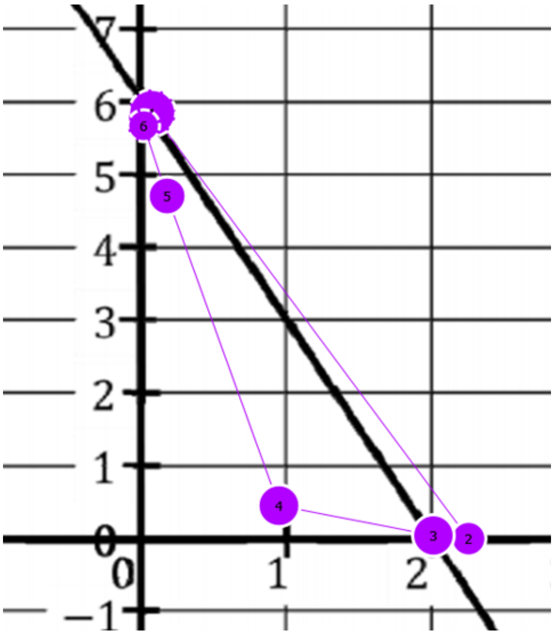


Figure 4.11.: Gaze path which goes from the x-axis intercept to the y-axis intercept.

the axis intercepts are fixated consecutively, one cathetus is very short compared to the other cathetus and the ratio differs from the ratio in slope triangles.

Results

We use a network-based approach that enables structure recognition on small data sets. To this extent, we analyze gaze data of 62 students who determined the slope of a linearly decreasing function. All students explained their solution strategy in an interview after completing the task using the gaze-cued retrospective think-aloud method. The algorithmic pattern recognition is validated using the performance metrics, recall and precision. Recall denotes the number of students who, according to the interview, used a slope triangle as a solution strategy and for whom the algorithm recognizes at least one slope triangle, divided by the number of all persons who claimed to use the slope triangle solution strategy in the interview.

In contrast to recall, for precision, the denominator changes to the number of all persons for whom the algorithm detected at least one slope triangle.

We appply Algorithm 3 to gaze data from 62 students solving item M8 from the test instrument by Ceuppens et al. (2019) (cf. Figure 4.5 on p. 105 for original version in english and Section A.1 in the Appendix for the version used here in German). In Figure 4.12 we show the number of cases in which the algorithm detected slope triangles and the performance metrics (precision and recall), for

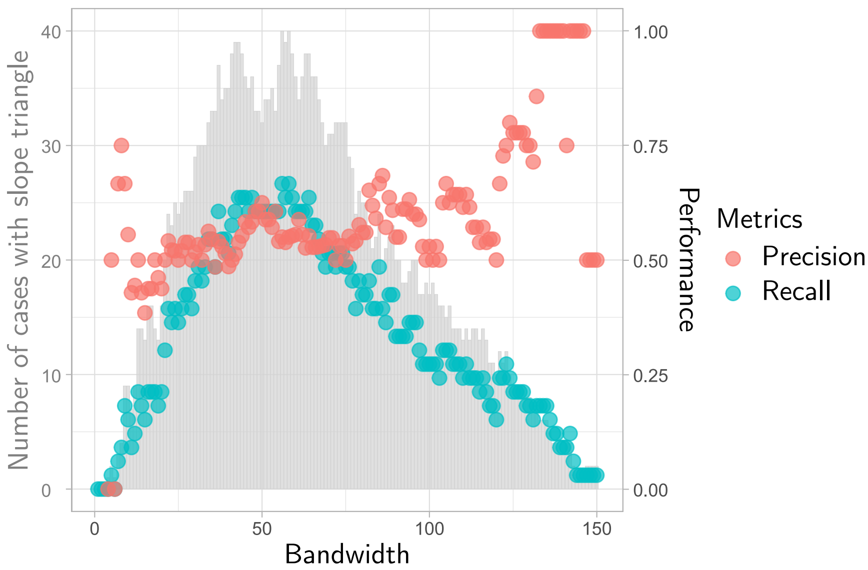


Figure 4.12.: Gray: number of cases with at least one slope triangle detected and performance metrics of the algorithm dependent on bandwidths. Red: precision (positive predictive value). Blue: recall (true positive rate). All results calculated for bandwidths from 1 to 150 in steps of 1.

the modulated bandwidth parameter of the mean shift clustering algorithm. As previously stated, very large and small bandwidths lead to low numbers of detected slope triangles (cf. Figure 4.12, gray). Since the number of people who used a slope triangle to solve the task is constant, the slope of recall values is similar to the one of the number of cases with slope triangles (cf. Figure 4.12, blue). This is due to the fact that a small number of detected slope triangles increase the precision. In summary, the algorithm for fixed bandwidths yields maximum recall values of 0.67 with bandwidths of 66 and 68 and an average precision of 0.60.

4.2. Suitability-Test of Network Metrics to Differentiate Between Correct and Incorrect Solvers

In the context of this section, we investigate whether certain network metrics are suitable to discriminate between correct and incorrect solvers based on gaze data alone. To this end, we describe an alternative, innovative analysis that uses network analysis methods. The gaze data are first grouped using mean shift clustering (cf. Section 12). The clustered gaze points are then transformed into a network on which various network metrics are examined. The results are divided into two groups (correct and incorrect solvers) and analyzed for statistically significant differences between the groups. Again, the goal is to use basic research to create initial approaches for integrating this technology into ALS to improve their diagnostic capabilities. The Python code for all the analyses in this chapter can be found in Knippertz (2024d). Note that this is an exploratory data analysis that will be treated statistically separately from the previous ones. This means, in particular, that the number of statistical tests performed here is not included in the corrections of the statistical analyses performed previously.

Methodical Considerations

We examine eight different tasks (= four IPs, namely IP 5, 6, 7, and 8) from the test instrument of Ceuppens et al. (2019), cf. Section 2.2. Only this selection of the entire test instrument is considered, as the item difficulty of these items is mostly in the medium range (see Section 3.1). To distinguish between correct and incorrect solvers, the items should have been solved correctly or incorrectly by approximately the same number of people. Furthermore, we aim to limit ourselves to one single type of representation—in this case—*graphs*—and to one type of task—in this case—*determine on graphs*—to avoid representation-specific and task-specific differences.

The sample consists of 131 students in the entry phase of upper secondary school, taking part in an ET-study, cf. Section 2.2.

Construction of the Network Out of ET Data

The construction of the network out of ET data was carried out in three steps:

1. First, a tsv file was created from the ET data in Tobii Studio for each participant (see Table 4.5). The tsv file contains information about FixationIndex (index of the fixation), GazeEventDuration (duration of the fixation in ms), FixationPointX and FixationPointY (coordinates of the fixation point), and AOI[task]Hit (information about on which area of the examined tasks of the test instrument the fixation belongs to). AOI[task]Hit can take the value 1 (fixation is in the graph area of the task), the value 0 (fixation is outside the graph area of the task, e.g., in the text area), or no numerical value (fixation belongs to another task). Only fixations in the graph area were considered. The data was read into Python using the *datareader.get_ all_ data function*. The function returns a Python dictionary that stores the gaze data in the graph region and the task solution (1:correct, 0:incorrect) for each person.

2. In a second step, a person's fixations for each task were clustered using the mean shift clustering algorithm (function *graphs.mean_ shift*). Again, the implementation from the scikit-Learn library was used (cf. Pedregosa

et al. (2011)). The clustering and subsequent analysis were performed for bandwidths h ranging from 3 to 50 with a step size of 1.

3. In a final step, based on the clustering, a graph was generated for each task and each student using the *graphs.graph_from_clustering* function. The nodes of the graph were formed by the clusters, the edges by the gaze path. For each saccade from one cluster to another, a directed edge was created between the two clusters. Since a cluster usually consists of several fixations depending on the bandwidth, there can be several saccades from one cluster to another. If a saccade in the gaze path created a directed edge to another cluster to which an edge already existed, the edge weight was increased instead. If the number of nodes in a graph after this procedure did not exceed 2, the graph was discarded, which could happen in some cases with large bandwidths. Such an extremely low number of clusters is an indication that not all fixations were captured by the eye tracker, which implies a low quality of the data. The cluster graph was visualized using the function *visualize.draw_clustered_graph*, which draws squares at the coordinates of the cluster centers. The area of the square is proportional to the sum of the gaze durations of the individual fixations that are part of the corresponding cluster. An example visualization for different bandwidths can be seen in Figure 4.13. For better visualization, the direction and weight of the edges are not shown. Note that the positions of the nodes are not a property of the graph, but serve only for visualization in the context of the ET task.

Calculation of Network Metrics and Analysis

For each task, the students were divided into two groups based on their results (correct/false), and a list of graphs was created for the correct and incorrect answers separately. For each graph, various metrics implemented in the NetworkX library (see Hagberg et al. (2008)) were calculated using the *metrics.significance_analysis_all_metrics* function. The 14 metrics analyzed are (cf. Section 2.1):

- Average Degree Connectivity
- Average Shortest Path Length
- Betweenness Centrality
- Closeness Centrality
- Clustering Coefficient
- Communicability
- Density
- Eigenvector Centrality
- Harmonic Centrality
- Katz Centrality
- Reciprocity
- Squares Clustering Coefficient
- s-Metric
- Transitivity

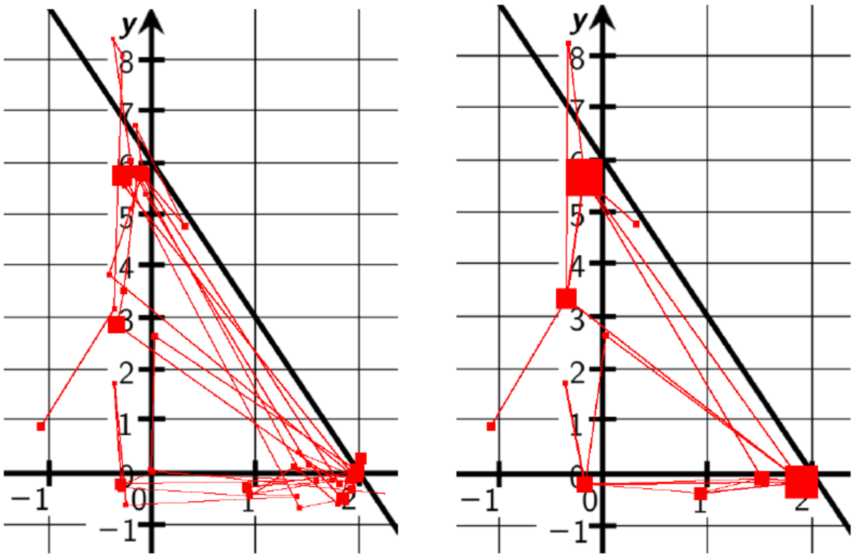


Figure 4.13.: Graph area of math item M8 from the validated test instrument by Ceuppens et al. (2019) where students had to determine the slope of the linear function. Cluster centers of the clustered fixations (red squares) and saccades between the clusters (red lines) can be seen. On the left, the cluster graph was formed with a bandwidth of 1, on the right with a bandwidth of 50.

For those metrics that are not defined for an entire graph, but for individual nodes, the mean value was calculated over the metric value of all nodes of the graph. This is a common procedure that is also used for other analyzed metrics that are defined for both individual nodes and the entire graph, such as the clustering coefficient or the density.

For each of the above-mentioned metrics, the *metrics.significance_analysis_all_metrics* function was also used to analyze whether the deviation of the values between the correct and incorrect answers was significant. As the results were not normally distributed for each group for a large number of the metrics, a t-test could not be performed. Instead, a one-sided Mann-Whitney U test (cf. Section 2.2) was used. The null hypothesis here was that it is equally likely that a value randomly selected from the distribution with the average higher values is greater than a value randomly selected from the distribution with the average lower values. Rejecting the null hypothesis therefore meant a significant deviation between the correct and incorrect answers. For the evaluation, a multidimensional array was created that contained the information on the tested bandwidth, task and metric, the group with the higher average value of the metric and the p-value (before error correction for multiple testing) for each metric.

Table 4.5.: Sample ET data. Each line represents a fixation.

Fixation Index	Gaze Event Duration	Fixation PointX	Fixation PointY	AOI[M5] Hit	AOI[K5] Hit	AOI[...] Hit
1	425	777	443	0		...
2	225	432	118	0		...
3	133	308	144	1		...
4	326	625	111	0	...	...
...	...	...	...	...	...	...
27	436	780	326		0	...
28	122	296	165		1	...
...	...	...	...	...	...	...

To correct the false discovery rate, the Benjamini-Hochberg error correction was applied to adjust the p-values (see Section 2.2) with multiple alpha $\alpha = 5\%$ across all 5376 tests combined. The number of tests results from the number of possible combinations of the 48 bandwidths, 8 tasks and 14 metrics examined.

The number of significant differences in the values of the metrics between the two groups of correct and incorrect solvers was examined, broken down by bandwidth, task and metric. This data was calculated using the functions of the *evaluation.py* module and output in both tabular and graphical form.

Results

Task-specific significant differences: Figure 4.14 shows the number of significant differences in the values of the metrics between the two groups of correct and incorrect solvers per task. For each task, 48 bandwidths with 14 metrics each were examined, so a maximum of 672 significant differences per task is possible.

The value is highest for the two items M8 (419 significant differences) and M7 (404 significant differences). Item M5 also shows a high number of significant differences (297). Among the items in the mathematical context, only item M6 produced a low value (24). Within the kinematics items, only item K6 shows a medium-high number of significant differences (154). The values of all other items in the kinematic context (K5, K7 and K8) are only in the low range, whereby K7 stands out with 48 significant differences compared to K5 and K6, which only have values in the single-digit range.

When the tasks are subdivided according to the concept to be recorded, there are only slight deviations in the sum of the significant differences of a group (y-axis intercept: 752 / slope: 603). When subdivided according to the type of slope, stronger deviations emerge (positive: 478 / negative: 877).

