A Digital Approach to Talent Development With Mathematical Proof

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Ed. Xenia Rea Viviane Stein

aus Göppingen

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	19.12.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Walther Paravicini
2. Berichterstatter/-in:	Prof. Dr. Jan-Philipp Burde

ACKNOWLEDGMENTS

On my journey from a blank page—figuratively representing an empty TeX file—several special people played significant roles. First, I would like to express my gratitude to my supervisors, Prof. Dr. Ulrich Trautwein and Prof. Dr. Walther Paravicini, for the opportunity to pursue a PhD in my two main passions: Giftedness and Mathematics. I also thank Prof. Dr. Jan-Philipp Burde for acting as a reviewer for my dissertation.

In particular, I would like to extend my gratitude to Prof. Dr. Jessika Golle and Prof. Dr. Katerina Tsarava, who supervised me from the very first day of my Master's thesis and continued to do so until the end of my PhD phase. Thank you for being such shining examples in challenging the 'glass ceiling'! - $\epsilon\upsilon\chi\alpha\rho\iota\sigma\tau\acute{\omega}$!

None of this would have been possible without my fellow PhD students: you made this time feel like a wonderful class trip! In both realms, the 'AG Mathematik und ihre Didaktik' and the HIB. Special thanks to David Weiler for co-working with me even across federal states. I am especially grateful to all (honorary) members of MET-Office 3, particularly Markus Kleinhansl, Fabienne Kremer, Dr. Lucas Stark, and my best-ever PhD buddy, Katrin Kunz!

I would also like to extend my gratitude to the HUB talent development, including the excellent HCAP coordination team led by Kristin Funcke, who consistently adds a spark of color to the otherwise grey academic structures. I am also thankful to Dr. Nina Udvardi-Lakos for her reel(!) support and for all the unconventional ideas we developed and implemented together, and to Dr. Armin Fabian for demonstrating to me how a successful interdisciplinary PhD journey can work out.

Furthermore, I wish to thank Dr. Benjamin Goecke for his expertise in test development, Dr. Wolfgang Wagner for his invaluable statistical consultations, and Alla Kutkina, Dr. Björn Rudzewitz, and Dr. Merlin Carl for talking to computers for me and solving every problem – even if located on the eighth layer.

Lastly, I want to express my heartfelt gratitude to those outside academia whose support has led me to this point: to my parents, for introducing me to the concepts of giftedness and enrichment from an early age and for always encouraging my ambitions; to my best friend Jan, for countless lunch breaks and endless conversations about anything and everything; and to my husband Jakob, for constantly reminding me of my own abilities and the humorous side of life.

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Explanation
BL	Boolean Logic
CFI	Comparative Fit Index
CT	Computational Thinking
e.g.,	for example
ENT	Elementary Number Theory
HCAP	Hector Children's Academy Program
H5P	HTML5 Package (interactive content creation tool)
i.e.,	that is
LD	Logical Detectives (name of intervention course)
M	Mean
n	Sample Size
OECD	Organisation for Economic Co-operation and Development
PC	Proof Competency
PfPT	Preformal Proving Test
RCT	Randomized Controlled Trial
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation
SRMR	Standardized Root Mean Square Residual
ST	Set Theory
α	Reliability coefficient alpha
ω	Reliability coefficient omega
χ^2	Chi-squared statistic

ABSTRACT

In the field of mathematics, proving is considered both the most crucial and the most challenging skill to acquire. The latter aspect is a likely reason why this topic is not emphasized in the (in this case, German) school curriculum. However, when it comes to fostering gifted children, extracurricular and challenging content is needed. Proving shows substantial overlap with the disciplines of reasoning and problem solving, which have often been successfully incorporated into fostering programs. And while some educators refrain from teaching proof to schoolchildren, even those who are gifted, others demand the integration of the topic into regular lessons for all age groups. This dissertation examines how gifted primary school children can be supported through an enrichment course designed to teach them the construction of mathematical proofs. The course was held as an asynchronous online course. Possible challenges associated with the format and topic were investigated in a pilot study of the course. The children participating in the study reported that they liked most parts of the course and were eager to engage with the content. They also identified parts that were too challenging from cognitive or technical perspectives. Following the study, a new test instrument was developed and validated to measure Proof Competency (PC) in children. This instrument was then used in an efficacy study of the revised course alongside motivational measures. This study did not find that the course had significant effects on PC, possibly due to the high number of non-finishers in the treatment group, which is typical for asynchronous online courses. Furthermore, the exceptionally high self-concept the children had before the course significantly declined. These findings led to the idea of a follow-up study to investigate how situational motivation evolves throughout the course. The outcomes revealed specific aspects of the course where additional scaffolding or facilitation could help prevent students from dropping out. Furthermore, embedding self-regulation training could be identified as a potential way to further enhance successful course participation.

ZUSAMMENFASSUNG

In der universitären Mathematik ist das Beweisen nicht nur die häufigste Tätigkeit, sondern wird auch als große Herausforderung für Studierende betrachtet. Aufgrund dieses Rufs findet das formale algebraische Beweisen in vielen Ländern – so auch in Deutschland – kaum Einzug in die schulischen Mathematik-Curricula. Für die Förderung hochbegabter Lernender sind jedoch genau diese außerschulischen und herausfordernden Inhalte bedeutsam. Auch mit den Tätigkeiten des Problemlösens und Argumentierens, die in der Fachliteratur immer wieder als Förderinhalte für diese Gruppe gelobt werden, hat das Beweisen große Gemeinsamkeiten. Es gibt zwar nach wie vor Stimmen, die das Beweisen als zu schwer für Schulkinder unabhängig von einer vorhandenen Begabung abtun, doch gleichzeitig wird eine Beweislehre schon im Grundschulalter von vielen Forschenden gefordert und von Lehrenden mitunter bereits erfolgreich durchgeführt. Die vorliegende Dissertation untersucht, inwieweit begabte Grundschulkinder mit einem asynchronen Onlinekurs zum mathematischen Beweisen gefördert werden können. Dieser Kurs wurde zunächst in einer Pilotstudie testweise durchgeführt, um Anpassungsbedarfe für die Zielgruppe zu identifizieren. Die Lernenden berichteten, dass sie an den meisten Kurselementen viel Spaß und gute Lernerlebnisse hatten, und bewerteten die Lernumgebung weitgehend positiv. Einzelne Inhalte, die inhaltlich oder technisch herausfordernd waren, wurden benannt und konnten in der Überarbeitung des Kurses adressiert werden. Um den Kurserfolg zu messen, wurde in zwei weiteren Studien ein Testinstrument für die präformale Beweiskompetenz von Grundschulkindern entwickelt und validiert. Dieses wurde in einer Wirksamkeitsstudie des überarbeiteten Kurses eingesetzt. Zusätzlich wurden in dieser Studie fachspezifische motivationale Effekte untersucht. Der Kurs zeigte keine Effekte auf die Beweiskompetenz der Lernenden, was mutmaßlich mit der mangelnden Interventionstreue zusammenhängt. Diese kommt durch die für asynchrone Kurse typische Rate an Lernenden zustande, die nicht den ganzen Kurs absolvierten. Auch zeigte sich eine signifikante Abnahme des zu Beginn häufig sehr hohen Beweis-Selbstkonzepts. Um diese Mechanismen genauer zu studieren, wurde in einer weiteren Studie die situationale Motivation während des Kurses untersucht. Dadurch ergaben sich Anhaltspunkte für die weitere Anpassung der Inhalte und des Formats an die Zielgruppe. Ausblickend soll auch die Möglichkeit erwähnt werden, durch ein eingebettetes Selbstregulationstraining die anhaltende Kursnutzung zu stützen.

1 Introduction

"convince yourself
convince a friend
convince an enemy"
(Mason, Burton, & Stacey, 1982, p.109)

This analogy illustrates vividly how a mathematical proof needs to evolve from a mere idea into a substantial assertion, and ultimately into a sound, formal proof. The product of this process, the mathematical proof, is what various scholars described as the most crucial aspect of mathematics: Proof has been characterized as a cornerstone (Dawkins & Weber, 2017), the heart (Bass, 2011), the soul (Schoenfeld, 2009), or the essence (Ross, 1998) of mathematics. Rav (1999) goes as far as to say that proofs, rather than the theorems they prove, are what constitute mathematics. Less poetically speaking, mathematical proof can be defined as the following: "A proof of a mathematical theorem is a sequence of steps that leads to the desired conclusion. The rules to be followed by such a sequence of steps were made explicit when logic was formalized early in this [the 20th] century, and they have not changed since"(Rota, 1997, p.34).

But proof is not only relevant for theoretical mathematicians to verify their claims: Proof has several other roles, namely explanation, systematization, discovery, intellectual challenge, and communication (Villiers, 2012). The role of explanation is outstanding, because no matter how confident one is about the truth of a claim, supported by many examples – only a proof can show *why* it is true (Villiers, 2012). Grabiner (2012) states that proof not only enables students to think logically in the classroom but also helps them become responsible citizens. She refers to several historical and real-world examples to illustrate the intersections of proof and society. For example, the proof-wording was used in the Declaration of Independence to ensure its logical flawlessness. Therefore, proof can be regarded as one of the skills included in mathematical literacy, which the OECD defines as "... an individual's capacity to reason mathematically and to formulate, employ, and interpret mathematics to solve problems in a variety of real-world contexts. It includes concepts, procedures, facts and tools to describe, explain and predict phenomena"(OECD, 2025).