Bandwidth-specific significant differences: The number of significant differences in the values of the metrics between the two groups of correct and in-

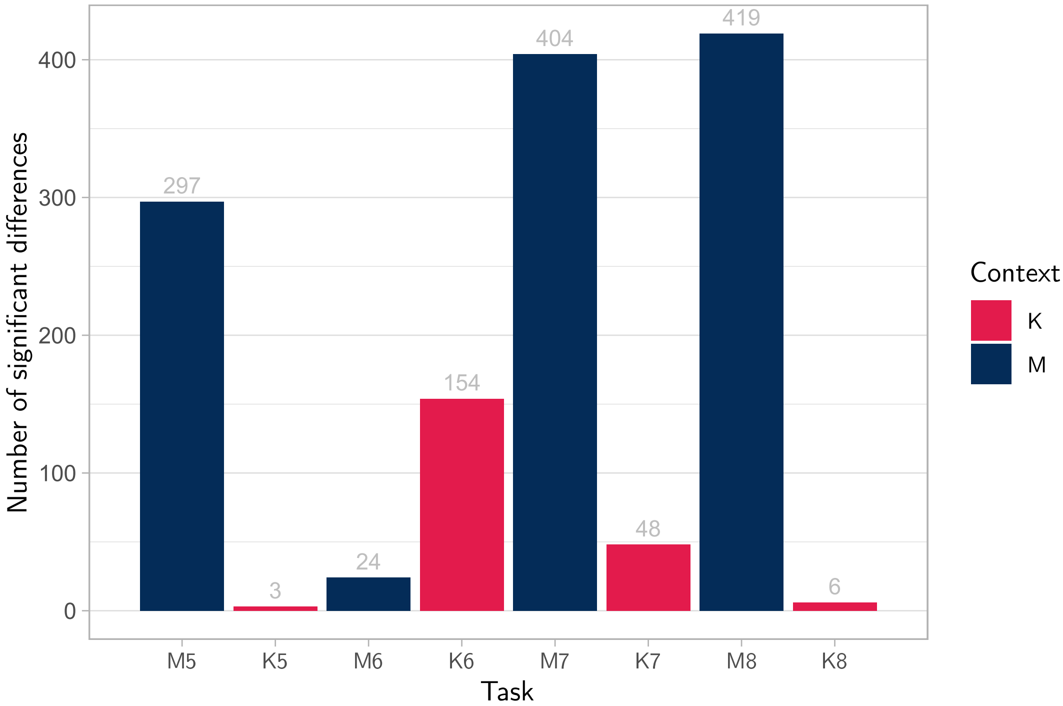


Figure 4.14.: Number of significant differences per task.

correct senders was analyzed according to the bandwidth h of the mean-shift clustering algorithm. The results are shown in Figure 4.15. For each bandwidth, a maximum of 112 significant differences are possible for the 8 tasks examined, each with 14 metrics.

For the bandwidths $h = 5, 7$ and 18, the value is the highest (33 significant differences). For $h = 47$ and 49, the number of significant differences is lowest with a value of 22. Overall, there is a slight decrease in the values toward the end of the analyzed spectrum, i.e., for a higher bandwidth.

Metric-specific significant differences: The frequency of significant differences depending on the metric examined is shown in Figure 4.16. In each case, it is shown which group (correct / incorrect) has a significant higher mean metric value. Each metric is examined in 8 tasks, each with 48 bandwidths, whereby a maximum of 384 significant differences are possible.

For the metrics Clustering Coefficient, Reciprocity and Squares Clustering Coefficient, there are only a small number of significant differences in the metric mean values (Clustering Coefficient: 23, Reciprocity: 9, Squares Clustering Coefficient: 34). However, if a significant difference occurs here, then the higher metric mean value is always over 90% in the group of incorrect respondents. The results show strong tendencies for six metrics (at least 75% of a total of over 50 significantly higher metric mean values for one group):

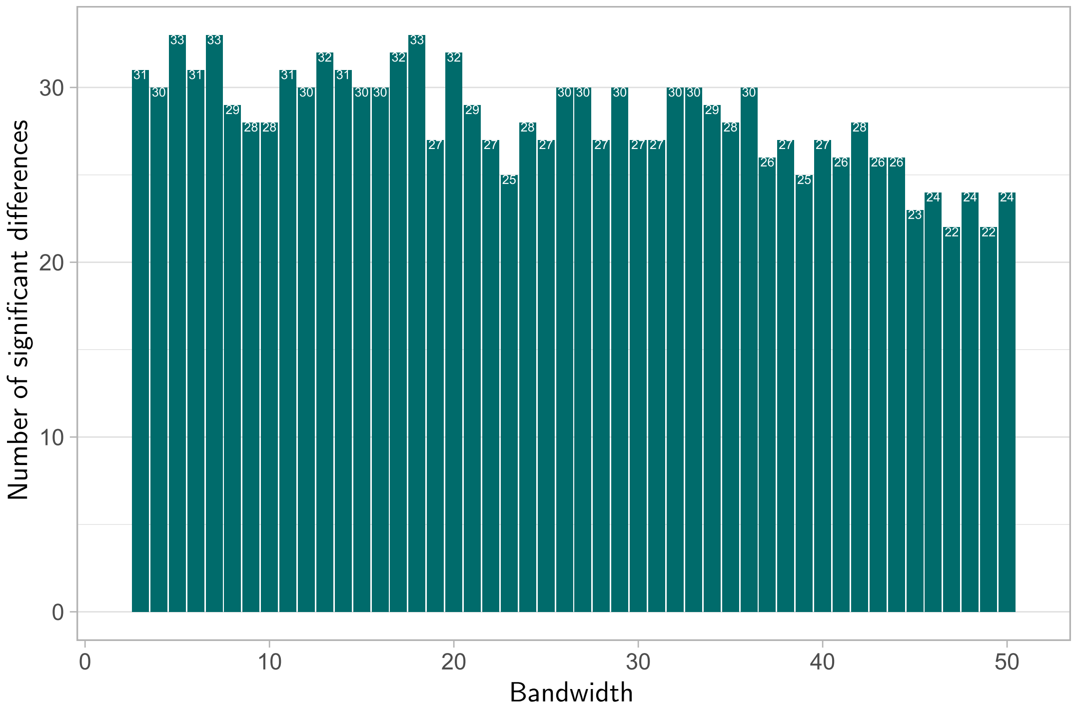


Figure 4.15.: Number of significant differences per bandwidth.

The metric mean for the group of correct respondents is higher in most cases for the following metrics: Closeness Centrality (81% of all 52 significantly higher metric mean values are among the correct solvers), Density (75% of 166) and Eigenvector Centrality (79% of 146).

The following metrics show a similar picture for the group of false solvers: Communicability (75% of all 71 significantly higher metric mean values are among the false positives), Harmonic Centrality (85% of 177) and Average Shortest Path Length (78% of 158). A very clear picture emerges for five other metrics (at least 90% of a total of over 50 significantly higher metric mean values lie with one group):

The mean metric value for the group of correct solvers is almost always higher for the following metrics than that of the incorrect respondents: Betweenness Centrality (100% of all 85 significantly higher metric mean values are among the correct solvers) and Katz Centrality (93% of 155).

For the group of false solvers, the following metrics show a similar pattern Picture: Average Degree Connectivity (100% of all 70 significantly higher metric mean values are among the false positives), s-Metric (100% of 150) and Transitivity (97% of 59).

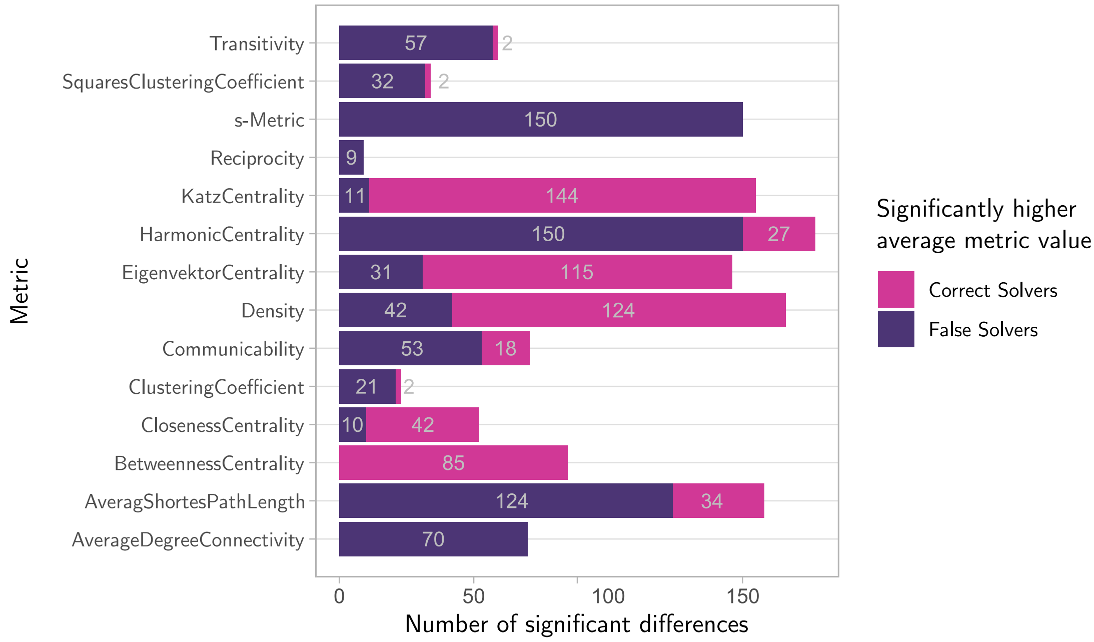


Figure 4.16.: Number of significant differences per metric.

4.3. Discussion and Summary

Detecting Geometrical Structures

RQ8: To sum up, we developed a method for modeling ET data by means of a temporal network G^T. Therefore, sequential information of the data is obtained and thus the 8th research question "How to model ET data as a network while maintaining sequential information of the data?" is answered. Furthermore, we presented an algorithm that transforms ET data into an ET-network in $\mathcal{O}(h \cdot n)$, where h is the length of the input list of the algorithm. Concerning kth order time-aggregated networks, we derived the upper bound $\min\{(n \cdot (n-1)^{k-}, T-k+2\}$ on the number of nodes and the upper bound $\min\{(n \cdot (n-1)^{k}, T-k+1\}$ on the number of edges in the kth order time-aggregated network. Moreover, we introduced an algorithm that derives a kth order time-aggregated network in $\mathcal{O}((T-k) \cdot k)$.

Thereupon, we show that common ET measures can be calculated by means of ET-networks, and therefore, an ET-network is an alternative for saving ET data, with the benefit of a clear and compact structure.

RQ9: We developed a method for detecting slope triangles in the subject's gaze and tested the corresponding algorithm. The procedure combines the investigation of kth order time-aggregated networks and the usage of geometrical criteria. We presented an automated alternative for placing AOIs by means of mean shift clustering and combined the automated AOI-allocation with the procedure for detecting slope triangles. The results show that it is possible to detect geometrical structures by means of ET data. The algorithm for fixed bandwidths yields maximum recall values of 0.67 with bandwidths of 66 and 68 and an average precision of 0.60. However, graph theoretical or geometrical criteria could be added in order to try to improve the results. For this, further investigation of geometrical structures from an educational science perspective could be helpful. In addition, the bandwidth yields potential for further improvement. A method for determining a suitable bandwidth for each subject individually could be developed in further research. Starting points for this could be, for example, the number of coordinates, the smallest distance between coordinates or the average distance of coordinates. Another possibility would be to first run the mean shift procedure with a larger bandwidth and then examine the clusters obtained using the previously mentioned criteria. On each cluster, the mean shift clustering could then be executed again with individual bandwidths.

Moreover, the low computational effort of the mean shift procedure could be used. The clustering and the detection procedure could be performed after certain time steps in the recording, which simulates a “real-time” application. Based on that, it could be investigated whether slope triangles or other geometrical structures are built at a specific point in time in the recording and whether it is possible to use this as a basis for an ALS.

Differenciating Network Metrics

RQ10: With regard to research question 10, “Are the methods of network analysis suitable for differentiating gaze data from correctly and incorrectly solving persons with regard to item difficulty, context and clustering?”, it can be seen that many network metrics are quite suitable to differentiate between the groups of correct and incorrect solvers in several of the tasks studied. This statement is supported by the high number of significant differences between the mean metric values of the two groups of people.

Due to the fact that the sizes of the groups of correct and incorrect respondents differ significantly when the item difficulty is very high or very low, it can be assumed that the Mann-Whitney U test (see Subsection 2.2) can be used to detect fewer significant differences between the two groups. This is due to the fact that the power of the test is usually lower (depending on the variances of the values of the samples) in the case of samples of very different sizes than in the case of equally distributed group sizes (Happ et al. (2019) and Zimmerman (1987)). The small size of a group leads to the fact that random deviations are more difficult to exclude and therefore, a stronger deviation must be present in order to be detected as significant with a low probability of error. With regard to the evaluation of the item difficulties (cf. Subsection **??**) and the number of significant differences between the metric mean values, sorted by task (see Figure 4.14 on p. 120), it can be seen that the items with a high number of significant differences (M5, M7, M8) each have a medium item difficulty. On the other hand, it is the case that not all items with a medium item difficulty are associated with a high number of significant differences. It is also noticeable that item K6 with the fourth highest value of significant differences between the metrics of the two groups of people even has the second lowest item difficulty across all tasks. It can therefore be said that a medium item difficulty is not a sufficient condition for a high number of significant differences. However, tasks with a high number of significant differences are often associated with a medium item difficulty.

As shown in Figure 4.14 on p. 120, for four tasks in the context of mathematics, the number of significant differences between the groups of people who solved correctly and incorrectly is significantly higher in the mathemtics than in the kinematics context. Here, too, there are outliers; not all mathematical items have a high value (see M6), and the value is not very low for all kinematic items (see K6). Overall, however, it can be assumed with regard to research question 11 that data for tasks in the context of mathematics are much better suited to be analyzed using network metrics. A possible explanation for this is, that solution strategies are used in a context-dependent manner. According to this, test subjects use a different solution strategy for the kinematics tasks than in the mathematics tasks.

Research question 10 can be answered to the effect that the bandwidth h of the mean shift clustering algorithm, which performs the preprocessing of the data, plays only a minor role in the number of significant differences between the metric values of the two groups of people. The highest number occurs at a bandwidth in the lower third of the analyzed spectrum, i.e. between bandwidths 3 and 18. Values in the middle third are slightly lower, followed by a further slight decrease in the number. One possible explanation for this is that the mean-shift- clustering for the high bandwidths results in graphs with only very few clusters or nodes and therefore so much information is lost that an analysis is more difficult. Overall, it can be seen that the relevance of clustering and the AOIs are of less importance for the analysis by network metrics than originally assumed, since relevant results also occur for very low values of the bandwidth h.