Nevertheless, in many countries, proof is not addressed in school, at least not before secondary school (e.g., Ministerium für Kultus, 2016a). In grammar schools, proof is a topic, but only geometrical proofs are featured and the transition to formal proof is often left out (e.g., Ministerium für Kultus, 2016b; National Council of Teachers of Mathematics, 2000). At university, proving is a skill necessary for most STEM subjects. But students often lack proof skills and require transition-to-proof

courses to make up for this (e.g., Glosauer, 2019; Grieser, 2017). Therefore, some educators argue for an earlier integration of formal proof. Some already demand and implement this in primary school (A. J. Stylianides, 2016; Zaslavsky, Nickerson, Stylianides, Kidron, & Winicki–Landman, 2011). Others reject these demands, arguing that formal proof is too challenging and not child-appropriate (Dreyfus, 1999; Kline, 1973; Sommerhoff & Ufer, 2019).

For mathematically gifted children, however, challenging content is necessary to ensure they do not lose their interest in mathematics (Rotigel & Fello, 2016). For this group specifically, learning how to prove can be especially worthwhile, as they tend to already have a formalized perception and a desire to generalize mathematical propositions (Krutetskii, 1976). Furthermore, proof activities are among the generally recommended activities for mathematically gifted children, similar to problem-solving and reasoning (Bardy & Bardy, 2020). Proving shares strong connections with these two prominent enrichment fields (G. Stylianides & Silver, 2007; Tall, 2009).

While several enrichment programs have successfully taught mathematical proof to gifted primary school children in Korea (Chang, Chong, & Song, 2006; Ko & Song, 2011; K. H. Lee, 2005; Na, 2011), to the best of my knowledge, there have been no such approaches targeting gifted primary school children in Germany. This is why this dissertation aims to explore the possibilities of incorporating mathematical proving into an enrichment program for German primary school children.

To do so, an intervention was developed in line with Nelson, Cordray, Hulleman, Darrow, and Sommer (2012), relying on – in this case: four – core components. Such core components are basic ideas from which the activities for an intervention are then derived. Two of these address the topic of proving, namely *Natural language proof writing* and *Iconic and symbolic logical reasoning*, and were derived from transition-to-proof courses and university interventions for proof novices (e.g., Carl, Lorenzen, & Schmitz, 2022; Glosauer, 2019; Grieser, 2017).

The other two core components address the learning format. To cope with the problems that on-site enrichment courses can suffer from, such as limited teacher attention per student (Diezmann & Watters, 2000) or inequality in nomination for limited course seats (Golle, Schils, Borghans, & Rose, 2022; McBee, Peters, & Waterman, 2014), the intervention is set up as an asynchronous online course with the core components of *Self-paced learning* and *Automated real-time feedback*. To enhance the possibilities for practice, Carl (2022) successfully implemented automated real-time feedback in university courses for proof novices. In other contexts, this method has shown potential to improve practice motivation, self-reflection, self-concept and performance (J. Schneider, Börner, van Rosmalen, & Specht, 2016). This kind of practice can easily be embedded into a self-paced learning environment, which is also beneficial for challenging tasks (Tullis & Benjamin, 2011). As proving is generally perceived as challenging (Bass, 2011), this can be a suitable setting for the

intervention.

Overall, in this dissertation, I strive to determine how enjoyable and effective an asynchronous proof course can be for talented primary school children. To do so, an online intervention was designed to foster the Proof Competency (PC) of talented primary school children, and the following research goals were set to guide the inquiry: (1) Investigate the feasibility of the intervention. (2) Develop a sound model of PC, sample test items, and investigate the feasibility of the test design and the nomological network of PC. (3) Study the achievement-related and motivational effects of the intervention. (4) Investigate the trajectories of dispositional and situational motivation during the intervention and their connection to parental support and dropout behavior.

The primary goal of this dissertation is to determine if and how PC can be fostered in an asynchronous enrichment course. Therefore, I will first focus on conceptualizations of mathematical proof, related competencies, and approaches to teaching and assessing PC. After that, I will consider the target group of mathematically gifted children and discuss the overlap between their fostering needs and the content of mathematical proving. Subsequently, I will lay out how the format of asynchronous online courses suits the demands for such a fostering program, before moving on to the research questions of the dissertation. In Section 3, I will describe the development of the asynchronous online course *Logical Detectives* that teaches mathematical proving to gifted primary school children. In Study 1, this course was piloted in a broad online study with talented primary school children. The study identified which course elements were already suitable for the target group and which needed further refinement. In Study 2, a new measurement instrument is presented that can be used with primary school children who have no prior knowledge of formal mathematics. Results from a pilot and a validation study are reported. In Study 3, the revised course was implemented as an intervention in a randomized controlled trial with a waiting group. To measure possible effects on PC, the test instrument from Study 2 was utilized. Furthermore, motivational effects were examined with self-reports. Finally, Study 4 evaluates the motivational trajectories during the asynchronous online course. As a final point, I will present the overarching results of the dissertation and discuss their relevance to research and practice.

2 Theoretical Framework

2.1 Theoretical Conceptualizations of Proof and Proving

2.1.1 Definitions of Mathematical Proof

In mathematical research, there are different schools of thought, each defining differently what a proof is and which ways of proving there are (Rota, 1997; Wilder, 1944). The school of formalists, known for their logical rigor, defines proof as

> ... a sequence of assertions, the last of which is the theorem that is proved and each of which is either an axiom or the result of applying a rule of inference to previous formulas in the sequence; the rules of inference are so evident that the verification of the proof can be done by means of a mechanical procedure. (Tall et al., 2011)

When this definition was given to a mathematics student, they would probably recognize some of the underlying thoughts and agree that it describes what they learned to be a proof. However, when asked to define proof, the answer would most likely be shaped in a more practical sense, like the one that Rota gives as an opening in his work about mathematical proof:

> Everybody knows what a mathematical proof is. A proof of a mathematical theorem is a sequence of steps which leads to the desired conclusion. The rules to be followed by such a sequence of steps were made explicit when logic was formalized early in this century, and they have not changed since. (Rota, 1997, p.34)

Indeed, when talking about proof in mathematics, formal approaches are often distinguished from more applied ones. In his book *Proofs and Contradictions*, Lakatos (2015) distinguished between *substantive mathematics*, which deals with actual objects and their representations and argues within the boundaries of these, and *formal mathematics*, which is disconnected from any real object, with propositions and conclusions solely serving as abstract ideas.

Similarly, proofs can be categorized into formal and practical proofs: Purely formal proofs match the proof definition by Hilbert provided above. They are unbroken chains of mathematical arguments, each represented in formal notation and deduced by strict application of the rules of inference, only adding presumed axioms (Snapper, 1979). Such proofs are written to verify theorems universally and demonstrate that a certain mathematical field is free from internal contradictions (Villiers, 2012). However, even professional mathematicians do not work on that level of formalization: As this formal approach would forbid the use of lemmata and propositions that others previously provided proofs for, all advanced proofs would be too lengthy for humans to engage with or even

read (Cadwallader Olsker, 2011). A *practical* mathematical proof, on the other hand, is informal, imprecise, and subjective, but serves the aim of being understood by the mathematical community and convincing them that a certain proposition holds (Cadwallader Olsker, 2011; Hersh, 1997).

Villiers (2012) differentiated between several distinct roles of proof, with different importance depending on the context: Verification – which is crucial for mathematical research, systematization, discovery, intellectual challenge, communication, and explanation. The latter role is the most important one within the school context, as a proof is the only way to see *why* a claim holds (Villiers, 2012). This leads us from the duality of formal and practical proof to the contrastive pair of proof in research and proof in school. Just as there are different definitions necessary to distinguish between formal and practical proofs, considering the various purposes of a proof, the mathematical classroom requires its own definition of proof.

> Proof is a mathematical argument ... with the following characteristics:
>
> 1. It uses statements accepted by the classroom community (set of accepted statements) that are true and available without further justification;
> 2. It employs forms of reasoning (modes of argumentation) that are valid and known to, or within the conceptual reach of, the classroom community;
> 3. It is communicated with forms of expression (modes of argument representation) that are appropriate and known to, or within the conceptual reach of, the classroom community. (A. Stylianides, 2007, p.291)

Thus, in the school context, very different chains of arguments can be considered a proof. In their work on the didactics of mathematics, Wittmann and Müller (1988) distinguish between different kinds of proofs, adequate for certain scenarios: 1. Formal-deductive proofs consist of logical argumentative chains expressed through formal symbols that hold under professional mathematical standards. 2. Substantive-illustrative proofs also use general mathematical operations, but merely those that can be intuitively recognized as universally applicable, allowing readers to acknowledge the generality of the claim immediately. 3. Experimental proofs demonstrate an argument through specific, but not universal, examples and are therefore not as rigorous but particularly suitable for proof novices.

In this work, I will refer to proofs as formal-deductive proofs. As the studies in this dissertation target gifted primary school children, I will use the proof definition by A. Stylianides (2007) as a basis. Nevertheless, the enrichment course described in 3.3 aims to expand the knowledge and skills of this specific *classroom community*, enabling the children to deal with proofs that also meet the more formal definition by Rota (1997).

I want to close this section with a quote by Ruben Hersh that illustrates the parallels between school and research, highlighting what we tend to forget about proof in both contexts and motivating this dissertation: "Proof is a tool in service of teacher and class, not a shackle to restrain them. In teaching future mathematicians, Proof is a tool in service of research, not a shackle on the mathematicians imagination"(Hersh, 1997, p.60).

2.1.2 Competences Associated with Mathematical Proof

After the previous clarification of the term *proof*, I will now turn to the act of proving. There are two core competences necessary for engaging with mathematical proofs: proof comprehension, which is the ability to make sense of a given proof, and proof construction, which refers to the ability to create proofs oneself – generally known as *proving* (Waluyo, Vidákovich, Ishartono, & Toyib, 2019). In this section, I will discuss models for both competences, and subsequently report on two cognitive skills, namely reasoning and problem solving, which are closely related to proving.

Proof Comprehension The process of proof comprehension according to the model by Yang and Lin (2008) consists of the following steps:

1. Knowing the meaning of terms, symbols or figures in this proof.
2. Identifying statements of this proof as definitions, premises, applied properties or conclusions.
3. Checking each step according to the deductive rules.
4. Identifying statements of this question as conditions or claimed conclusions.
5. Finding the critical ideas of this proof.
6. Judging the validation of this proof.
7. Judging the truth of this proposition.
8. Judging the generality of this proposition.
9. Finding conflict between this proof and one's anticipated proof.
10. Revising this proposition.
11. Giving plausible explanation for some statements. [p.61]

The model illustrates how proof comprehension begins on a symbolic level and goes via logical deduction to the investigation of generality in the mathematical world. Thus, one can assume that the necessary abilities to comprehend proofs include mathematical knowledge on a formal (Step 1) and content level (Step 2), as well as reasoning abilities (Steps 3,4,6-8).

Proof Construction For proof construction, Boero (1999) suggests six steps of writing a mathematical proof:

1. Finding a conjecture within a mathematical problem area,
2. Formulation of the conjecture according to customary standards,
3. Exploration of conjecture with the limits of its truth; ... Identifying appropriate arguments in support of the conjecture
4. Selection of arguments that can be organized into a proof chain in a deductive chain,
5. Fixation of the chain of argumentation according to current mathematical standards
6. Formal proof. (par. 4)

When deriving competencies from these steps, proof construction requires content knowledge (especially Steps 1 and 3), argumentation skills (Steps 3, 4, and 5), and formal mathematical knowledge (Step 6). In contrast to proof construction, the mathematical content is investigated in the first place, followed by steps of reasoning and argumentation, ending with the formalization of the proof. This progression is antiparallel to the progression in proof comprehension that starts on the symbolic level.

For the target group of gifted primary school students, I decided to address only proof construction. Here, the children can apply their mathematical knowledge and reasoning skills and then proceed with learning the formal symbols and formalizing their thoughts. Thus, I will from here on use the term *proving* for proof construction, the main competence investigated in the studies included in this work.

Reasoning Reasoning can be defined as the process of evaluating and combining knowledge into a conclusion (Stenning & van Lambalgen, 2010). Considering the model of proving presented in the previous section, it becomes clear that proving requires exactly this competence. Also, reasoning is relevant to the classroom definition of proof cited above (A. Stylianides, 2007).

In the didactical literature, this relationship is well-known and has been broadly investigated (Bleiler, 2009; G. Stylianides & Silver, 2007; Thompson, Senk, & Johnson, 2012). G. Stylianides (2008) states that reasoning is a prerequisite for mathematical proving. He has established the term *reasoning-and-proving* for school activities that encourage children to justify their answers and thereby support the transition from reasoning to proving. Gutierrez and Jaime (1998) have worked out an assessment of the van Hiele levels of reasoning, a prominent model for rating how elaborately one can argue in a geometrical context. They have defined four processes of reasoning,

corresponding to the van Hiele levels: Recognition, Definition, Classification, and Proof. Thus, proving can be seen as a very advanced way of reasoning. Nevertheless, it is important to stress that even though proving is related to reasoning, not every form of reasoning can be considered a proof (Selden, 2013). Also, in several mathematics curricula, reasoning and proving are mentioned as one common field for learning goals (Ministerium für Kultus, 2016b; National Council of Teachers of Mathematics, 2000).

Therefore, it can be assumed that learners doing well in reasoning tasks have better prerequisites for proving. This relation will be especially relevant to Section 2.2.3, in which the suitability of proving for mathematically gifted children is shown. Furthermore, in Study 2 (see: Section 5), psychological measures of reasoning will be applied to investigate the convergent validity of a proof test.

Problem Solving In a similar way, a connection between proving and problem solving was drawn by several scholars:

Firstly, the work of George Polya highly influenced both the disciplines of proving and problem solving (Hanna & Knipping, 2020). In his four-step model of problem solving (understanding the problem, devising a plan, carrying out the plan, looking back), the last step demands a proof of the solution one has devised for the problem (Polya, 2014). The full four steps can be seen as a structure very similar to the process of proving (Hanna & Knipping, 2020).

Secondly, Weber (2005) regards the perspective of proof construction as a problem solving task. He refers to Schoenfeld's (1985) definition of a mathematical problem as a task with no obvious solution or order of mathematical actions to be applied. This can be due to multiple ways of solving it or because the starting point is not a standard situation (Weber, 2005). Therefore, he concludes that proving is a problem solving task, as there are many different ways to construct a proof in most cases. Furthermore, he highlights that just like in problem solving, the heuristics and strategies that learners apply in a proof define which competences are trained.

Lastly, according to Selden (2013), proving consists of one problem solving part and one formal part. In the problem solving part, informal reasoning is included, while the formal part only relies on applying formal logical laws and symbols (Selden, 2013).

Thus, there is much theoretical evidence for the entanglement of problem solving and proving. I will revisit this relation in Section 2.2.3 and also in Study 2 (see: Section 5). There, I will consider correlations between PC and Computational Thinking (CT), as the latter is highly correlated to problem solving (Tsarava et al., 2019) and also applicable to different problem solving contexts (Román–González, Pérez–González, Moreno–León, & Robles, 2018).

2.1.3 Mathematical Proving in the German Educational System

In the 1960s, Western countries sought to address the lack of PC that university students displayed in all mathematical fields except geometry (Hanna & Knipping, 2020). Therefore, the New Mathematics movement introduced a novel curriculum to include formal notation and rigorous proving at school level already (Hanna & Knipping, 2020; Kline, 1973). This approach was soon criticized by teachers for being too formal and by parents for being too far away from what they learned at school, and thus not continued any further (Hanna & Knipping, 2020; Kline, 1973).

To describe the present situation, I will focus on the German state of Baden-Württemberg, as curricula differ between the federal states. Here, the mathematical curriculum is comprised of two dimensions: It features five overarching procedural competences, and for each school year, five guiding ideas with several sub-topics, i.e., mathematical content to which these competences are applied. In describing how proving is taught, I will go from primary school to higher education.

In the primary school curriculum, which ranges from year 1 to 4, the words 'proof' or 'proving' do not appear (Ministerium für Kultus, 2016a). Still, one of the procedural competences established in the curriculum is *argumentation*. The learning goals for this competency include, among others, justifying one's own solutions and striving to understand the reasons behind observed rules. The procedural competences are applied to all content that is listed in the curriculum.

In the curriculum for grammar school (Gymnasium), the procedural competence of *argumentation* is extended to *argumentation and proof*. It features three subordinate learning goals that explicitly contain the word proof: "10. Understanding and reproducing proofs, 11. Tracing the argumentation in mathematical proofs back to the underlying basis of reasoning ... 13. Checking statements for their validity and writing proofs"(Ministerium für Kultus, 2016b, p.12). Here, the duality of proofreading and proofwriting discussed in 2.1.2 can be observed again. Remarkably, learning goal 13, which is the only one explicitly concerned with proof writing, is only linked to one guiding idea, namely *geometry*. Thus, algebraic proving is not explicitly part of the curriculum. Furthermore, the curriculum does not specify any requirements regarding the formal level of the proofs as described in the model by Wittmann and Müller (1988) that was cited in Section 2.1.1.

The transition to proof can be particularly challenging for first year university students, many university courses demand rigorous algebraic proof writing, even though it is not taught in the final years of school (Glosauer, 2019; Kempen, 2019). Therefore, universities offer classes for new students to bridge the gap, and several books have been written to help individuals acquire this skill unsupervised (e.g., Glosauer, 2019; Grieser, 2017). Typically, these transition-to-proof courses cover the disciplines of Boolean Logic, Set Theory, and Elementary Number Theory, as these form

a foundation for writing proofs in calculus, algebra, and subsequent mathematical lectures (Carl, 2022). Boolean Logic (BL) is formal logic based on binary truth values (Stoll, 2012), Set Theory (ST) focuses on determining which mathematical elements belong to or are excluded from a set of mathematical objects (Stoll, 2012), and Elementary Number Theory (ENT) features universal conjectures involving natural numbers (e.g., divisibility, number parity) (Carl, 2022).