RQ11: With regard to research question 11: "If there are differentiating network metrics, which network metrics are particullarly suitable for differentiating between correctly and incorrectly solving persons?", we will discuss the suitability of the different network metrics in detail:

- Average Degree Connectivity: The value is on average higher for the group of false solvers. Possible explanations for this are that ...
  - they make more saccades and thus create more edges between the nodes, resulting in a higher average degree. This is in line with Höffler, Leutner (2007) who state that more saccades occur in the gaze data of people with less prior knowledge.

  - the degrees of the nodes are more evenly distributed between them, and the vacancy in the degrees is lower than in the group of correct solvers. This was consistent with the research results of Zhu, Feng (2015), that correct solvers make more strategic transition and saccades of incorrect solvers occur randomly.

- Average Shortest Path Length: The value is on average higher for the group of false solvers. This means that the shortest paths in their gaze data are shorter on average. One possible explanation for this is that there are generally more saccades in the gaze data of the false solvers and the resulting more frequent direct connections between nodes mean that the shortest paths between these nodes are shorter on average. This would also be consistent with the assumption that the gaze data of people with less prior knowledge shows more saccades Höffler, Leutner (2007).

- Betweenness Centrality: The value is on average higher for the group of correct respondents. One possible explanation for this is that there are more central nodes in the clustered gaze data of the correct solvers, through which many of the shortest paths run.

- Closeness Centrality: The value is on average higher for the group of correct solvers. Since the Closeness Centrality for connected graphs is the reciprocal of the Average Shortest Path Length, this result shows an internal consistency between the results of the two metrics.

- Clustering coefficient: The value is higher on average for the group of incorrect respondents. This means that in the gaze data of the correct respondents, the average proportion of existing triangles in relation to all possible triangles is higher. One possible explanation for this is that in the studied group of people, nodes in rather irrelevant areas of the task are more strongly connected and more triangles are formed at these points than in the group of correct solvers. However, it should be noted that a significantly higher mean metric value occurs only in a few cases; therefore, the assertion just made should be viewed with caution.

- Communicability: The value is higher on average for the group of false respondents. Possible explanations for this are that ...
  - the shortest paths between two nodes in the gaze data graphs of the group of incorrectly solving persons are generally shorter (see Average Shortest Path Length, which would confirm the internal consistency between the two metrics).
  - There are on average more paths between two nodes, for example because the graph generally consists of more nodes. This could be explained by the large number of studies, which show that false solvers

generally look at more irrelevant areas of the stimulus, which is why more clusters are created overall (Bayazit et al. (2014), Brückner et al. (2020), Kekule (2014), Klein et al. (2019), and Tsai et al. (2012) and others)

- Density: The value is higher on average for the group of correct respondents. This means that the ratio of actually existing edges compared to potentially possible edges is higher for the group of correct solvers. This can be explained by the above-mentioned fact that correct solvers focus their attention more on the relevant points of the task and tend to ignore irrelevant points in the task, which leads to fewer clusters overall.

- Eigenvector Centrality: The value is on average higher for the group of the correct solvers. This means that nodes in the gaze data of this group are on average adjacent to more central nodes (according to the eigenvector centrality metric). One possible explanation for this is that correct solvers look more frequently into the areas relevant for processing the task and make strategic transitions between these areas, which is why the nodes in these relevant areas have a high centrality.

- Harmonic Centrality: The value is higher on average for the group of false solvers. Possible explanations for this are that ...
  - there are on average more nodes in the gaze data graphs of the false solvers. This would also support the large number of study results mentioned in Section 3.3.4, i.e., that wrong-setters generally look more often at irrelevant areas of the stimulus, which is why more clusters are created overall (Bayazit et al. (2014), Brückner et al. (2020), Kekule (2014), Klein et al. (2019), and Tsai et al. (2012) and others).
  - the shortest paths between two nodes in the gaze data graphs of the group of incorrectly reading persons are generally shorter (see Average Shortest Path Length, which would confirm the internal consistency between the two metrics).

- Katz Centrality: The value is on average higher for the group of correct respondents. Since Katz Centrality is a generalization of Eigenvector Centrality, this result shows an internal consistency between the results of the two metrics.

- Reciprocity: The value is on average higher for the group of false solvers. This means that more edges that exist between the nodes in the gaze data graphs of the false solvers, point in both directions than in the graphs of the true solvers. This supports the assumption that more saccades occur in the gaze data of people with less prior knowledge and that these occur more randomly. However, it should also be noted here that a significantly higher mean metric value only occurs in a few cases and the assertion just made should therefore be viewed with caution.

- Squares Clustering Coefficient: The value is on average higher for the group of false solvers. Since the squares clustering coefficient queries a similar systematic property of the graphs as the clustering coefficient, this result shows an internal consistency between the results of the two metrics.
- s-Metric: The value is higher on average for the group of false solvers. One possible explanation for this is that ...
  - they look back and forth between different clusters more often and therefore create more edges between them, which increases the average degree.
  - they direct their gaze to more different areas of the stimulus than correct solvers, which leads to more clusters overall.
- Transitivity: The value is higher on average for the group of false solvers. Since this metric also asks for a similar systematic property of the graphs as the clustering coefficient, this result shows an internal consistency between the results of the two metrics. The better suitability of transitivity to capture differences between the graphs of the groups may be due to the fact that it directly takes an overall view of the proportion of triangles present in the entire graph, rather than forming an average over the value of individual nodes.

This explorative study has some limitations: We only considered line graphs. As the age of the subjects is exclusively between 15 and 19 years, it would be interesting to see in a follow-up study whether the observed results also occur in another age group or generally in all age groups. It would also be interesting to investigate whether and how the results differ for subjects with more or less prior knowledge or expertise, also in order to draw more conclusions about the relevance of item difficulty.

The existence of significant differences between a large proportion of the mean values of the metrics investigated indicates that there are structural differences between the gaze data graphs of the correct and incorrect learners that can be measured with the same metrics. It can therefore be the subject of a follow-up study to create a predictor based on the metrics highlighted in Subsection 4.2 which discriminate very well between the correct and incorrect solving groups of people, and which predicts only on the basis of a person's gaze data. For this purpose, the values of the aforementioned metrics can be used to classify the person into the group of correct or incorrect learners. In addition, it would be interesting to investigate whether the significant differences occur while the task is being solved by analyzing the incomplete gaze data graph. In this case, the results could be used, for example, to provide early support in learning contexts to a person who has been classified by the predictor as a likely incorrect learner during the task. Such integration into ALS could thus make a positive contribution to the person's learning process.

5. Conclusion and Future Research

Finally, we can answer the title question about context dependency of linear functions with yes. We showed that linear function problems are solved very differently depending on the context. Students use completely different solution strategies to solve similar linear function problems in mathematics and physics. This is despite the fact that these pairs of tasks require the same solution strategy.

To this end, a study of similar linear function tasks in mathematics and physics was conducted by Ceuppens et al. (2019). The tasks from their test instrument require the same solution strategy regardless of the context. We replicated the study of Ceuppens et al. (2019) for $N = 249$ 9th graders. In addition, we investigated whether the difficulties identified by Ceuppens et al. (2019) with physical tasks, especially with negative speed, persist in the entry phase of high school ($N = 298$). Furthermore, we examined the gaze behavior of 131 students when solving graph and formula tasks from Ceuppens' test instrument. In a next step, basic research for possible implementations in ALS was investigated. On the one hand, it was investigated whether it is algorithmically possible to recognize slope triangles from gaze data. On the other hand, network metrics were exploratively investigated for their ability to discriminate between correct and incorrect solvers.

The results of the replication study show that German students have an average item difficulty that is about 30% lower than that of Belgian students of the same age. This means that German students have more problems solving mathematics and physics problems correctly than Belgian students. The results of the Ceuppens study are therefore not directly transferable between countries. Differences can also be seen between the different grades. Here, the performance of German students improved by 88%, while the increase in performance in physics tasks was only 56%. This result shows that gaps in understanding in physics are often not closed during the school years.

Students' gaze behavior is completely different for similar tasks in math and physics. The triangulation of ET and interview data has shown that students solve problems from different contexts that require the same solution strategies differently depending on the context. In our case, the math problems were often solved correctly and the physics problems were mostly solved incorrectly. Thus, there often is a lack of transfer of mathematical content to other contexts. A major problem is the appropriate interpretation of the physics content. A closer integration of mathematical and physical content could help here.

To overcome these challenges for students, especially in the transfer from mathematics to other disciplines, we suggest emphasizing the correspondences and relationships between graphs and formulas in mathematics and in other disciplines. Concretely, this can be done through contrasting cases. ET data from experts demonstrating efficient processing of formulas or graphs can also help here. It would be exciting to empirically investigate the extent to which these suggestions are able to improve students' transfer skills.

In this thesis we used the test instrument by Ceuppens et al. (2019), which is not optimized for ET research because the item pairs consist of different surface properties. Unfortunately, there is no comparable validated test instrument specifically for ET research. In the future, attention should be paid to a comparable surface structure when creating test instruments so that they can be used appropriately in ET research. Existing test instruments should be revised and revalidated accordingly.

Furthermore, it is worth investigating whether the failure to transfer mathematical solution procedures is also a major source of error in other contexts and forms of representation. This could include not only physics but also the other natural sciences, since the degree of mathematization in biology or chemistry is comparable to that in physics. Moreover, it would be interesting to investigate empirically to what extent teacher training has an influence on the use of mathematical solution procedures; for example, do teachers who teach both mathematics and physics succeed better in linking mathematics and physics to solve physical problems?

Our teacher training has shown that math and physics teachers, in particular, already incorporate many elements of interlinking into their teaching. However, according to the teachers, this is not sustainable enough, because after a few weeks, the content is no longer present in the students' minds, or they cannot recall material that has already been covered. Is the problem the 45-minute teaching cycle in which different subjects are taught side by side every day without any connection? Or is it more a problem of the school system, which does not leave enough room for sufficient repetition and consolidation of learning material due to full curricula? Is it due to the adolescence of the students, which often makes it difficult to absorb logical content and puts other things than school at the forefront of young people's minds? All these concerns were expressed in the teacher training course. It is probably a mixture of all these problems. In addition, teachers who teach mathematics and a subject other than physics should be trained so that they too can integrate cross-curricular parallels from mathematics to physics into their teaching. Teacher training has made a promising contribution to this, as the survey results show.

Due to the small number of participants, we were only able to evaluate this training qualitatively. It is planned to test and evaluate the training again with a larger target group. Since the training material is available online in German and English (cf. Knippertz (2024c)), the course can also be used by other people in the future.

In the future, ALS may be able to take over repetition and practice tasks, thus reducing the workload of teachers. In this work, we have shown that mathematical structures are suitable for the algorithmic recognition of geometric patterns in gaze data. This made it possible to algorithmically detect solution strategies associated with certain gaze patterns, such as slope triangles. We used temporal networks capable of storing temporal sequences of fixations. In addition, third-order time-aggregated networks can select temporal paths consisting of three points. Together with geometric criteria, this allows the detection of slope triangles. We believe that this approach using temporal networks is completely new. So far, we detected slope triangles with a precision 0.60 of and a recall of 0.67. These values are not yet suitable for use in schools, but future optimized systems can support teachers here. In the future, it will be possible to investigate which other gaze patterns are suitable for detection using temporal networks. The specific use and possible added value for schools must also be empirically investigated.

In addition to detecting geometric gaze patterns, we conducted an exploratory study to test the ability of network metrics to discriminate between correct and incorrect solvers. First network metrics have been identified. On this basis, predictors can be developed in the future that can recognize students' problems by calculating network metrics and, if possible, provide individual help based on them. This will provide optimal support for the learning process. In addition, it would be interesting to investigate whether the significant differences occur while the task is being solved by analyzing the incomplete gaze data graph.

Our basic research has provided results and valuable ideas for the future of ALS. It is worth noting that our basic research on adaptive systems does not involve artificial intelligence at all. This has the advantage that the processes for recognizing geometric patterns or classifying correct and incorrect learners based on network metrics remain completely transparent. Whether AI systems can also perform these characterizations remains to be seen. However, it is already clear that they cannot show exactly how this characterization was performed. It was precisely this "exactly" that was too important to us in the initial phase, which is why AI was not used.

It should be particularly emphasized that all data sets and a large part of the analysis of the research presented here are available to the reader (cf. Knippertz (2024a)(replication study), Becker, Knippertz, et al. (2022) (ET study on graphs), Knippertz et al. (2024) (ET study on formulas), Knippertz (2024c) (material of teacher training), Knippertz (2024b) (detection of slope triangles), Knippertz (2024d) (differentiating network metrics)). This enables a transparent presentation of the research results and space for reproduction and replication.

In summary, this work has shown that the adequate transfer of mathematical content to physics is a problem that affects countries and grade levels, and is also manifested in ET data. To solve this problem, a teacher training course has been developed that sensitizes teachers to transfer difficulties and provides new ideas for good teaching. To improve future learning systems in this area, two perspectives have been developed as part of this work.

In a nutshell, this thesis consists of three main research findings:

1. We show that students' difficulties with linear functions (e.g. problems with the physical interpretation of negative slopes) persist across two countries and grade levels.

2. Using triangulation of ET and RTA data, we show that cognitive processes in solving linear function problems are context-dependent, i.e., they differ in mathematics and physics.

3. For the first time, we have used network analysis methods to algorithmically detect solution strategies based on geometric patterns (e.g. slope triangles) from gaze data and to differentiate between false and correct solvers – potential tools for improving ALS.

A. Appendix

A.1. Questionaire and Full Test Instrument (in German)

Questionaire about demographic data in German

The following variables were asked in the questionnaire on the next page:

0. Number of test person,
1. age in years,
2. gender (m,f,d),
3. mother tongue (German: yes, no),
4. repetition of the 11th school year (yes, no),
5. choice of math course (basic, advanced),
6. choice of physics course (basic, advanced, neither),
7. last grade in math (grade scale: 1-very good to 6-insufficient),
8. last grade in physics (grade scale: 1-very good to 6-insufficient),
9. last grade in German (grade scale: 1-very good to 6-insufficient),
10. if you wear glasses (yes, no),
11. motivation (very motivated, rather motivated, rather not motivated, not motivated at all).