2.1.4 Approaches to Teaching Mathematical Proving

As discussed in Section 2.1.1, between the poles of a professional mathematical definition (Rota, 1997) and a classroom definition (A. Stylianides, 2007) there is much room: While the purpose of a proof in research is mostly to ensure the validity of a new statement, in school, the explanatory role of proof is central (Villiers, 2012). This also affects the ways in which proving is taught and shifts the focus from formalism to reasoning. Biehler and Kempen (2016) have developed a framework to classify approaches to teaching mathematical proof depending on the mode of didactical reduction. For their categorization, they use four criteria (Biehler & Kempen, 2016): 1. Mode of representation: Are formal symbols replaced by another representation? Which? 2. Generality review: If formalism is left out, how is generality ensured? 3. Justification of the term 'proof': How does the approach include reflection on how it differs from a formal mathematical proof? 4. Didactical intention: What is the target group and learning goal? Depending on the respective target group and learning goal, educators must choose different approaches associated with different representations (Tall, 1999).

To illustrate these differences, I will apply this framework to one transition-to-proof course (Grieser, 2017) and compare it to the categorization that Biehler and Kempen (2016) undertook of Wittmann and Müller's (1988) substantive-illustrative proof approach. To not exceed the scope of this dissertation, I will limit myself to these two very different examples. For further reading, Biehler and Kempen (2016) list several historical approaches and corresponding categorizations. Table 1 displays this comparison:

Table 1

Two Approaches to Proving

Approach	Substantive-illustrative proof	Transition-to-proof course
Source	Approach by Wittmann and Müller (1988), classification cited from Biehler and Kempen (2016)	Approach by Grieser (2017), own classification
Representation	Non-formal; actual numbers and patterns as examples	Formal symbols from Boolean Logic, Set Theory and Elementary Number Theory
Generality review	The authors argue that the applied mathematical operations are general, but leave open how the reader can recognize this.	The author states that a mathematical proof is needed to show the generality of a claim.
Justification of the term 'proof'	The authors undertake an elaborate reflection on what can be regarded as a proof and and lay out why a substantive-illustrative proof is a proof.	The chapter starts with a definition of proof and distinguishes the terms 'proof' and 'example'.
Didactical intention	Teaching proof to school children but also teaching proof approaches to pre-service teachers.	Teaching the primary model of formal mathematical proof to bachelor's students in mathematical subjects.

These examples illustrate well how approaches vary depending on the didactical intention and target group: Substantive-illustrative proof is designed for proof learners in school who are not yet familiar with formalism (Wittmann & Müller, 1988). Therefore, the representation uses established mathematical objects like numbers as examples. This leads to a divergence from formal mathematical proof standards. Thus, to convey this difference, educators must include reflection on generality and proof characteristics in the lesson (Biehler & Kempen, 2016; Wittmann & Müller, 1988). The transition-to-proof course by Grieser (2017) is a practical example of a proof course for learners who are moving on from school to university. As a consequence, its goal is to teach formal proof writing skills that meet research standards. To this end, the author includes a section on formal language before he leads over to the section on proving. This duality aligns with the model by

Selden (2013) which distinguishes between the reasoning component of proof and the aspect related to formal rules. In the section on proving, Grieser (2017) also reflects on what distinguishes proofs from examples and how only proofs can show the generality of a claim. In both examples, this reflection can help the learners understand what comprises a mathematical proof. Moreover, their obvious differences show the opposite poles of what can be regarded as proof, with the formal, research-related perspective on one side and the practical school-related perspective on the other.

2.1.5 Assessment Tools for Proof Competency

To determine the effectiveness of any proof intervention, researchers rely on valid instruments to measure PC. According to Waluyo et al. (2019), PC test instruments can be categorized into the categories of proof comprehension and proof construction. To provide a comprehensive picture, I will report on both types of assessment, although this work primarily focuses on fostering proof construction in children.

Proof Comprehension Mejía–Ramos, Lew, de la Torre, and Weber (2017) have developed and validated short, multiple-choice scales around three different mathematical proofs to assess proof comprehension in undergraduate students. The goal was to develop a reliable tool to help university instructors in formative assessment. To achieve this, the researchers conducted structured interviews, revised the items accordingly, and validated them with a sample of 200 students, resulting in an internal consistency of $\alpha \geq .7$. An example item of this tool is:

> MC2. Which of the following are examples of finite sets? Please select all that apply.
> (a) The set with the following elements: 1, 2, 3.
> (b) The set of real numbers between -2 and 2.
> (c) The set of all fractions 1/r where r is a natural number.
> (d) The set of integers greater than 4 and smaller than 10.
> (Mejía–Ramos et al., 2017, p.139)

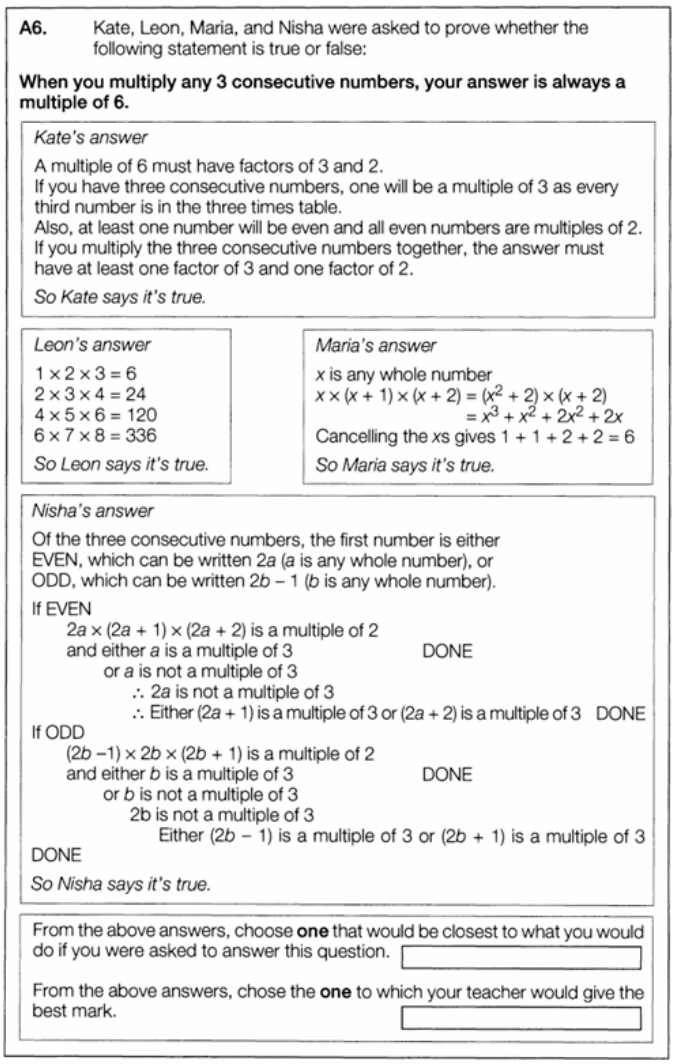
Remarkably, with items like these, the complex cognitive process of proof comprehension can be assessed via closed questions. These are preferable to open questions, as they are time-effective and easier to evaluate (Goecke, Staab, Schittenhelm, & Wilhelm, 2022). Even though the test shows good internal consistency, it only includes three different proofs, so it does not cover a broad spectrum of proof comprehension.

Healy and Hoyles (2000) have developed a test for school children that presents different proofs for the same claim, written by fictive children, and asks the students which version comes closest to the

method they themselves would employ and which would be preferred by their teacher. Additionally, the test has the children assess the generality, validity, and explanatory power of the given proofs (Healy & Hoyles, 2000). However, this test was evaluated with 14-15-year-old students, so the mathematical content might also be unsuitable for assessing proof comprehension in primary school children. An example item is depicted in Figure 1.

Figure 1

Example item from the proof assessment by (Healy & Hoyles, 2000, p.401).



A6. Kate, Leon, Maria, and Nisha were asked to prove whether the following statement is true or false:

When you multiply any 3 consecutive numbers, your answer is always a multiple of 6.

Kate's answer

A multiple of 6 must have factors of 3 and 2.
If you have three consecutive numbers, one will be a multiple of 3 as every third number is in the three times table.
Also, at least one number will be even and all even numbers are multiples of 2.
If you multiply the three consecutive numbers together, the answer must have at least one factor of 3 and one factor of 2.

So Kate says it's true.

Leon's answer

$1 \times 2 \times 3 = 6$
$2 \times 3 \times 4 = 24$
$4 \times 5 \times 6 = 120$
$6 \times 7 \times 8 = 336$

So Leon says it's true.

Maria's answer

x is any whole number
$x \times (x + 1) \times (x + 2) = (x^2 + 2) \times (x + 2)$
$= x^3 + x^2 + 2x^2 + 2x$
Cancelling the xs gives $1 + 1 + 2 + 2 = 6$

So Maria says it's true.

Nisha's answer

Of the three consecutive numbers, the first number is either EVEN, which can be written $2a$ (a is any whole number), or ODD, which can be written $2b - 1$ (b is any whole number).

If EVEN
$2a \times (2a + 1) \times (2a + 2)$ is a multiple of 2
and either a is a multiple of 3 DONE
or a is not a multiple of 3
$\therefore$ $2a$ is not a multiple of 3
$\therefore$ Either $(2a + 1)$ is a multiple of 3 or $(2a + 2)$ is a multiple of 3 DONE
If ODD
$(2b - 1) \times 2b \times (2b + 1)$ is a multiple of 2
and either b is a multiple of 3 DONE
or b is not a multiple of 3
2b is not a multiple of 3
Either $(2b - 1)$ is a multiple of 3 or $(2b + 1)$ is a multiple of 3
DONE

So Nisha says it's true.

From the above answers, choose **one** that would be closest to what you would do if you were asked to answer this question.

From the above answers, chose the **one** to which your teacher would give the best mark.

Proof Construction In the following, I will elaborate on assessment tools that have the participants *construct* proofs:

Kempen (2019) has developed a method of measuring the proof construction abilities of first-year university students. The test consists of a single, open assignment: "The sum 11 + 17

is an even number. Is this true for the sum of any two odd numbers? – Justify your answer convincingly!"(Kempen, 2019, p. 116 – translated from the German original)

To evaluate the answers via qualitative content analysis, the author has developed a categorical system based on previous models of proof quality and refined it in a pilot study (Kempen, 2019). This standardized procedure allows for a rather unbiased comparison of the students' proofs and the categories grounded on several years of research on proof-writing. Nevertheless, open questions are very time-consuming for both the test-takers and the evaluators (Goecke et al., 2022). Additionally, they bear a higher risk of coding errors compared to single- and multiple-choice tests. Lastly, the robustness of the test could be improved by including more than one item. For the target group of primary school children, this task might be comprehensible. Still, the categorical system for the raters will most likely not be applicable for this age group and would have to be adapted in line with the classroom definition of proof by A. Stylianides (2007).

Senk (1989) has created the CDASSG Proof Test. The test consists of six items, with four full proofs to be written and two short-answer questions, all to be completed within 35 minutes, and shows good internal consistency ($\alpha = .85$) (Senk, 1989). As with the other tests described above, this assessment is designed for proof novices who have acquired particularly more algebraic knowledge than an average primary school child. Thus, adapting this instrument to the target group seems unrealistic.

All tools discussed here can be used to evaluate PC in learners who have prior knowledge of formal reasoning and the use of algebraic symbols. This is why enrichment courses for this age group so far have only been evaluated by assessing the PC of the learners through observation or by assessing their constructed proofs based on rating schemes (Ball & Bass, 2003; Ko & Song, 2011; G. Stylianides, 2008). For a large-scale efficacy study of a primary school proof intervention, it is thus necessary to develop a novel test instrument.

2.2 Enriching Gifted Primary School Children with Mathematical Proving

2.2.1 Characteristics of (Mathematically) Gifted Children

The prevalence of individuals with high abilities has given rise to various models of giftedness. To design adequate education for gifted children, it is essential to consider these models and derive thoughtful actions from them. We will, from here on, use the terms 'gifted children' and 'talented children' in reference to the same group, as they are generally used as equivalents (Ziegler, 2008). One of the most traditional definitions of giftedness is the one proposed by Terman (1922) which describes giftedness as a single cognitive factor, namely high cognitive ability (i.e., IQ). In contrast,

the enrichment triad model does not only take cognitive ability into account but also creativity and passion for a task (Renzulli, 2016). It especially emphasizes the intersections between these three dimensions.

The multiple intelligences theory proposed by Gardner (1983) states that there are seven specific kinds of intelligence: linguistic, logical-mathematical, musical, bodily-kinesthetic, spatial, interpersonal, and intrapersonal. According to that model, giftedness is defined as excellence in at least one of these areas. As indicators for giftedness in the logical-mathematical domain, Gardner (1983) describes many characteristics that he derives from the vitae of prominent mathematicians. Käpnick (1998) condensed these into the following four abilities: "(1) Ability to flexibly apply rules of logic, (2) Ability to grasp and retain mathematical concepts, (3) Ability to recognize patterns, (4) Ability to identify and solve problems"[p.72].

Tannenbaum's (1983) model of giftedness encompasses both this domain-specific factor and a general factor of intelligence, as well as psychosocial skills, a supportive environment, and chance.

Later authors also describe giftedness as a multidimensional construct with a general cognitive factor and different domains of extraordinary performance, but additionally include the development from potential to talent (K. Heller & Schofield, 1993; K. A. Heller, Mönks, Subotnik, & Sternberg, 2000; Worrell, Subotnik, Olszewski–Kubilius, & Dixson, 2019). In these modern models of giftedness, the general factor is usually represented by fluid intelligence (gf), which corresponds to the ability to solve novel problems ad hoc, together with crystallized intelligence (gc), the cultivated knowledge component of intelligence (W. Schneider & McGrew, 2012). Consequently, the term 'mathematically gifted' describes students with potential in the mathematical domain (Bicknell, 2008). According to Foth and van der Meer (2013), there is much interplay between mathematical giftedness and general factors of giftedness: Mathematically gifted children are more likely to reach higher levels of gf and gc than other intellectually gifted children.

One model combining the approaches from several talent development models is the *Talent Development Mega Model (TDMM)* that arranges the components of several other prominent models into one (Subotnik, Olszewski–Kubilius, & Worrell, 2011): It integrates individual cognitive abilities as the basis of talent and their development over time, the role of the environment and psychological strengths as well as eminence and contribution to the domain as the final outcome (Subotnik et al., 2011). This integrative model has then been adapted and refined by Preckel et al. (2020), who state that talent development takes different trajectories depending on the specific domain. This resulted in the framework *Talent Development in Achievement Domains* (TAD framework), which suggests different predictors and influencing circumstances for talent development in the various domains, of which mathematics is one. They found that previous mathematical development research had

mostly neglected diagnosing mathematical giftedness in primary school children and instead focused on the identification of children with developmental deficits. Therefore, broad measurements of children's mathematical performance related to cardinality and ordinality were undertaken, as well as numerical tasks, and spontaneous focusing on numerosity (SFON) (Preckel et al., 2020). All of these turned out to be useful predictors of mathematical giftedness, as high performance in these tasks is associated with higher mathematical performance later in life (Preckel et al., 2020) and mathematical performance is one main characteristic of mathematical giftedness (Koshy, Ernest, & Casey, 2009).

Nevertheless, insights into mathematical giftedness are not limited to psychological research on giftedness. The prevalence of mathematically exceptional individuals has led to many approaches within mathematics education research to model this specific talent. The most prominent model of mathematical giftedness goes back to Krutetskii. He lists the following characteristics by which to identify mathematically gifted youths:

1. Formalized perception of mathematical structures, i.e., the ability to abstract from content and grasp only the formal structure of a given mathematical problem;
2. (Fast and wide) generalization of mathematical content and problems (i.e., a concrete problem is recognized as a special case of a more general problem);
3. Shortening of a train of thought and thinking in super-ordinate structures;
4. Flexibility in thought processes that allow an easy and quick switch from one thought operation to another qualitatively different;
5. Striving for clarity, simplicity and also elegance of a solution;
6. permanent and fast recall of mathematical knowledge

(Krutetskii, 1968, p.46, according to Bardy & Bardy, 2020, translated into English by the author).

Several aspects in this model seem familiar to the four indicators of logical-mathematical giftedness cited above (Käpnick, 1998), suggesting an overlap between the psychological and didactical perspectives on mathematical giftedness. Further characteristics described by current authors include advanced logical thinking, dealing with numbers and symbols, complex reasoning, and recognition of patterns (Deal & Wismer, 2010; Koshy et al., 2009; Leikin, Koichu, & Berman, 2009). All of these characteristics influenced this dissertation and the corresponding intervention.

2.2.2 Fostering Mathematically Gifted Children

Fostering gifted and talented children is not only crucial for their personal development: These children can be future decision makers, contribute novel methods, and solve societal problems if

they get the opportunity to unfold their gifts (Diezmann & Watters, 2000; Koshy et al., 2009; Renzulli, 2016). For mathematically gifted children, the content of curricular math lessons is not challenging enough (Diezmann & Watters, 2000; Rotigel & Fello, 2016). To approach this issue, Ziegler (2008) lists four pillars of possible fostering measures: acceleration, enrichment, pull-out programs, and ability grouping. While acceleration allows an individual to impel their intellectual career by enrolling earlier for primary school or skipping school years, enrichment comprises additional learning opportunities with content outside the school curriculum or deeper insights into school topics (Subotnik et al., 2011). In pull-out programs, students can skip certain lessons and instead attend enrichment courses, lessons for higher grades , or academic programs (Ziegler, 2008). Ability grouping is a measure affecting the whole age group within a school: Here, students are, at least in certain subjects, differentiated by their performance in that domain and instructed in more heterogeneous groups (Käpnick, 1998). Käpnick (1998) argues that in most cases of mathematically gifted children enrichment is the most appropriate action, as they need a vivid social environment, which could be endangered by the other forms of fostering. Furthermore, mathematically gifted children do not necessarily perform extraordinarily in all subjects, which would be an obstacle to, for example, acceleration (Käpnick, 1998). Also, for young children at primary school level, , enrichment is especially important, as at this stage of talent development, they need to explore different intellectual fields so they can discover which disciplines they would like to embrace (Renzulli, 2016).