Persönliche Angaben, Testperson Nr.: ______

Persönliche Angaben

1. **Wie alt sind Sie?** ______ Jahre.
2. **Ihr Geschlecht:** ○ männlich ○ weiblich ○ divers
3. **Ist Deutsch Ihre Muttersprache?** ○ Ja. ○ Nein.
4. **Wiederholen Sie gerade das 11. Schuljahr?** ○ Ja. ○ Nein.
5. **Mathekurs:** ○ Grundkurs ○ Leistungskurs
6. **Physikkurs:** ○ Grundkurs ○ Leistungskurs ○ weder noch
7. **Ihre letzte Zeugnisnote in Mathematik:** ______ Notenskala 1-6 (ohne +/-)
8. **Ihre letzte Zeugnisnote in Physik:** ______ Notenskala 1-6 (ohne +/-)
9. **Ihre letzte Zeugnisnote in Deutsch:** ______ Notenskala 1-6 (ohne +/-)
10. **Tragen Sie eine Brille?** ○ Ja. ○ Nein.
11. **Wie motiviert sind Sie für den folgenden Test?**
    - ○ sehr motiviert
    - ○ eher motiviert
    - ○ eher nicht motiviert
    - ○ überhaupt nicht motiviert

Figure A 1: Questionaire about demographic data in German.

Full test instrument in German

This test instrument was used for the ET study. It contains all 24 items from the validated test instrument by Ceuppens et al. (2019) in randomized order, starting with the mathematics items. Each item requires either a written solution on paper (cf. Figure A 3) or clicking on the correct solution with the mouse (cf. Figure A 8). After each item, the test person is asked how confident he/she was in answering the question (possible answers: very confident, confident, uncertain, guessed, cf. Figure A 4).

A1) Die Abbildung zeigt den Graphen einer Funktion f.
Bestimmen Sie die Steigung.

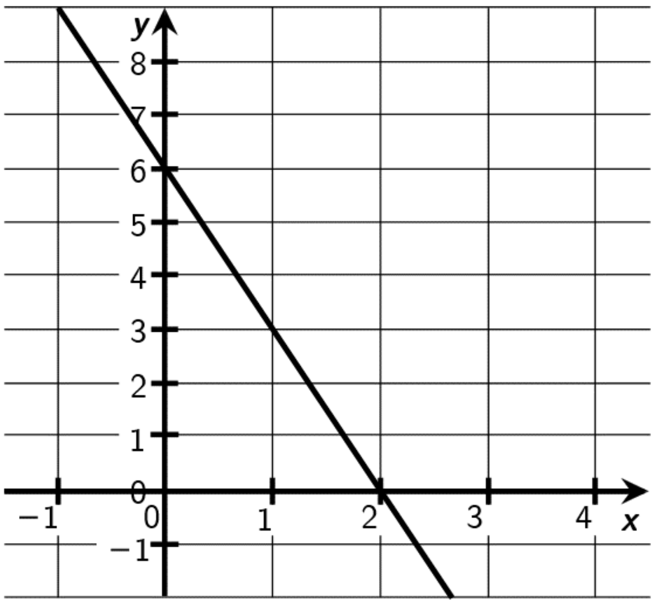


Figure A 2: Item A1 of the ET study translated into German (corresponds to item M8 of the test instrument by Ceuppens et al. (2019)).

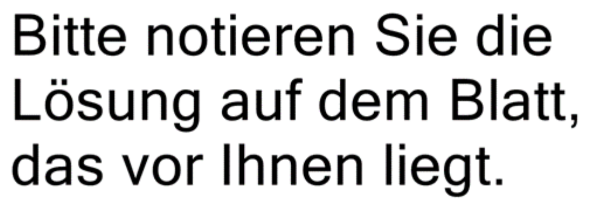


Figure A 3: Request for a written solution to the ET study in German.

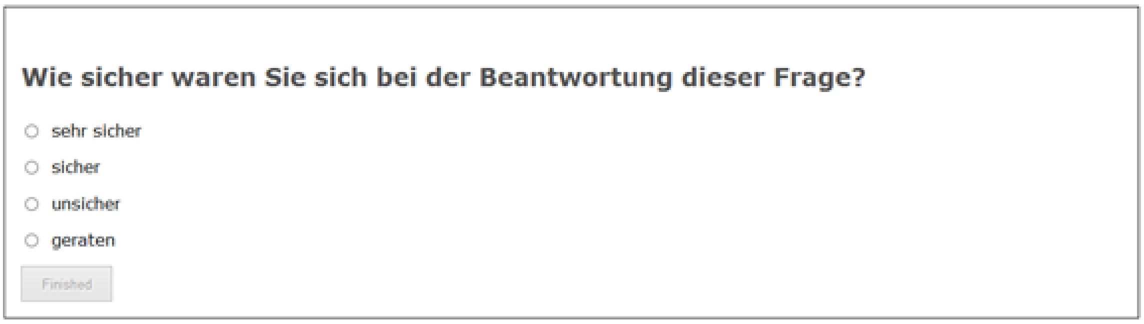


Figure A 4: Indication of certainty when answering the question in German.

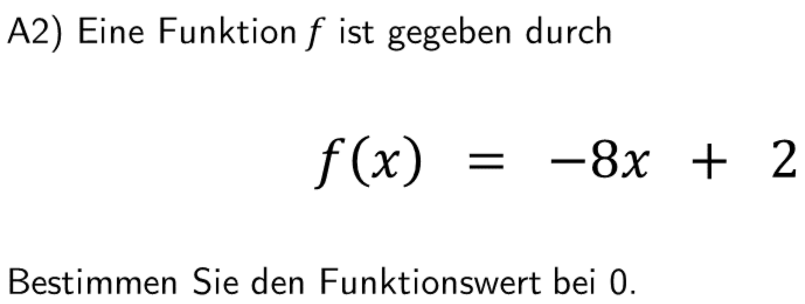


Figure A 5: Item A2 of the ET study translated into German (corresponds to item M11 of the test instrument by Ceuppens et al. (2019)).

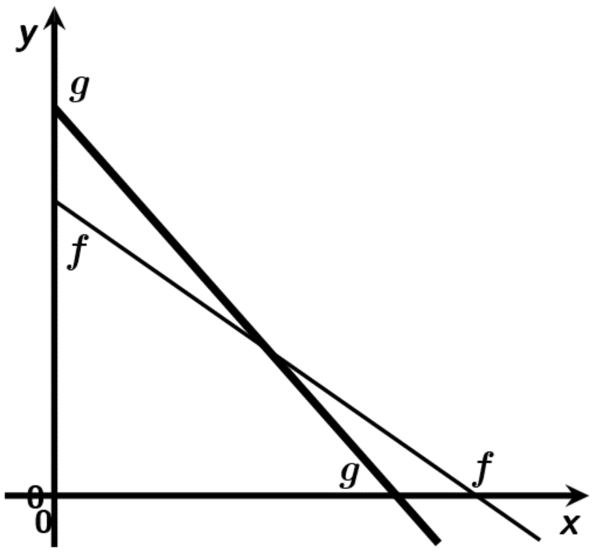
A3) Eine Funktion f ist gegeben durch

$$f(x) \;=\; -6x \;+\; 12$$

Bestimmen Sie die Steigung.

Figure A 6: Item A3 of the ET study translated into German (corresponds to item M12 of the test instrument by Ceuppens et al. (2019)).

A4) Die Abbildung zeigt die Graphen von zwei Funktionen f und g. Welche dieser Funktionen hat den größten Funktionswert bei 0?



a.) Funktion f

b.) Funktion g

c.) Funktion f und g haben den gleichen Funktionswert bei 0

Figure A 7: Item A4 of the ET study translated into German (corresponds to item M3 of the test instrument by Ceuppens et al. (2019)).

Figure A 8: Specification of the solution by mouse click for item A4.

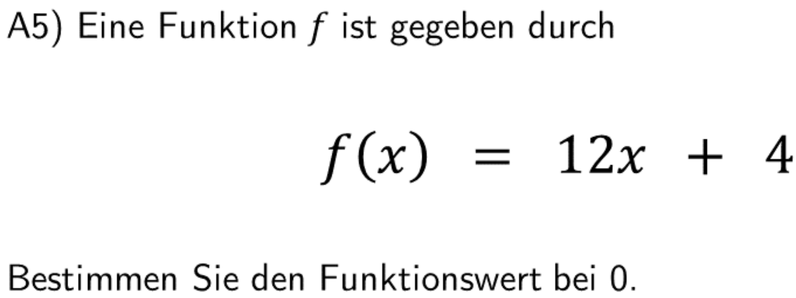


Figure A 9: Item A5 of the ET study translated into German (corresponds to item M9 of the test instrument by Ceuppens et al. (2019)).

A6) Die Abbildung zeigt den Graphen einer Funktion f.
Bestimmen Sie die Steigung.

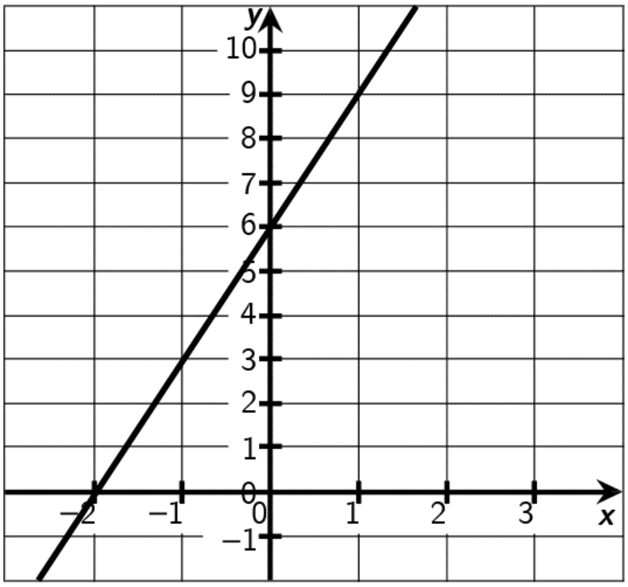
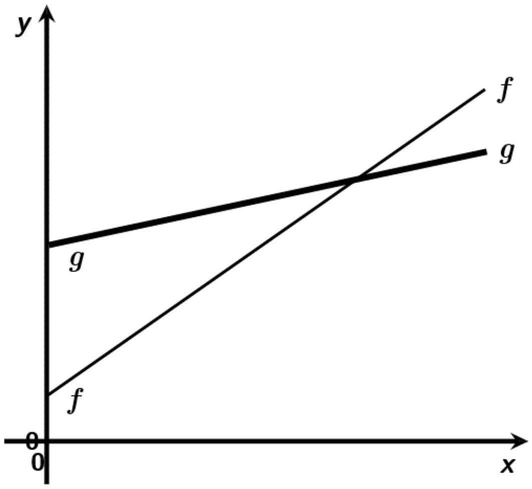


Figure A 10: Item A6 of the ET study translated into German (corresponds to item M6 of the test instrument by Ceuppens et al. (2019)).

A7) Die Abbildung zeigt die Graphen von zwei Funktionen f und g.
Welche dieser Funktionen hat den größten Funktionswert bei 0?



a.) Funktion f

b.) Funktion g

c.) Funktion f und g haben den gleichen Funktionswert bei 0

Figure A 11: Item A7 of the ET study translated into German (corresponds to item M1 of the test instrument by Ceuppens et al. (2019)).

A8) Die Abbildung zeigt den Graphen einer Funktion f.
Bestimmen Sie den Funktionswert bei 0.

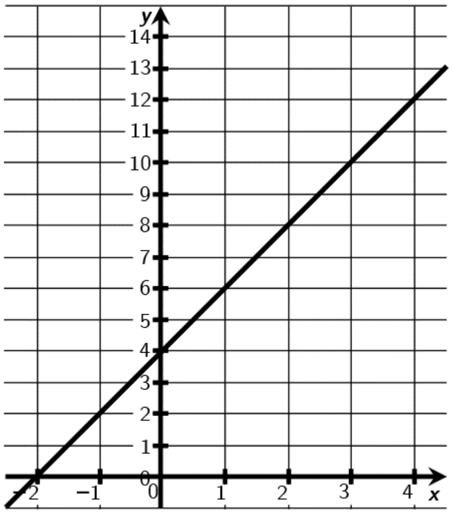


Figure A 12: Item A8 of the ET study translated into German (corresponds to item M5 of the test instrument by Ceuppens et al. (2019)).

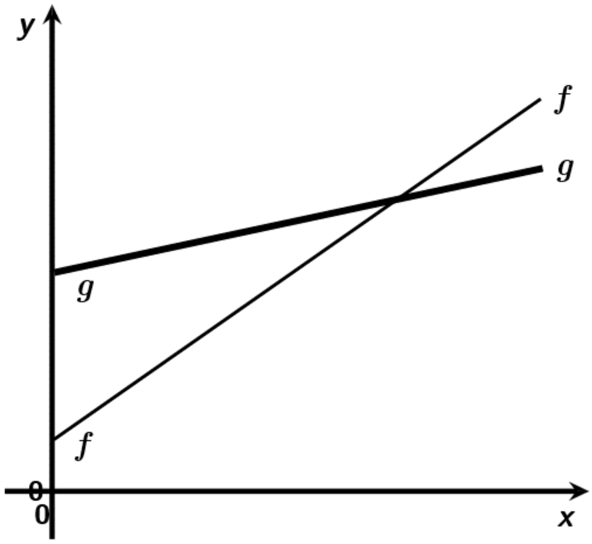
A9) Eine Funktion f ist gegeben durch

$$f(x) = 6x + 2$$

Bestimmen Sie die Steigung.

Figure A 13: Item A9 of the ET study translated into German (corresponds to item M10 of the test instrument by Ceuppens et al. (2019)).

A10) Die Abbildung zeigt die Graphen von zwei Funktionen f und g. Welche dieser Funktionen hat die größte Steigung?



a.) Funktion f

b.) Funktion g

c.) Funktion f und g haben die gleiche Steigung

Figure A 14: Item A10 of the ET study translated into German (corresponds to item M2 of the test instrument by Ceuppens et al. (2019)).