For enrichment in the domain of mathematical talent, Bardy and Bardy (2020) curated a number of fostering goals that match the needs and characteristics of the target group:

- Encouraging the use of heuristic aids such as informative sketches, tables, or variables;
- Teaching general problem-solving strategies such as systematic trial and error, working backwards, recording relationships using equations or inequalities, forming analogies, and changing representations;
- Promoting structuring, abstraction, and generalization;
- Promoting logical thinking (reasoning in Thurstone's sense as a primary factor of intelligence);
- Promoting (rational) argumentation and reasoning;
- Introduction to (mathematical) proof;
- Promoting spatial awareness;
- Teaching an adequate picture of mathematics ... ;

- Introduction to expert culture (for the young people being supported)

(Bardy & Bardy, 2020, p.165, translated into English by the author)

The aspects of problem solving and reasoning were also discussed in Section 2.1.2 as competences related to proving. Building on this, I will more closely examine ways of enrichment related to these two aspects before discussing proof-related enrichment in a dedicated section.

Problem Solving Bardy and Bardy (2020) name problem solving and the corresponding strategies as one fostering goal. In a section of their book on mathematical giftedness, they describe how problem solving can be taught with so-called *problem fields*, a term established – among others – by Zimmermann (1991). A problem field is an open collection of mathematical problems, similar in terms of content or building on each other, allowing learners to develop strategies on one problem and reuse them for others (Bardy & Bardy, 2020). Mathematically gifted children tend to come up with their own variations of given problems within a problem field (Kiesswetter, 1985). Though this can be a challenging task for young students, problem variation has great potential as a task for problem solving courses, as it also fosters the children's understanding of a problem, their flexibility in thought, and their motivation (Bardy & Bardy, 2020). Furthermore, solving an exercise that fellow children perceive as a mathematical problem can be trivial for mathematically gifted students while posing new problems of the same sort can still be an appropriate challenge (Singer, Sheffield, Freiman, & Brandl, 2016). Interventions featuring problem solving and problem posing had positive effects on the mathematical creativity and talent of the target group (Singer et al., 2016). Leikin et al. (2009) even synthesized research on giftedness and problem solving by connecting the giftedness model by Renzulli (2016) with different styles of approaching a problem, leading to new ways of identifying gifted children based on how they solve problems.

Reasoning Similarly, Jablonski and Ludwig (2022) describe how reasoning can serve both as a means of identification and as suitable content for fostering measures targeted at gifted students. Nevertheless, Øystein (2011) has found that high-achieving students need additional guidance in order to demonstrate creative and extraordinary ways of reasoning. Therefore, enrichment programs should feature reasoning activities so the students can practice this skill. Furthermore, mathematically gifted children tend to overlook that explicit reasoning is needed to support a claim, as it might seem trivial to them that it applies (Bardy & Bardy, 2020). Thus, enrichment activities should also evoke a need for justification in the children before teaching them reasoning techniques (Bardy & Bardy, 2020). Also, as pointed out above, mathematically gifted children need challenging tasks (Diezmann & Watters, 2000; Rotigel & Fello, 2016). A challenging task is most

of all characterized by its authenticity, which depends on its complexity, the unavailability of a ready-made solution, and the necessity of high-level thinking and reasoning (Diezmann & Watters, 2000).

2.2.3 Enriching Gifted Primary School Children with Mathematical Proving

The previous section has illustrated that many researchers recommend problem solving and reasoning for mathematical enrichment. The familiarity between these fields and proof, which I have described in Section 2.1.1, indicates that proof could be a valuable activity in this context as well. In this section, I will discuss the suitability of authentic mathematical proving for gifted primary school children.

In the list of fostering activities for the mathematically gifted cited in Section 2.2.2, Bardy and Bardy (2020) explicitly list *Mathematical proof* as one field of fostering, but most of the other mentioned aspects also correspond to proving: Reflections on *generalization* and *abstraction* are characteristic for any approach to teaching proving (Biehler & Kempen, 2016), and structuring a chain of *arguments* is one step in a proof according to Boero (1999). Similarly, *logical thinking* is necessary for constructing a proof (Rota, 1997). Finally, both an *adequate picture of mathematics* and a mathematical *expert culture* require proving, as it is central to mathematical research and the main activity of professional mathematicians (Carl, 2022; Rav, 1999; Villiers, 2012).

Considering how late in the curriculum the art of proving is introduced to students, it stands to reason that proving is maybe *too* challenging and not suitable for primary school students. However, there are considerable reasons for teaching proving to gifted primary school children: Firstly, enrichment affords extracurricular content (Subotnik et al., 2011). Therefore, the absence of proof lessons in school speaks for their suitability for mathematical enrichment. Secondly, Gadanidis, Hughes, Minniti, and White (2017) argue that higher education topics (in their case: programming) should not be perceived as unsuitable for children simply because they are not part of the curriculum. Gadanidis et al. refer to so-called *what-might-be-settings* to argue that a lack of example cases should not deter educators from teaching unfamiliar content to children. Lastly and notably, mathematical proving with all the challenges it entails is well-suited to the characteristics of mathematically gifted children: Comparing the model by Boero (1999), which describes the steps of a proof, and the mathematical giftedness model by Krutetskii (1976), we find that mathematically gifted children possess abilities particularly applicable to the art of proving: They have a formalized perception, strive for generalization and shorten thought chains to reach elegant solutions (Krutetskii, 1976). These characteristics may also help them overcome the difficulties that average children had with the New Mathematics curriculum in the past century.

Furthermore, international examples demonstrate that formal proving activities can be successfully implemented even in primary school lessons. Several interventions with gifted primary school children in South Korea serve as promising examples of how to integrate geometric and algebraic proofs into mathematical enrichment lessons (Chang et al., 2006; Ko & Song, 2011; K. H. Lee, 2005; Na, 2011). But even when the target group is not gifted, didactical research from the US indicates that primary school students can reach fairly advanced levels of proof if the topic is introduced carefully (Ball & Bass, 2003; Bleiler, 2009; A. J. Stylianides, 2016). These arguments and examples encouraged the approach of developing a proof course for gifted primary school children. In the next part of this work, I will further explain why I chose the format of an asynchronous online course for this enrichment course.

2.3 Asynchronous Online Courses for Gifted Children

2.3.1 Challenges for On-Site Enrichment Courses

Classical enrichment courses, held in a synchronous, on-site class setting, provide several benefits, including peer-to-peer interaction (Käpnick, 1998). However, not all talented children can take advantage of these opportunities, as several obstacles pose themselves (Lohaus & Wild, 2021). Firstly, high-quality on-site courses often come with high tuition fees and additional material costs. Additionally, transporting the child to the course site can cause logistical effort and further expenses because young children cannot commute independently (Weaver, Shaul, & Lower, 2022). These aspects are particularly challenging for low-income families, resulting in inequality in talent development. For families that can afford many extracurricular activities, conflicting schedules can still be a problem (Fölling–Albers, 2000), forcing them to choose courses based on time preferences rather than by achievement domain. Another problematic aspect is the great heterogeneity among talented children. For statistical reasons, the first decile of a normal distribution covers a diverse group of people than the second or third. Thus, teaching a class for gifted children requires the teacher to accommodate the variable levels. This is hard to accomplish, as teachers only have a limited amount of time per student (Diezmann & Watters, 2000). Lastly, most enrichment courses have limited capacities, which means there is only a defined number of course seats. This requires a selection mechanism, such as a test or a nomination process. While teacher nominations are biased against girls and children with an immigrant background (Golle et al., 2022), tests are expensive to administer and also discriminate against certain groups if the items are not culture fair. Thus, it is necessary to complement the enrichment landscape with more flexible courses. In the next section, I will outline how asynchronous online enrichment can fill this gap.

2.3.2 Chances for Asynchronous Online Enrichment Courses

Asynchronous online courses, i.e., online courses that do not require the students to study at a specific time, are very popular in today's educational landscape, as they offer several benefits: First, flexibility and convenience allow for the integration of asynchronous online courses into the everyday lives, even of very busy students (Lin & Gao, 2020; Weaver et al., 2022). Thus, students can choose their courses solely based on content preferences without having to consider time preferences. Digital learning provides easier access to interesting content and allows students to study according to their individual needs (Leikin, 2021). Second, both students and educational institutions benefit from the geographical freedom and cost-effectiveness that come with the absence of a physical classroom (Weaver et al., 2022). These aspects promote a more diverse audience in terms of socio-economic status and living area (Lohaus & Wild, 2021; Weaver et al., 2022). Furthermore, asynchronous online courses permit self-paced learning (Lin & Gao, 2020). In this format, learners can choose what to review and how much time to allocate to each task, which is particularly beneficial when it comes to difficult tasks and also results in better, but not longer, study sessions (Tullis & Benjamin, 2011). Another aspect worth mentioning is that automated real-time feedback can be smoothly embedded into asynchronous online courses. This feedback can improve practice motivation, the ability to review one's own performance, self-concept, and post-training performance (J. Schneider et al., 2016). Lastly, asynchronous online courses can foster the development of critical thinking skills: Students can then engage in thoughtful discussions and reflect on their responses (Aloni & Harrington, 2018). This kind of reflection, can be helpful for gifted children, especially in the field of proof learning and argumentation, as many of them still need to discover the necessity of providing a proof for a claim (Bardy & Bardy, 2020).

The fit between these aspects and gifted education becomes apparent when considering the model of technology use in gifted education by J. Chen, Yun Dai, and Zhou (2013). The model lists three main functions: The first function, *enable*, describes the improvement of access to enrichment programs. Here, the authors see potential in digital technologies to maximize the number of course seats and to provide access to children in rural areas. Secondly, *enhance* describes the function of digital tools to refine fostering programs, for example by allowing for self-paced learning. Lastly, *transformation* summarizes novel, previously unthinkable approaches, such as individualized learning or a community knowledge base. All these aspects are immanent in asynchronous online courses, as described at the beginning of this section.

A similar picture arises when considering Pyryt's enrichment matrix that he applies to "[evaluate technologies] in terms of their capacity to accelerate the Pace of learning, facilitate the development

of higher-order [P]rocess skills, allow students to pursue their Passion areas, develop a variety of [P]roducts, and interact with intellectual Peers"(Pyryt, 2009, p.1173). Especially the first of Pyryt's five P of enrichment, namely Pace, indicates the suitability of asynchronous online courses for gifted learners, but the remaining aspects can be integrated into such courses as well.

Widening the perspective, we find that asynchronous online learning is not only beneficial for out-of-school enrichment. As described by Ziegler (2008), pull-out courses can be useful for the mathematically gifted. With the development of new technologies, novel attempts have been made at developing so-called digital pull-out courses in which gifted children have the possibility to take an asynchronous online course while their fellow students attend their usual lesson (Huth, Pollok, & Schreiber, 2024; Kohnen & Fischer–Ontrup, 2023). Such technological set-ups can also cater to the different needs in a heterogeneous gifted classroom (Leikin, 2021).

2.4 Research Questions

Proving is a core component of the mathematical world (Dawkins & Weber, 2017; Ross, 1998; Schoenfeld, 2009) and crucial to understanding why something is true (Villiers, 2012). It is also closely linked to the 21st century skill of problem solving. Still, proof and the necessary competencies can rarely be found in school curricula. This is why this dissertation investigates the extent to which proof can be taught in primary school.

The population for this investigation were gifted and talented primary school students interested in mathematics. Firstly, to foster gifted children, extracurricular and appropriately challenging content is needed. Mathematical proving covers both dimensions. Secondly, most proof courses are designed for adult learners. Thus, there is a massive gap to bridge in the didactical reduction of the topic. Therefore, a course designed for highly capable primary school students can be a meaningful first step and the feedback of this group is crucial to the ongoing development process. Lastly, learning how to construct mathematical proofs can be especially relevant for gifted and talented learners, as these individuals are likely to become future decision-makers or developers of cutting-edge technologies. Thus, it seems worthwhile to study how gifted children respond to primary proof education.

The overarching goal was to determine how enjoyable and effective proof learning is in an asynchronous online course for talented primary school children. To this end, a dedicated intervention was designed to enhance the mathematical PC of talented primary school children (see: Section 4). In line with the considerations from Section 2.3.2, I chose the format of an asynchronous online course as it suits the demands of (digital) enrichment courses well.

In the first study (*Development and Pilot of an Asynchronous Online Enrichment Course on*

Mathematical Proving), I describe and evaluate the course design. As the content and learning format are fairly unusual for primary education, the main aim was to determine which aspects of the course already aligned with the needs of the target group and which needed further refinement. Therefore, the first version of the course was rolled out in a pilot study. The main research question in Study 1 was: *Is an asynchronous online mathematical proof course a feasible option for fostering gifted children?* To thoroughly assess this question, I considered the perspectives of children and parents as well as their behavioral traces. The results of these investigations informed the further revision of the course.

After piloting and revising an intervention, an efficacy study should be conducted to monitor how the targeted skill changes in the treated sample (Nelson et al., 2012). Here, the relevant competence was mathematical proving. Thus, a test instrument for conducting pre- and post-measurements of PC was necessary. However, at this point in the research, test instruments for proofreading and proofwriting only existed for strikingly older target groups. Therefore, we developed a new test instrument – the Preformal Proving Test (PfPT). To validate this assessment, we conducted an online study as well as a proctored on-site study, which we will subsume under Study 2 (*Development and Validation of a Preformal Test for Mathematical Proof Competency*) and discuss in Section 5. In Study 2, we aimed to develop a psychometrically sound model of mathematical PC to sample items for the final test from the initial item set and to investigate the feasibility of the test design both online and in class. Furthermore, we intended to study the nomological network of PC.

In Study 3 (*A Self-Paced Online Course to Introduce Mathematical Proof to Talented Primary School Children: A Randomized Controlled Trial*), we then conducted an efficacy study corresponding to the intervention from Study 1. To this aim, we implemented a randomized controlled trial (RCT) with a wait-list control group (see: Section 6). In a pre- and posttest, we measured PC using the PfPT (Stein, Tsarava, & Goecke, 2025). Expecting a positive treatment effect on PC, we preregistered one affirmative research question: *What is the effect of attending the course compared to not attending the course on children's proof skills?* (Registry of Efficacy and Effectiveness Studies [#20667.1v1]). Additionally, we included several variables of domain-specific motivation to monitor possible effects. Regarding the non-cognitive variables, we studied the exploratory research question: *Is there an effect of attending the course compared to not attending the course on children's domain-specific self-concept, interest, attainment value, utility value, and persistence?*

In Study 4 (*Investigating Motivational Fluctuation and Dropout Among Talented Primary School Children in a Self-Paced Online Mathematics Course*), we aimed to investigate the motivational trajectories of the children during the proof course more deeply. These insights can help to further scaffold the learning processes in future interventions with this challenging content and format.

To do so, we analyzed the non-cognitive data from the pre- and post-tests of Study 3 as well as additional data on the children's motivational state that we had collected using single-item indicators throughout the course. First, we investigated the extent of motivation fluctuation in the children during the course (RQ1). Secondly, we examined the relationship between dispositional (trait-like) motivation and situational motivation, which is state-like and may vary throughout the course (RQ2). Our research also analyzes changes in situational motivation throughout the course (RQ3) and the impact of situational parental support on changes in children's motivation (RQ4). Finally, the study seeks to identify which factors (dispositional motivation, situational motivation, or parental assistance) can reduce the risk of dropout, and to what extent (RQ5).

3 Materials and Methods

3.1 Context and Development Procedure

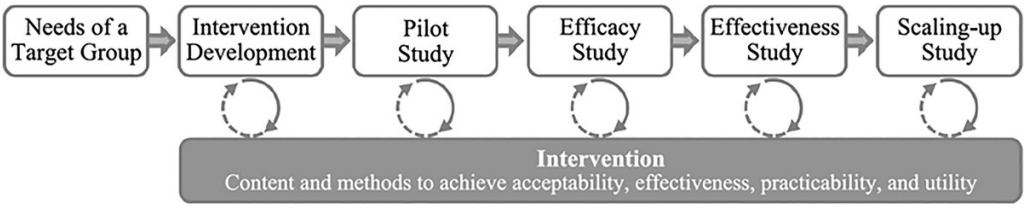
The course was developed within the Hector Children's Academy Program (HCAP), a statewide enrichment program for gifted and talented primary school students in the German federal state of Baden-Württemberg (Trautwein, Golle, Jaggy, Hasselhorn, & Nagengast, 2023). As described in Section 2.2.2, it is crucial for a society to provide its young talents with talent development measures (Subotnik et al., 2011). Still, it is challenging for teachers to include such measures in a regular classroom due to tight schedules, heterogeneous classes, and lack of professional knowledge (Dimitriadis, 2016).

Therefore, the HCAP offers extracurricular courses for K-4 children who demonstrate exceptional interest and dedication. At all primary schools in Baden-Württemberg, school teachers can nominate the top ten percent of their class to participate in the program. If the child and their legal guardian agree to participation, the child will be enrolled at the nearest of the 70 HCAP sites and can choose from the local course register freely. The courses feature mostly STEM topics and have a focus on fostering interest and motivation with extracurricular content (Trautwein et al., 2023). To cover a broad variety of topics, experts from all STEM domains can offer their own courses at their local HCAP site. However, individuals without a pedagogical background tend to lack practical instructional knowledge. This is why the scientific monitoring of the HCAP has developed a professional training program for course instructors to increase the quality of their courses (Trautwein et al., 2023). Nevertheless, enrichment programs cannot work if they are not grounded in giftedness research and theory (Dimitriadis, 2016). Therefore, the scientific support of the HCAP is constantly monitoring and improving the so-called Hector Core Courses (HCCs) – courses with a theoretically sound foundation that are empirically tested before they are rolled out for all

HCAP sites (e.g., Herbein et al., 2018; Rebholz et al., 2017; Schiefer et al., 2021; Stark, 2025). The development procedure for HCCs is grounded on the recommendations for intervention development by Humphrey et al. (2016) , visualized by Herbein et al. (2018) as depicted in Figure 2.