A11) Die Abbildung zeigt den Graphen einer Funktion f.
Bestimmen Sie den Funktionswert bei 0.

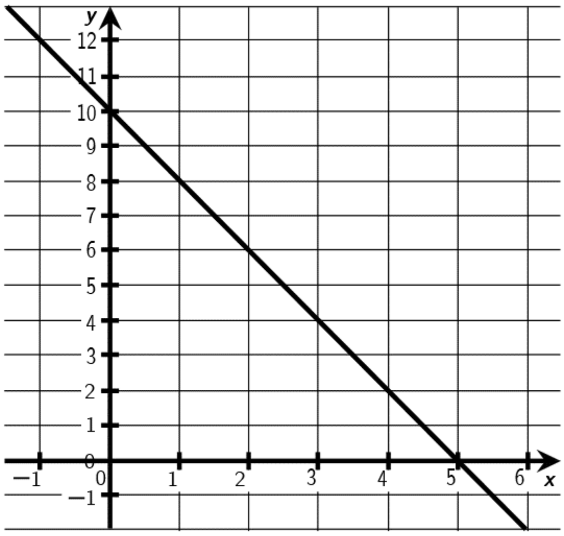
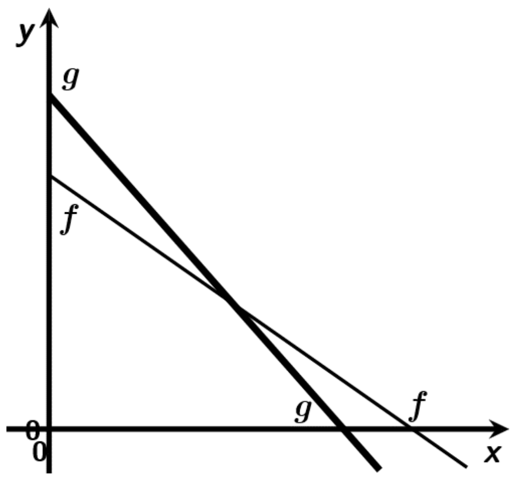


Figure A 15: Item A11 of the ET study translated into German (corresponds to item M7 of the test instrument by Ceuppens et al. (2019)).

A12) Die Abbildung zeigt die Graphen von zwei Funktionen f und g.
Welche dieser Funktionen hat die größte Steigung?



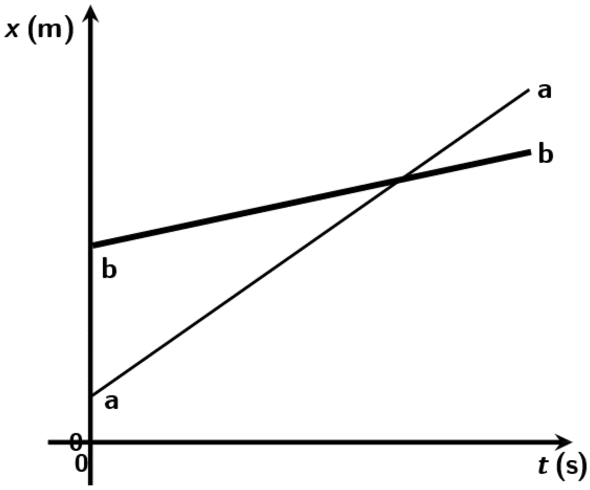
a.) Funktion f

b.) Funktion g

c.) Funktion f und g haben die gleiche Steigung

Figure A 16: Item A12 of the ET study translated into German (corresponds to item M4 of the test instrument by Ceuppens et al. (2019)).

A13) Zwei Autos a und b fahren auf einer geraden Straße. Der Graph zeigt die Position x der Autos als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Welches Auto hat die größte Geschwindigkeit?



a.) Auto a

b.) Auto b

c.) Auto a und b haben die gleiche Geschwindigkeit

Figure A 17: Item A13 of the ET study translated into German (corresponds to item K2 of the test instrument by Ceuppens et al. (2019)).

A14) Ein Radfahrer fährt auf einem geraden Radweg. Der Graph zeigt die Position x des Radfahrers als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Bestimmen Sie die Ausgangsposition des Radfahrers.

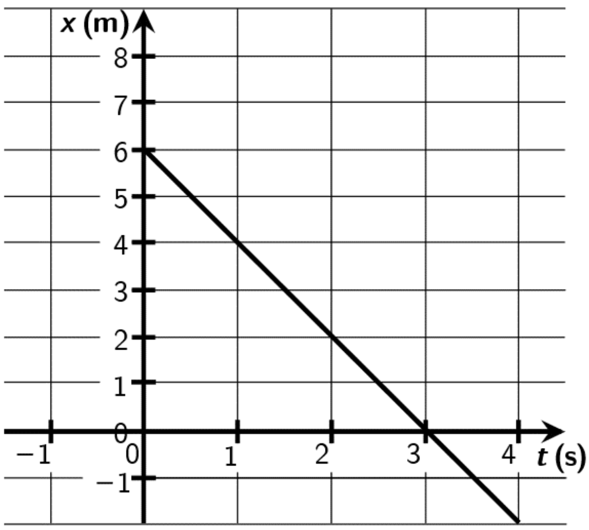


Figure A 18: Item A14 of the ET study translated into German (corresponds to item K7 of the test instrument by Ceuppens et al. (2019)).

A15) Ein Auto fährt auf einer geraden Straße. Die Position x des Autos als Funktion der Zeit t ist gegeben durch

$$x = 5 - 10t$$

Die Position ist in Metern, die Zeit in Sekunden.
Bestimmen Sie die Ausgangsposition des Autos.

Figure A 19: Item A15 of the ET study translated into German (corresponds to item K11 of the test instrument by Ceuppens et al. (2019)).

A16) Ein Radfahrer fährt auf einem geraden Radweg. Die Position x des Radfahrers als Funktion der Zeit t ist gegeben durch

$$x = 2 + 6t$$

Die Position ist in Metern, die Zeit in Sekunden.
Bestimmen Sie die Ausgangsposition des Radfahrers.

Figure A 20: Item A16 of the ET study translated into German (corresponds to item K9 of the test instrument by Ceuppens et al. (2019)).

A17) Ein Auto fährt auf einer geraden Straße. Der Graph zeigt die Position x des Autos als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Bestimmen Sie die Geschwindigkeit des Autos.

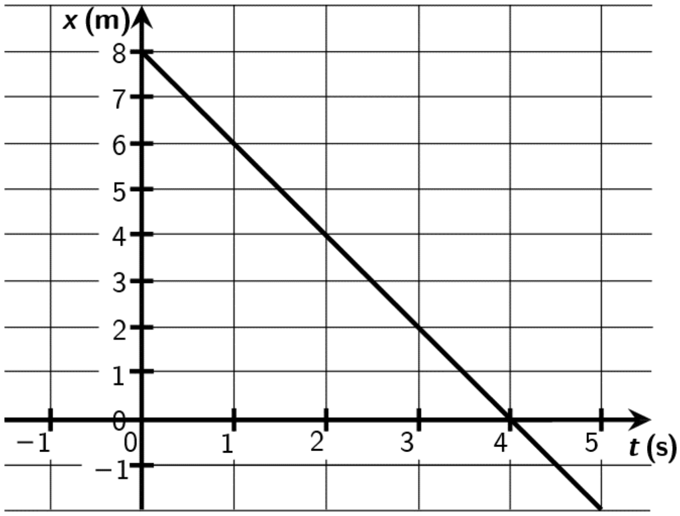
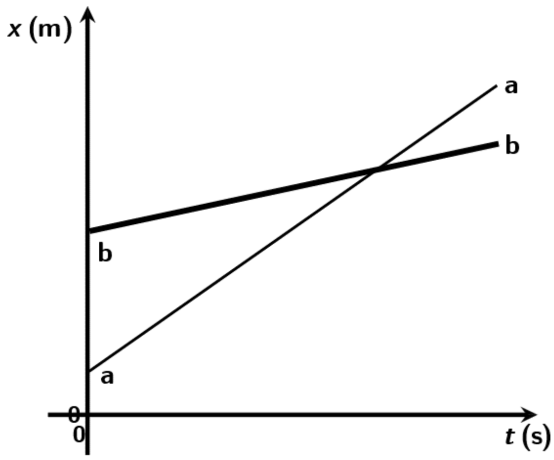


Figure A 21: Item A17 of the ET study translated into German (corresponds to item K8 of the test instrument by Ceuppens et al. (2019)).

A18) Zwei Radfahrer a und b fahren auf einem geraden Radweg. Der Graph zeigt ihre Position x als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Welcher Radfahrer hat die größte Ausgangsposition?



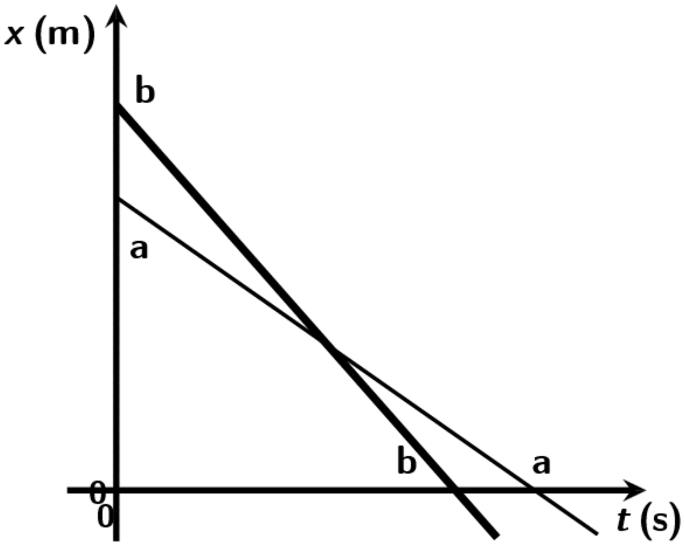
a.) Radfahrer a

b.) Radfahrer b

c.) Radfahrer a und b haben die gleiche Ausgangsposition

Figure A 22: Item A18 of the ET study translated into German (corresponds to item K1 of the test instrument by Ceuppens et al. (2019)).

A19) Zwei Radfahrer a und b fahren auf einem geraden Radweg. Der Graph zeigt ihre Position x als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Welcher Radfahrer hat die größte Geschwindigkeit?



a.) Radfahrer a

b.) Radfahrer b

c.) Radfahrer a und b haben die gleiche Geschwindigkeit

Figure A 23: Item A19 of the ET study translated into German (corresponds to item K4 of the test instrument by Ceuppens et al. (2019)).

A20) Ein Zug fährt auf einer geraden Schiene. Die Position x des Zuges als Funktion der Zeit t ist gegeben durch

$$x = 4 + 8t$$

Die Position ist in Metern, die Zeit in Sekunden.
Bestimmen Sie die Geschwindigkeit des Zuges.

Figure A 24: Item A20 of the ET study translated into German (corresponds to item K10 of the test instrument by Ceuppens et al. (2019)).

A21) Ein Auto fährt auf einer geraden Straße. Der Graph zeigt die Position x des Fahrzeugs als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Bestimmen Sie die Ausgangsposition des Autos.

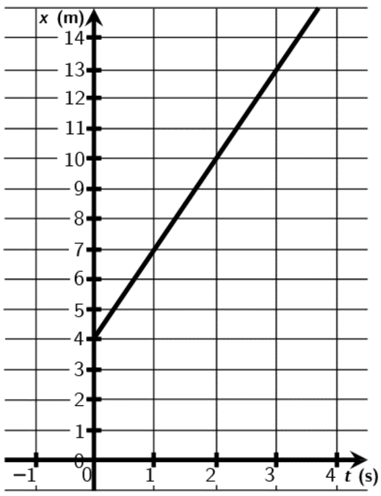


Figure A 25: Item A21 of the ET study translated into German (corresponds to item K5 of the test instrument by Ceuppens et al. (2019)).

A22) Ein Zug fährt auf einer geraden Schiene. Der Graph zeigt die Position x des Zuges als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Bestimmen Sie die Geschwindigkeit des Zuges.

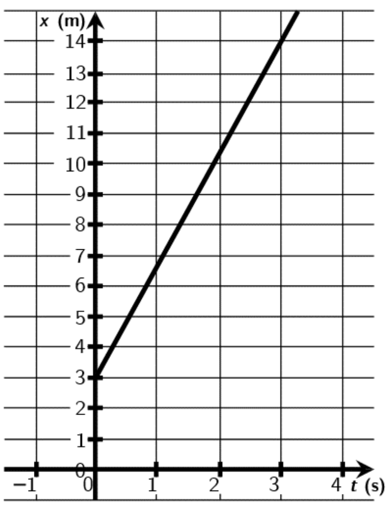


Figure A 26: Item A22 of the ET study translated into German (corresponds to item K6 of the test instrument by Ceuppens et al. (2019)).

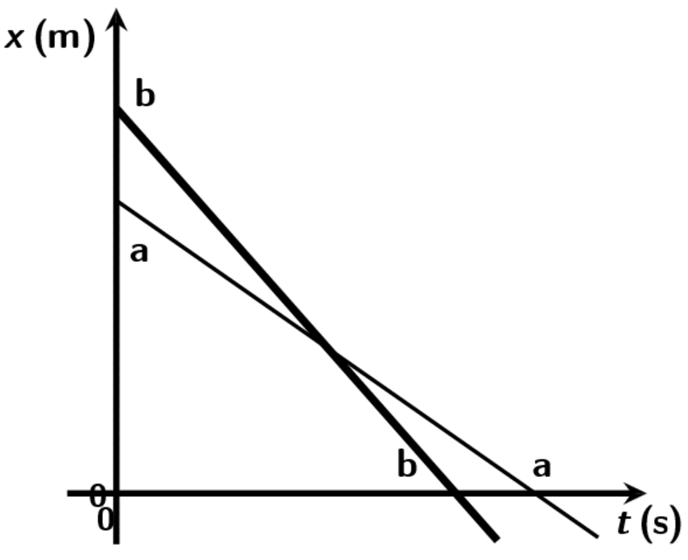
A23) Ein Auto fährt auf einer geraden Straße. Die Position x des Autos als Funktion der Zeit t ist gegeben durch

$$x = 3 - 12t$$

Die Position ist in Metern, die Zeit in Sekunden.
Bestimmen Sie die Geschwindigkeit des Autos.

Figure A 27: Item A23 of the ET study translated into German (corresponds to item K12 of the test instrument by Ceuppens et al. (2019)).

A24) Zwei Radfahrer a und b fahren auf einem geraden Radweg. Der Graph zeigt ihre Position x als Funktion der Zeit t. Die Position ist in Metern, die Zeit in Sekunden. Welcher Radfahrer hat die größte Ausgangsposition?



a.) Radfahrer a

b.) Radfahrer b

c.) Radfahrer a und b haben die gleiche Ausgangsposition

Figure A 28: Item A24 of the ET study translated into German (corresponds to item K3 of the test instrument by Ceuppens et al. (2019)).

A.2. Categorization Scheme, Descriptive Results and GEE Analysis

Categorization Scheme by Ceuppens et al. (2019)

The categorization scheme is split by concept (y intercept or initial position and slope or velocity). The first seven categories are shared between both concepts followed by two and four concept specific categories. Answers usually fall in only a single category, but a non-negligible number of questions had answers which made use of multiple strategies, in which case they were put into all suitable categories.

I1/S1: Location in an equation: A coefficient in the equation is marked by underlining it or encircling it. Alternatively, the student explicitly writes that, e. g., "6" is the answer because it is located before the "x."

I2/S2: Identification of the x intercept: Instead of identifying the y intercept or slope they identify the x intercept (graphical or symbolic).

I3/S3: Construct an equation: Change of representation. When presented with a graph, an equation was constructed.

I4/S4: Construct a graph: Change of representation. When presented with an equation, a graph was constructed. This need not be a detailed one, a simple sketch suffices.

I5/S5: Construct a table: Change of representation. A table is constructed with two or more data points.

I6: Intersection with vertical axis: When a graph is given or constructed, the y intercept is marked by, e. g., encircling it or drawing an arrow toward it. Often this is accompanied by written explanations such as "This is where the line/graph/cyclist starts/begins."

I7: Calculate a specific function value: When an equation is given or constructed, the value of the dependent variable is calculated for a specific value of the independent variable, often with $x = 0$ or $x = 1$.

S6: Reasoning with "steepness": Qualitative reasoning concerning the steepness of the graph. Often wording similar to "Graph a is steeper than graph b" is used.

S7: Drawing a triangle on a graph: Drawing a triangle with a vertical and a horizontal side under or above the line of a graph to indicate the step change in both directions. Often these triangles are accompanied by the change in value along each axis with notations such as "+1" and "-2." The size of the triangle does not matter.

S8: Ratio of differences: Calculation of the slope using Δy Δx or $(y_2 - y_1)/(x_2 - x_1)$, or written explanations such as "In 1s the cyclist traveled 2m;" "If x increases with 1, then y increases with 3;" "If you go one unit to the right, then you go 3 units up." A specific case in this category is when the ratio of the y value of the y intercept over the x value of the x intercept is calculated, which is essentially the ratio of the differences from the intercepts with respect to the origin. More qualitative explanations also fit in this category, such as: "In the same time, cyclist b traveled more meters;" "Cyclist b starts later, but still overtakes cyclist a."

S9: Ratio of coordinates: The ratio y/x of some coordinate (x, y) is calculated. This includes cases in which the correct formula is written, e. g., $v = \Delta x/\Delta t$, but actually $v = x/t$ is calculated, signaling the disregard or the misunderstanding of Δ in the formula. Often $f(1)/1$ or $F(x = 1)$ is calculated. Also written explanations such as "The y value of function f is larger than that of function g for the same x value" with comparison questions are in this category.

I99/S99: Other: All answers which are less frequent and do not fit in any of the other categories. For example, confusion of y intercept and slope; calculations with the coefficients of a formula such as $a \cdot b$ or $a + b$ in $f(x) = ax + b$; identifying the slope as $6x$ in $f(x) = 6x + 2$; in the case of a y intercept question, providing coordinates such as $(0, 2)$ given the equation $f(x) = 6x + 2$, which is always considered incorrect since the questions ask for the function value or the initial position; or in case of a y intercept question, providing a coordinate constructed of the x intercept and the y intercept such as $(-\frac{1}{3}, 2)$ for $f(x) = 6x+2$.

Remark from Ceuppens et al. (2019): Categories I3/S3, I4/S4, and I5/S5 all describe a change in representation, these categories are always accompanied by another strategy to fully answer the question. We deliberately keep these categories separate so we can compare their prevalences. Categories S7 and S8 are both essentially a ratio of differences, but our data strongly suggest that it is useful to distinguish the two.

Descriptive Results of the Eye-Tracking Study

Figure A.1 shows the results of the item analysis of all 24 test items. The test items are divided into 12 item pairs, each consisting of a mathematics item and a kinematics item (all item pairs can be viewed in the Appendix A.1). The item characteristics of the test items are described on the top and right-hand sides of the diagram. For example, in item pair no. 3 (IP 3), participants were asked to compare the y-axis intercepts of two linearly decreasing functions in mathematics, or the initial positions of two linearly decreasing $x(t)$-graphs in kinematics, respectively. In IP 10, participants were asked to determine the slope from a linear function equation in mathematics, or the velocity of an object from a linear equation of motion in kinematics. Except for IP 1 and IP 3, the item difficulties (proportion of participants who solved the item correctly) of the mathematics items are higher than those of the respective similar kinematics items. Extremely difficult items are kinematic items dealing with negative velocities and with the

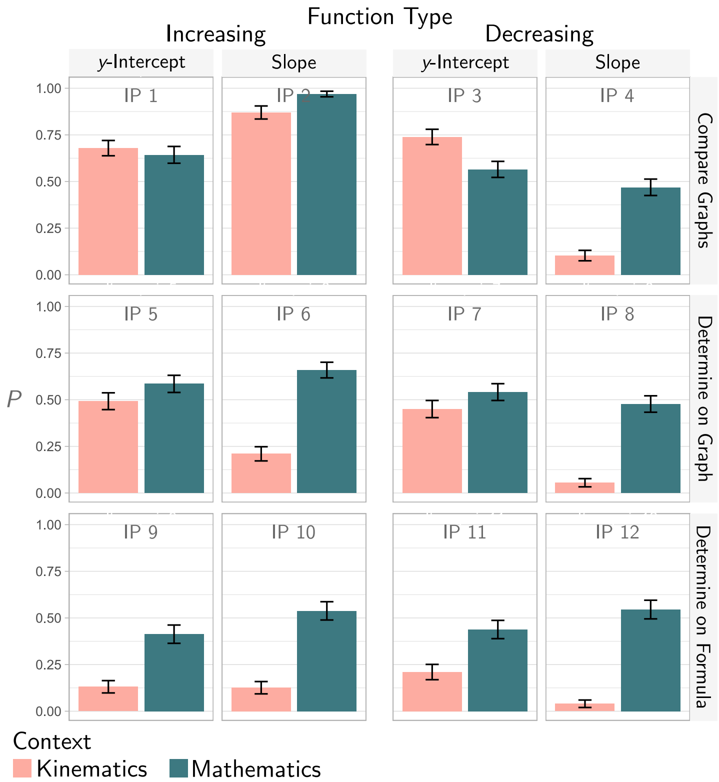


Figure A.1.: Comparison of item difficulty level P of all item pairs (IP) in the test instrument.

representation form formula ($P_{\text{K4}} = 0.10$, $P_{\text{K8}} = 0.06$, $P_{\text{K9}} = 0.13$, $P_{\text{K10}} = 0.13$, $P_{\text{K11}} = 0.21$, $P_{\text{K12}} = 0.04$, Bortz, Schuster (2010)). Whereas the items K2 and M2 are extremely easy items ($P_{\text{K2}} = 0.87$, $P_{\text{M2}} = 0.97$, Bortz, Schuster (2010)). Here, the slope of a linearly increasing function (M2) or the velocity of a linearly increasing time-position function (K2) had to be determined.

Context: The item difficulty is higher in the mathematical context than in the kinematic context for all pairs of items examined (e. g. $P_{\text{M5}} = 0.57$, $P_{\text{K5}} = 0.47$). In concrete terms, this means that the mathematics task was always solved correctly more often than the similar kinematics task.

Slope: Other things being equal (context, calculation), a task with a positive slope was always solved correctly more often than the corresponding item with a negative slope (e. g. $P_{\text{M5}} = 0.57$, $P_{\text{M7}} = 0.53$).

Concept: The effect of the concept to be recorded (y-intercept or slope) differs between the two contexts: The item difficulty of the kinematics tasks for calculating the initial position of the cyclist (y-axis intercept) is always higher than that of the corresponding task for calculating the speed (slope) (e. g. $P_{\text{K5}} = 0.47$, $P_{\text{K6}} = 0.19$). In the mathematical context, it depends on the type of slope (positive / negative) whether the item difficulty of the slope tasks (for positive slopes;

$P_{\mathrm{M5}} = 0.57$, $P_{\mathrm{M6}} = 0.64$) or that of the y-intercept task is higher (for negative slopes; $P_{\mathrm{M7}} = 0.53$, $P_{\mathrm{M8}} = 0.46$).

GEE Analysis

Table A.1.: Coefficients section of GEE analysis with basic model and correlation structure *exchangeable* for 11th graders.

Coefficients	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-1.17489	0.30684	14.661	<0.001
genf	0.48277	0.32305	2.233	0.135
genm	0.85450	0.31782	7.229	<0.010
contextm	0.96133	0.06682	207.001	<0.001
concepty-intercept	0.43554	0.06210	49.194	<0.001
slope_sign+	0.66932	0.04017	277.625	<0.001
Qtypedetermine_formula	-1.12063	0.06758	274.939	<0.001
Qtypedetermine_graph	-0.91396	0.05734	254.087	<0.001

Table A.2.: Coefficients section of GEE analysis with basic model without gender and correlation structure *exchangeable* for 9th graders.

Coefficients	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-1.0082	0.0790	163.08	<0.001
contextm	0.2492	0.0872	8.17	0.010
concepty-intercept	0.2752	0.0858	10.29	0.010
slope_sign+	0.7043	0.0515	187.26	<0.001
Qtypedetermine_formula	-1.8383	0.1145	257.61	<0.001
Qtypedetermine_graph	-1.3723	0.0940	212.90	<0.001

Table A.3.: Coefficients section of GEE analysis with basic model without gender and correlation structure *exchangeable* for 11th graders.

Coefficients	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	-1.0082	0.0790	163.08	<0.001
contextm	0.2492	0.0872	8.17	<0.010
concepty-intercept	0.2752	0.0858	10.29	<0.010
slope_sign+	0.7043	0.0515	187.26	<0.001
Qtypedetermine_formula	-1.8383	0.1145	257.61	<0.001
Qtypedetermine_graph	-1.3723	0.0940	212.90	<0.001

A.3. Alternative Visualization of a Temporal Network

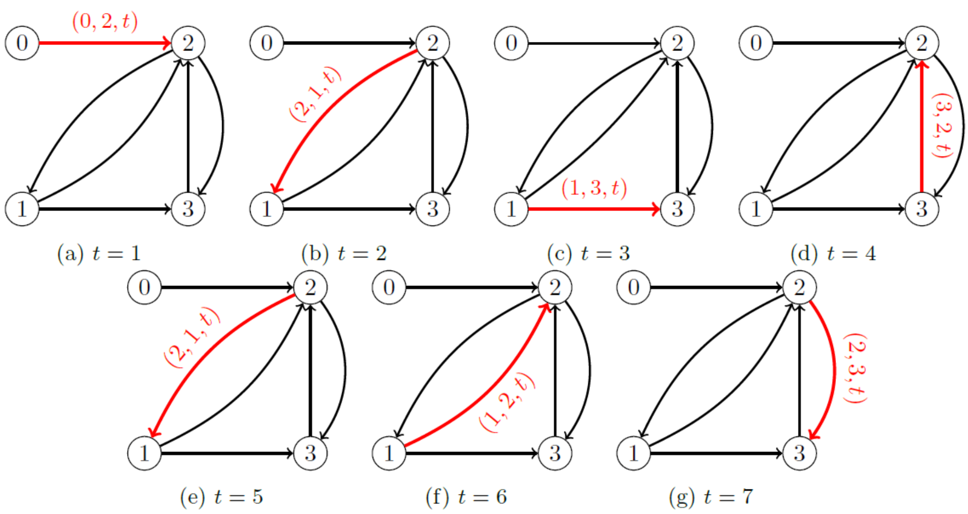


Figure A.2.: Alternative visualization of the temporal network based on sample data in Table 4.1.

Bibliography

Acevedo Nistal, A., Van Dooren, W., Verschaffel, L. (2012). "What counts as a flexible representational choice? An evaluation of students' representational choices to solve linear function problems." *Instructional Science* 40, pp. 999–1019.

Ainsworth, S. (2006). "DeFT: A conceptual framework for considering learning with multiple representations." *Learning and Instruction* 16.3, pp. 183–198.

Alemdag, E., Cagiltay, K. (2018). "A systematic review of eye tracking research on multimedia learning." *Computers and Education* 125, pp. 413–428.

Alhadreti, O., Elbabour, F., Mayhew, P. (2017). "Eye tracking in retrospective think-aloud usability testing: Is there added value?" *Journal of Usability Studies* 12.3, pp. 95–110.

Andrá, C., Lindström, P., Arzarello, F., Holmqvist, K., Robutti, O., Sabena, C. (2015). "Reading mathematics representations: An eye-tracking study." *International Journal of Science and Mathematics Education* 13.2, pp. 237–259.

Arge, M. (2016). "Generalized estimating equations: an overview and application in IndiMed study." MA thesis. University of Tartu, Faculty of Science, Technology, Institute of Mathematics, and Statistics.

Baddeley, A. (1992). "Working memory." *Science* 255.5044, pp. 556–559.

Bayazit, A., Askar, P., Cosgun, E. (2014). "Predicting learner answers correctness through eye movements with random forest." *Educational data mining: Applications and trends*, pp. 203–226.

[Dataset] Becker, S., Knippertz, L., Ruzika, S., Kuhn, J. (2022). *Data from an eye-tracking based study of solution strategies in similar mathematical and physical tasks.* Zenodo. DOI: `10.5281/zenodo.7426806`.

Becker, S., Knippertz, L., Ruzika, S., Kuhn, J. (2023). "Persistence, context, and visual strategy of graph understanding: Gaze patterns reveal student difficulties in interpreting graphs." *Physical Review Physics Education Research* 19.2, pp. 1–12.

Becker, S., Küchemann, S., Klein, P., Lichtenberger, A., Kuhn, J. (2022). "Gaze patterns enhance response prediction: More than correct or incorrect." *Physical Review Physics Education Research* 18.2, pp. 1–11.

Beichner, R. J. (1994). "Testing student interpretation of kinematics graphs." *American Journal of Physics* 62.8, pp. 750–762.

Benjamini, Y., Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300.

Berg, C. A., Phillips, D. G. (1994). "An investigation of the relationship between logical thinking structures and the ability to construct and interpret line graphs." *Journal of research in science teaching* 31.4, pp. 323–344.

Bicer, A., Bicer, A. (2022). "Understanding young students' mathematical creative thinking processes through eye-tracking-stimulated recall interview." *Mathematics Education Research Journal* 35, pp. 1–39.

Bieber, G., Egyptien, E. L., Klein, G., Oechslein, K., Pikowsky, B. (2018). "Positionspapier der Landesinstitute und Qualitätseinrichtungen der Länder zum Transfer von Forschungswissen." *Deutscher Bildungsserver* 10.

Bortz, J., Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. 7. Auflage. Springer-Lehrbuch. Springer Berlin Heidelberg.

Brandes, U. (2008). "On variants of shortest-path betweenness centrality and their generic computation." *Social Networks* 30.2, pp. 136–145.

Brennan, R. L., Prediger, D. J. (1981). "Coefficient kappa: Some uses, misuses, and alternatives." *Educational and Psychological Measurement* 41.3, pp. 687–699.

Brückner, S., Schneider, J., Zlatkin-Troitschanskaia, O., Drachsler, H. (2020). "Epistemic network analyses of economics students' graph understanding: An eye-tracking study." *Sensors* 20.23, pp. 6908–6941.

Burton, P., Gurrin, L., Sly, P. (1998). "Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling." *Statistics in Medicine* 17.11, pp. 1261–1291.

Carli, M., Lippiello, S., Pantano, O., Perona, M., Tormen, G. (2020). "Testing students ability to use derivatives, integrals, and vectors in a purely mathematical context and in a physical context." *Physical Review Physics Education Research* 16.1, pp. 1–17.

Carrera-Fernandez, M. J., Guardia-Olmos, J., Peró-Cebollero, M. (2014). "Qualitative methods of data analysis in psychology: An analysis of the literature." *Qualitative Research* 14.1, pp. 20–36.

Ceuppens, S., Bollen, L., Deprez, J., Dehaene, W., Cock, M. de (2019). "9th grade students' understanding and strategies when solving $x(t)$ problems in 1d kinematics and $y(x)$ problems in mathematics." *Physical Review Physics Education Research* 15.1, pp. 1–22.

Chambers, J., Hastie, T. (1992). "Statistical models. Chapter 2 of statistical models in S." *Statistical models, Wadsworth and Brooks/Cole* 73.

Chandler, P., Sweller, J. (1991). "Cognitive load theory and the format of instruction." *Cognition and instruction* 8.4, pp. 293–332.

Christensen, W., Thompson, J. (2012). "Investigating graphical representations of slope and derivative without a physics context." *Physical Review Special Topics - Physics Education Research* 8.2, pp. 1–5.

Cohen, J. (1960). "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement* 20.1, pp. 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences second edition.* Second. Lawrence Erlbaum Associates, pp. 285–288.

Comaniciu, D., Meer, P. (2002). "Mean shift: A robust approach toward feature space analysis." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5, pp. 603–619.

Dilling, F. (2019). "Vektoren in der mathematik-und physikdidaktischen Forschung – Stoffdidaktischverbindende Ansätze zwischen linearer Algebra und klassischer Mechanik." *PhyDid B-Didaktik der Physik-Beiträge zur DPG-Frühjahrstagung.*

Ding, L., Beichner, R. (2009). "Approaches to data analysis of multiple-choice questions." *Physical Review Special Topics—Physics Education Research* 5.2, pp. 1–17.

Doran, R. L. (1980). *Basic measurement and evaluation of science instruction.* ERIC.

Duval, R. (2006). "A cognitive analysis of problems of comprehension in a learning of mathematics." *Educational Studies in Mathematics* 61.1-2, pp. 103–131.

Duval, R. (2017). *Understanding the mathematical way of thinking-the registers of semiotic representations.* Springer.

Efron, B. (1987). "Better bootstrap confidence intervals." *Journal of the American Statistical Association* 82.397, pp. 171–185.

Elling, S., Lentz, L., Jong, M. de (2011). "Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1161–1170.

Engelhardt, P. V. (2009). "An introduction to classical test theory as applied to conceptual multiple-choice tests." *Getting started in PER* 2.1, pp. 1–40.

Ericsson, K. A. (2017). *A companion to cognitive science.* Wiley Online Library. Chap. Protocol Analysis, pp. 425–432.

Even, R. (1998). "Factors involved in linking representations of functions." *The Journal of Mathematical Behavior* 17.1, pp. 105–121.

Eye, A. von (2006). "An alternative to Cohen's κ." *European Psychologist* 11.1, pp. 12–24.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., Tutz, G. (2016). *Statistik: Der Weg zur Datenanalyse.* Springer-Verlag.

Ferguson, G. A. (1949). "On the theory of test discrimination." *Psychometrika* 14.1, pp. 61–68.

Fitting, N., Hornung, G. (2021). *HyperDocSystems: Die Idee.* `https://didaktik.chemie.uni-kl.de/` [Accessed: 01.08.2024].

Fitzmaurice, G. M., Laird, N. M., Ware, J. H. (2012). *Applied longitudinal analysis*. John Wiley and Sons.

Ford, C. (2022). *Getting started with generalized estimating equations*. `https://library.virginia.edu/data/articles/getting-started-with-generalized-estimating-equations` [Accessed: April 3, 2024)].

Frege, G. (1970). “The whole number.” *Mind* 79.316, pp. 481–486.

Fritz, C. O., Morris, P. E., Richler, J. J. (2012). “Effect size estimates: current use, calculations, and interpretation.” *Journal of Experimental Psychology: General* 141.1, pp. 2–18.

Fukunaga, K., Hostetler, L. (1975). “The estimation of the gradient of a density function, with applications in pattern recognition.” *IEEE Transactions on information theory* 21.1, pp. 32–40.

Gegenfurtner, A., Lehtinen, E., Jarodzka, H., Säljö, R. (2017). “Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis.” *Computers and Education* 113.15, pp. 212–225.

Geyer, M.-A., Pospiech, G. (2016). “Diagramme im Physikunterricht. Hintergründe und Anregungen zur Förderung des Umgangs mit Diagrammen.” *Naturwissenschaften im Unterricht Physik. Mathematik im Physikunterricht* 153.154, pp. 36–42.

Gray, E. M., Tall, D. O. (1994). “Duality, ambiguity, and flexibility: A “proceptual” view of simple arithmetic.” *Journal for Research in Mathematics Education* 25.2, pp. 116–140.

Green, H. J., Lemaire, P., Dufau, S. (2007). “Eye movement correlates of younger and older adults’ strategies for complex addition.” *Acta Psychologica* 125.3, pp. 257–278.

Guan, Z., Lee, S., Cuddihy, E., Ramey, J. (2006). “The validity of the stimulated retrospective think-aloud method as measured by eye tracking.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1253–1262.

Guillon, Q., Afzali, M. H., Rogé, B., Baduel, S., Kruck, J., Hadjikhani, N. (2015). “The importance of networking in autism gaze analysis.” *PLoS one* 10.10, pp. 1–14.

Hagberg, A., Swart, P., S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hahn, L., Klein, P. (2022). "Eye tracking in physics education research: A systematic literature review." *Physical Review Physics Education Research* 18.1, pp. 1–35.

Han, J., Kamber, M., Pei, J. (2012). "Data mining: Concepts and techniques." Morgan Kaufmann Publishers. Chap. 10 - Cluster Analysis: Basic Concepts and Methods, pp. 443–495.

Handl, A., Kuhlenkasper, T. (2018). *Einführung in die Statistik: Theorie und Praxis mit R*. Springer-Verlag.

Happ, M., Bathke, A. C., Brunner, E. (2019). "Optimal sample size planning for the Wilcoxon-Mann-Whitney test." *Statistics in Medicine* 38.3, pp. 363–375.

Hartmann, M. (2015). "Numbers in the eye of the beholder: What do eye movements reveal about numerical cognition?" *Cognitive Processing* 16.1, pp. 245–248.

Hegarty, M., Mayer, R. E., Monk, C. A. (1995). "Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers." *Journal of Educational Psychology* 87.1, pp. 18–32.

Höffler, T. N., Leutner, D. (2007). "Instructional animation versus static pictures: A meta-analysis." *Learning and instruction* 17.6, pp. 722–738.

Hoffman, J. E., Subramaniam, B. (1995). "The role of visual attention in saccadic eye movements." *Attention, Perception, and Psychophysics* 57, pp. 787–795.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Hyrskykari, A., Ovaska, S., Majaranta, P., Räihä, K.-J., Lehtinen, M. (2008). "Gaze path stimulation in retrospective think-aloud." *Journal of Eye Movement Research* 2.4, pp. 1–18.

Ieva Miseviciute Ph.D., T. A. (2024). *Learn article how do eye trackers work? A tech-savvy walk-through*. `https://www.tobii.com/resource-center/learn-articles/how-do-eye-trackers-work` [Accessed: 01.08.2024].

Inglis, M., Alcock, L. (2012). "Expert and novice approaches to reading mathematical proofs." *Journal for Research in Mathematics Education* 43.4, pp. 358–390.

International GeoGebra Institute Wolfauerstr 90 4040 Linz, A. (2021). *Geogebra - dynamic mathematics for everyone (version 5.0.803)*. http://www.geogebra.org/ [Accessed: 01.08.2024].

Ivanjek, L., Susac, A., Planinic, M., Andrasevic, A., Milin-Sipus, Z. (2016). "Student reasoning about graphs in different contexts." *Physical Review Physics Education Research* 12.1, pp. 1–13.

Jang, M. J. (2011). "Working correlation selection in generalized estimating equations." PhD thesis. The University of Iowa.

Jansen, A. R., Marriott, K., Yelland, G. W. (2007). "Parsing of algebraic expressions by experienced users of mathematics." *European Journal of Cognitive Psychology* 19.2, pp. 286–320.

Just, M. A., Carpenter, P. (1976). "Eye fixations and cognitive processes." *Cognitive Psychology* 8, pp. 441–480.

Kekule, M. (2014). "Students' approaches when dealing with kinematics graphs explored by eye-tracking research method." *Proceedings of the Frontiers in Mathematics and Science Education Research Conference, FISER*, pp. 108–117.

Kendall, M. G. (1962). "Rank correlation methods."

Kiili, K., Ketamo, H., Koivisto, A., Finn, E. (2014). "Studying the user experience of a tablet based math game." *International Journal of Game-Based Learning* 4.1, pp. 60–77.

Klein, P., Küchemann, S., Brückner, S., Zlatkin-Troitschanskaia, O., Kuhn, J. (2019). "Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students." *Physical Review Physics Education Research* 15.2, pp. 1–17.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. Methuenm London, p. 144.

[Dataset] Knippertz, L. (2024a). *Dataset and analysis of a study on linear functions in mathematics and physics*. Zenodo. DOI: 10.5281/zenodo.14354871.

[Dataset] Knippertz, L. (2024b). *Material for detecting slope triangles from et data based on automatic aoi placement and network analysis*. Zenodo. DOI: 10.5281/zenodo.14332928.

[Dataset] Knippertz, L. (2024c). *Material for teacher training on transfer difficulties*. Zenodo. DOI: 10.5281/zenodo.14328495.

[Dataset] Knippertz, L. (2024d). *Material from an exploratory study on the differentiation of correct and incorrect solvers using network metrics*. Zenodo. DOI: 10.5281/zenodo.14330911.

[Dataset] Knippertz, L., Becker-Genschow, S., Ruzika, S., Küchemann, S., Kuhn, J. (2024). *Dataset for paper "Equations – more a matter of context than text"*. Zenodo. DOI: 10.5281/zenodo.10829912.

Kohlhase, A., Kohlhase, M., Ouypornkochagorn, T. (2018). "Discourse phenomena in mathematical documents." *Intelligent Computer Mathematics: 11th International Conference, CICM 2018, Hagenberg, Austria, August 13-17, 2018, Proceedings 11*. Springer, pp. 147–163.

Kraemer, H. C. (1980). "Extension of the kappa coefficient." *Biometrics*, pp. 207–216.

Kuo, E., Wieman, C. E. (2016). "Toward instructional design principles: Inducing Faraday's law with contrasting cases." *Physical Review Physics Education Research* 12.1, pp. 1–14.

Kustov, A. A., Robinson, D. L. (1996). "Shared neural control of attentional shifts and eye movements." *Nature* 384.6604, pp. 74–77.

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., Lee, M.-H., Chiou, G.-L., Liang, J.-C., Tsai, C.-C. (2013). "A review of using eye-tracking technology in exploring learning from 2000 to 2012." *Educational Research Review* 10, pp. 90–115.

Landis, J. R., Koch, G. G. (1977). "The measurement of observer agreement for categorical data." *Biometrics* 33, pp. 159–174.

Leinhardt, G., Zaslavsky, O., Stein, M. K. (1990). "Functions, graphs, and graphing: Tasks, learning, and teaching." *Review of Educational Research* 60.1, pp. 1–64.

Liang, K.-Y., Zeger, S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika* 73.1, pp. 13–22.

Lohman, D. F., Dennis, I., Tapsfield, P. (1996). *Human abilities: Their nature and measurement.*

Madsen, A. M., Larson, A. M., Loschky, L. C., Rebello, N. S. (2012). "Differences in visual attention between those who correctly and incorrectly answer physics problems." *Physical Review Special Topics-Physics Education Research* 8.1, pp. 1–13.

Mallery, P., George, D. (2000). *SPSS for windows step by step.* Allyn and Bacon, Inc.

Malone, S., Altmeyer, K., Vogel, M., Brünken, R. (2020). "Homogeneous and heterogeneous multiple representations in equation-solving problems: An eye-tracking study." *Journal of Computer Assisted Learning* 36.6, pp. 781–798.

Maslowska, E., Segijn, C. M., Vakeel, K. A., Viswanathan, V. (2020). "How consumers attend to online reviews: an eye-tracking and network analysis approach." *International Journal of Advertising* 39.2, pp. 282–306.

Mautone, P. D., Mayer, R. E. (2001). "Signaling as a cognitive guide in multimedia learning." *Journal of educational Psychology* 93.2, p. 377.

Mayer, R. E. (2009). *Multimedia learning.* 2nd ed. New York: Cambridge University Press.

Mayer, R. E. (2005). *The cambridge handbook of multimedia learning.* Cambridge university press.

Mayer, R. E. (2008). "Applying the science of learning: evidence-based principles for the design of multimedia instruction." *American Psychologist* 63.8, pp. 760–769.

Mayring, P. (2014). "Qualitative content analysis: theoretical foundation, basic procedures and software solution."

Mayring, P. (2019). "Qualitative content analysis: Demarcation, varieties, developments." *Forum: Qualitative Social Research.* Vol. 20. 3. Freie Universität Berlin.

McDermott, L. C., Rosenquist, M. L., Van Zee, E. H. (1987). "Student difficulties in connecting graphs and physics: Examples from kinematics." *American Journal of Physics* 55.6, pp. 503–513.

Merkley, R., Ansari, D. (2010). "Using eye tracking to study numerical cognition: The case of the ratio effect." *Experimental Brain Research* 206, pp. 455–460.

Mock, J., Huber, S., Klein, E., Moeller, K. (2016). "Insights into numerical cognition: Considering eye-fixations in number processing and arithmetic." *Psychological Research* 80.3, pp. 334–359.

Neuendorf, K. (2002). "The content analysis guidebook sage publications, inc." *Library of Congress. CA: United States.*

Newmann, M. (2010). "Networks: an introduction." *Oxford University Press, ed* 1, p. 720.

Nieminen, P., Savinainen, A., Viiri, J. (2010). "Force concept inventory-based multiple-choice test for investigating students' representational consistency." *Physical Review Special Topics-Physics Education Research* 6.2, pp. 1–12.

OECD (2019). *Pisa 2018 results (volume i)*, p. 354. DOI: https://doi.org/10.1787/5f07c754-en. URL: https://www.oecd.org/en/publications/pisa-2018-results-volume-i_5f07c754-en.html.

OECD (2023). *Pisa 2022 results (volume i)*, p. 491. DOI: https://doi.org/https://doi.org/10.1787/53f23881-en. URL: https://www.oecd-ilibrary.org/content/publication/53f23881-en.

Olsen, A., Smolentzov, L., Strandvall, T. (2010). "Comparing different eye tracking cues when using theretrospective think aloud method in usability testing." *Proceedings of HCI 2010* 24, pp. 45–53.

Paivio, A. (1990). *Mental representations: A dual coding approach.* Oxford university press.

Pan, W. (2001). "Akaike's information criterion in generalized estimating equations." *Biometrics* 57.1, pp. 120–125.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). "Scikit-learn: Machine learning in python." *the Journal of Machine Learning Research* 12, pp. 2825–2830.

Peter, X.-K. S., Song, K. (2007). *Correlated data analysis: modeling, analytics, and applications*. Springer.

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, pp. 1393–1414.

Pollock, E. B., Thompson, J. R., Mountcastle, D. B. (2007). "Aip conference proceedings." Vol. 951. 1. American Institute of Physics, pp. 168–171.

Potgieter, P., Blignaut, P. (2017). "Using eye-tracking to assess the application of divisibility rules when dividing a multi-digit dividend by a single digit divisor." *Proceedings of the South African Institute of Computer Scientists and Information Technologists*, pp. 1–10.

Prokop, M., Pilař, L., Tichá, I. (2020). "Impact of think-aloud on eye-tracking: A comparison of concurrent and retrospective think-aloud for research on decision-making in the game environment." *Sensors* 20.10, pp. 2750–2771.

Radford, L. (2002). "The seen, the spoken and the written: A semiotic approach to the problem of objectification of mathematical knowledge." *For the Learning of Mathematics* 22.2, pp. 14–23.

Redish, E. F., Kuo, E. (2015). "Language of physics, language of math: Disciplinary culture and dynamic epistemology." *Science and Education* 24.5-6, pp. 561–590.

Redish, E. F., Smith, K. A. (2008). "Looking beyond content: Skill development for engineers." *Journal of Engineering Education* 97.3, pp. 295–307.

Rello, L., Bautista, S., Baeza-Yates, R., Gervás, P., Hervás, R., Saggion, H. (2013). "One half or 50%? An eye-tracking study of number representation readability." *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part IV 14*. Springer, pp. 229–245.

Salvucci, D. D., Anderson, J. R. (2001). "Automated eye-movement protocol analysis." *Human-Computer Interaction* 16.1, pp. 39–86.

Salvucci, D. D., Goldberg, J. H. (2000). "Identifying fixations and saccades in eye-tracking protocols." *Proceedings of the 2000 Symposium on Eye tracking Research and Applications*, pp. 71–78.

Scheiter, K., Schubert, C., Schüler, A., Schmidt, H., Zimmermann, G., Wassermann, B., Krebs, M.-C., Eder, T. (2019). “Adaptive multimedia: Using gaze-contingent instructional guidance to provide personalized processing support.” *Computers and Education* 139.3, pp. 31–47.

Schindler, M., Lilienthal, A. J. (2019). “Domain-specific interpretation of eye tracking data: towards a refined use of the eye-mind hypothesis for the field of geometry.” *Educational Studies in Mathematics* 101.1, pp. 123–139.

Schneider, B., Pea, R. (2014). “Toward collaboration sensing.” *International Journal of Computer-Supported Collaborative Learning* 9.4, pp. 371–395.

Schneider, E., Maruyama, M., Dehaene, S., Sigman, M. (2012). “Eye gaze reveals a fast, parallel extraction of the syntax of arithmetic formulas.” *Cognition* 125.3, pp. 475–490.

Scholtes, I., Wider, N., Garas, A. (2016). “Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities.” *The European Physical Journal B* 89.61, pp. 1–15.

Schreier, M. (2012). “Qualitative content analysis in practice.” *Qualitative content analysis in practice*, pp. 1–280.

Schüler, A. (2017). “Investigating gaze behavior during processing of inconsistent text-picture information: Evidence for text-picture integration.” *Learning and Instruction* 49, pp. 218–231.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., Chin, D. B. (2011). “Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer.” *Journal of Educational Psychology* 103.4, pp. 759–776.

Schwarz, J. (2023). *Einfaktorielle Varianzanalyse (ohne Messwiederholung).*

Sfard, A. (1991). “On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin.” *Educational Studies in Mathematics* 22.1, pp. 1–36.

Strohmaier, A. R., MacKay, K. J., Obersteiner, A., Reiss, K. M. (2020). “Eye-tracking methodology in mathematics education research: A systematic literature review.” *Educational Studies in Mathematics* 104.2, pp. 147–200.

Suppes, P., Cohen, M., Laddaga, R., Anliker, J., Floyd, R. (1983). "A procedural theory of eye movements in doing arithmetic." *Journal of Mathematical Psychology* 27.4, pp. 341–369.

Susac, A., Bubic, A., Kazotti, E., Planinic, M., Palmovic, M. (2018). "Student understanding of graph slope and area under a graph: A comparison of physics and nonphysics students." *Physical Review Physics Education Research* 14.2, p. 020109.

Susac, A. N., Bubic, A., Kaponja, J., Planinic, M., Palmovic, M. (2014). "Eye movements reveal student's strategies in simple equation solving." *International Journal of Science and Mathematics Education* 12.3, pp. 555–577.

Titscher, S., Jenner, B. (2000). *Methods of text and discourse analysis: In search of meaning*. Sage.

Tobii AB (2017). *Tobii studio user's manual*. English. Version 3.4.8. tobii pro. 172 pp.

Tsai, M.-J., Hou, H.-T., Lai, M.-L., Liu, W.-Y., Yang, F.-Y. (2012). "Visual attention for solving multiple-choice science problem: An eye-tracking analysis." *Computers and Education* 58.1, pp. 375–385.

Uhden, O., Karam, R., Pietrocola, M., Pospiech, G. (2012). "Modelling mathematical reasoning in physics education." *Science and Education* 21, pp. 485–506.

Uygun, T., Guner, P., Simsek, I. (2022). "Examining students' geometrical misconceptions by eye tracking." *International Journal of Mathematical Education in Science and Technology*, pp. 1–26.

Van den Eynde, S., Kampen, P. van, Van Dooren, W., De Cock, M. (2019). "Translating between graphs and equations: The influence of context, direction of translation, and function type." *Physical Review Physics Education Research* 15.2, pp. 1–13.

Van der Weijden, F. A., Kamphorst, E., Willemsen, R. H., Kroesbergen, E. H., Hoogmoed, A. H. van (2018). "Strategy use on bounded and unbounded number lines in typically developing adults and adults with dyscalculia: An eye-tracking study." *Journal of Numerical Cognition* 4.2, pp. 337–359.

Van Gog, T., Paas, F., Van Merriënboer, J. J., Witte, P. (2005). "Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting." *Journal of Experimental Psychology: Applied* 11.4, pp. 237–244.

Vekiri, I. (2002). "What is the value of graphical displays in learning?" *Educational Psychology Review* 14, pp. 261–312.

Was, C., Sansosti, F., Morris, B. (2016). *Eye-tracking technology applications in educational research.* IGI Global.

Witte, K., Witte, K. (2019). "Statistische Verfahren zur Überprüfung von Unterschiedshypothesen bei zwei Stichproben." *Angewandte Statistik in der Bewegungswissenschaft (Band 3)*, pp. 63–78.

Zhu, M., Feng, G. (2015). "An exploratory study using social network analysis to model eye movements in mathematics problem solving." *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pp. 383–387.

Zimmerman, D. W. (1987). "Comparative power of student t test and mann-whitney u test for unequal sample sizes and variances." *The Journal of Experimental Education* 55.3, pp. 171–174.

Curriculum Vitae

Work Experience

since 03/2020 **Research assistant**
at the Department of Mathematics, RPTU Kaiserslautern-Landau, Location Kaiserslautern.

04/2019–09/2019 **Student assistant**
at the Department of Physics, RPTU Kaiserslautern-Landau, Location Kaiserslautern.

10/2015–03/2016, 10/2018–03/2019 **Student assistant**
at the Department of Mathematics, RPTU Kaiserslautern-Landau, Location Kaiserslautern.

Education

since 04/2020 **PhD**
University of Kaiserslautern-Landau, Location Kaiserslautern. Title: Linear Functions – a Matter of Context? Toward Understanding Students' Learning Processes in Mathematics and Physics. Department: Mathematics. Supervisor: Prof. Dr. Stefan Ruzika.

10/2018–03/2020 **Master of Education**
Teacher training program for secondary school teachers, University of Kaiserslautern-Landau, Location Kaiserslautern. Subjects: Mathematics, Physics. Title of the Master's Thesis: Design and construction of optical tweezers for nanodiamonds and glass spheres. Supervisor: Prof. Dr. Artur Widera.

04/2017–09/2018 **Bachelor of Education**
Teacher training program for secondary school teachers, University of Kaiserslautern-Landau, Location Kaiserslautern. Subjects: Mathematics, Physics.

10/2013–09/2017 **Bachelor of Science**
Bachelor's degree in Mathematics, RPTU Kaiserslautern-Landau, Location Kaiserslautern. Title of the Bachelor Thesis: The Riemann-Hurwitz Formula. Supervisor: Prof. Dr. Andreas Gathmann.

10/2011–03/2012	**Early studies in mathematics** Linear Algebra, Trier University, Trier
2004–2013	**Graduation (Abitur)** Episcopal Angela-Merici-Grammar school, Trier.

Wissenschaftlicher und beruflicher Werdegang

Berufserfahrung

seit 03/2020	**wissenschaftliche Mitarbeiterin** am Fachbereich Mathematik, RPTU Kaiserslautern-Landau, Standort Kaiserslautern.
04/2019–09/2019	**wissenschaftliche Hilfskraft** am Fachbereich Physik, at the Department of Physics, RPTU Kaiserslautern-Landau, Standort Kaiserslautern.
10/2015–03/2016 10/2018–03/2019	**wissenschaftliche Hilfskraft** am Fachbereich Mathematik, RPTU Kaiserslautern-Landau, Standort Kaiserslautern.

Ausbildung

seit 04/2020	**Promotion** RPTU Kaiserslautern-Landau, Standort Kaiserslautern. Titel: Lineare Funktionen – eine Frage des Kontexts? Zum Verständnis der Lernprozesse von Schüler:innen in den Fächern Mathematik und Physik. Fachbereich: Mathematik. Doktorvater: Prof. Dr. Stefan Ruzika.
10/2018–03/2020	**Master of Education** Lehramtsstudium (LAG), RPTU Kaiserslautern-Landau, Standort Kaiserslautern, Fächer: Mathematik, Physik. Titel der Masterarbeit: Konstruktion und Aufbau einer optischen Pinzette für Nanodiamanten und Glaskugeln. Betreuer: Prof. Dr. Artur Widera.
04/2017–09/2018	**Bachelor of Education** Lehramtsstudium (LAG), RPTU Kaiserslautern-Landau, Standort Kaiserslautern. Fächer: Mathematik, Physik.
10/2013–09/2017	**Bachelor of Science** Bachelorstudium Mathematik, RPTU Kaiserslautern-Landau, Standort Kaiserslautern. Titel der Bachelorarbeit: Die Riemann-Hurwitz-Formel. Betreuer: Prof. Dr. Andreas Gathmann

10/2011–03/2012	**Frühstudium** Linear Algebra, Universität Trier, Trier.
2004 – 2013	**Allgemeine Hochschulreife** Bischöfl. Angela-Merici-Gymnasium, Trier.