Figure 2

Steps of intervention development and validation (Herbein et al. 2018, p.177, based on Humphrey et al. 2016; Lendrum and Wigelsworth 2013).



First, the learning goals and content field for the course are chosen based on the needs of the target group. The foundation for this choice can be built based on previous literature about the potentials and interests of gifted primary school children or on own investigations of the HCAP among the nominated children. In the next step, core components that target these learning goals are developed for the course (Nelson et al., 2012). From these core components, the researchers in charge then derive the specific activities for the intervention course and shape them into a number of eight to 12 course sessions. As soon as the course materials are drafted, a pilot study is conducted with a group of children in one or two academies. In this pilot study the feasibility of the course activities for children is investigated and the content receives feedback from the children regarding suitability and appeal. The course is then adapted accordingly. After that, the course is offered in a randomized wait-list control group setting to a larger number of children to investigate the effects on the target constructs. In this efficacy study, the course is usually taught directly by the developer. If this leads to significant positive effects, an effectiveness study takes place in which trained course instructors teach the course. This makes it possible to observe if the effects persist even if the course is taught by different teachers. Finally, teachers from all HCAP sites can register for the training that qualifies them to teach a specific HCC. After that, a scaling-up study can be conducted in which the course is offered in a large-scale setting at many HCAP sites.

Since 2020, the HCAP also maintains a statewide Learning Management System (Moodle LMS) to offer online courses across locations. As described in Section 2.3.2, asynchronous online courses can be a beneficial form for an enrichment course. Therefore, the intervention *Logical Detectives* was developed as the first asynchronous online HCC in the HCAP. After considering the needs and characteristics of mathematically gifted students (see: Section 2.2.1), I decided to focus on mathematical proving as a course topic, more specifically: on basic proofs in Boolean Logic, Set

Theory, and Elementary Number Theory. In the following paragraphs, I will lay out how *Logical Detectives* was developed and how the four research articles included in this dissertation correspond to the process.

3.2 Core Components of the Course

Following the approach of Nelson et al. (2012), I based the course development on four theory-driven core components: *Self-paced learning, Automated real-time feedback, Iconic and symbolic logical reasoning* and *Natural language proof writing* (Stein, Tsarava, Fabian, et al., 2025). The first two components address the mathematical content of the course: According to Bronkhorst, Roorda, Suhre, and Goedhart (2021), *iconic and symbolic logical reasoning* helps learners acquire proof skills in Boolean Logic and Set Theory. This is especially important, as it matches two of the three basic proof topics described in Section 3.3. But logical understanding is not sufficient as a prerequisite for writing proof. According to the model of proof composition by Boero (1999) cited in Section 2.1.2, writing a proof requires forming a chain of arguments and putting it into formal expressions. *Natural language proof writing* is, therefore, crucial to bridging the gap between the children's (informal) prior reasoning skills and the formal logic that mathematicians use. In addition, repeatedly writing their own proofs in a drill-and-practice environment allows proof learners to develop necessary routines and formalism (Carl, 2022). Nevertheless, course design should not only address questions of subject matter didactics but also create a fruitful learning environment. Thus, the other two core components target the cultivation of a fruitful learning environment within the asynchronous online setting: *Self-paced learning* means that the children can freely choose how long to work on which course activity, without a teacher or a class schedule determining the pace. This can improve learning results, especially when there are difficult tasks and the learners decide to allocate much time to these (Tullis & Benjamin, 2011). Still, children need constructive support to learn from possible mistakes. Thus, the design team implemented *automated real-time feedback* throughout the course, with feedback prompts that appear directly after the learner completes a task. This kind of immediate feedback can have a positive impact on learners' meta-cognition, performance, and self-concept (J. Schneider et al., 2016).

3.3 Development and Design of the Course Chapters

Introductory courses on mathematical proving in university contexts most commonly feature the disciplines of Boolean Logic, Set Theory, and Elementary Number Theory (e.g., E. Brunner, 2014; Carl et al., 2022; Glosauer, 2019). Students discover the meaning of logical symbols as well as the

rules of transforming equivalent expressions into one another. This formal knowledge is one part of proof, with reasoning skills being the other one (Selden, 2013). Students must learn how to apply both, as a proof is a chain of mathematical arguments, which is then formalized (Boero, 1999).

Interestingly, the logical operations from these three mathematical fields (e.g., union, intersection) that are needed for writing a proof constitute algebraic structures similar to the arithmetic functions taught in primary school (e.g., addition, subtraction). Examples of how this content was implemented in curricula for K-4 children can be found in the New Math movement from the 1960s or in today's Early Algebra curricula (Engledowl, 2020; Hanna & Knipping, 2020). Hence, I decided to also address Boolean Logic, Set Theory, and Elementary Number Theory in the intervention course. I designed three content-based course chapters, each focusing on one of the three content pillars (Chapters 2 to 4). Additionally, I developed a chapter to introduce formal proving and to practice reasoning (Chapter 1). These exercises were meant for transferring the duality of formalism and reasoning described by Selden (2013). An introductory chapter was placed before Chapter 1 to familiarize the children with the asynchronous online course (Chapter 0), as unfamiliar online interfaces can increase extraneous cognitive load (Hollender, Hofmann, Deneke, & Schmitz, 2010; Skulmowski & Xu, 2022). In the following paragraphs, I will present the general structure of the course chapters, describe the two technological instruments primarily employed for this project, and describe the content of every chapter.

3.3.1 Structure of the chapters

To provide the children with scaffolding for the course, each chapter is structured in the same way. For completing the main activities in a chapter, the children could earn digital badges, as this is generally perceived as fun and motivating (Amaefule, Britzwein, Yip, & Brod, 2025). In the following, I will present each course activity on a general level and, after that, give insight into the development and content of these activities in the single chapters.

At the start of each chapter, the children watched a *preview video* in which HCAP mascot Hasel provides an overview of the chapter's topic and exercises, because advanced organizers and pedagogical agents can promote self-regulated learning (Taub, Mudrick, & Azevedo, 2018). Following Mayer's (2009) multimedia principles, I combined spoken information with a graphic depiction: In the videos, Hasel is standing next to a tiny monitor with a screen-cast that offers a glimpse into each element of the respective chapter. This is meant to motivate the children to start the chapter and to demonstrate the use of the interface and the navigation of the activities.

The second course element was called *exploration* and is implemented as an H5P interactive presentation (H5P Group, 2013). In every exploration, the children are presented with a picture

related to the overarching detective story (e.g., Hasel's office, a secret notebook). In this picture, they can click on several objects to see what lies behind. They can find riddles, hints, or definitions of new expressions. They can always navigate back to the primary picture. This allows the children to choose their own path through the exploration and repeat exercises as they like, thus making use of the advantages of self-paced learning (Tullis & Benjamin, 2011).

In the middle of every chapter, a *case* is presented, a little puzzle for children to solve by employing logical reasoning. This puzzle is meant for rehearsing logical thinking and repeating the new content introduced in the exploration. To keep the children from being overwhelmed, the case does not contain any new input. It also aims to give the children a sense of achievement by solving a short yet challenging task. Such successes are needed to keep children motivated in self-paced settings (Amaefule et al., 2025). Also, reasoning exercises are beneficial for both proof learning (A. J. Stylianides, 2016) and mathematical enrichment (Bardy & Bardy, 2020).

After each *case*, the *proof* of the chapter is presented. It takes the form of an H5P interactive video (H5P Group, 2016a) in which mascot Hasel narrates a detective story that ends with him explaining an example of a proof corresponding to the story. The children are then asked to complete one or two similar proofs in a drag-and-drop task. In this way, they can apply their knowledge from the previous activities to a real proof text, but still obtain enough scaffolding and feedback. Thus, the *proof* serves as a preparatory exercise for the *proof collections*.

The *proof collections* at the end of Chapters 2-4 are the core activities for practice and deeper understanding. Here, the children apply their knowledge from the whole chapter and write full natural language proofs for given conjectures. They receive instant feedback on their use of syntax and logic so they can revise their answers. The learning environment for these exercises is an adaptation of the drill-and-practice proof tool Diproche by Carl (2022). A detailed description of the original tool as well as the implemented changes is contained in Section 3.3.3. I embedded the tool into the Moodle course using iframes, so the children did not have to switch between platforms.

Lastly, the chapters that contain a *proof collection* also feature a *notebook*. More of a repository and less of an activity, the *notebook* provides the possibility to look up all the symbols and definitions introduced before, so the students have them at hand for their natural language proofs. These were implemented using the Moodle format *Glossary* (Moodle, 2023), as it can help learners to retain new expressions in self-paced settings (Ratz, 2016).

Additionally, some chapters contain little *games* that allow the children to repeat symbols or new words from the respective chapter, for example in a memory game or an adaptation of one of the simpler activities within Diproche.

3.3.2 H5P Elements

Digital learning allows for the integration of a multitude of different formats to create learning opportunities (Leikin, 2021). As described above, the intervention course uses different H5P content types for different parts of the chapters. In this section, I will briefly describe each content type and explain how the different types can encourage the users through feedback and interaction.

Interactive Video (H5P Group, 2016a). This type refers to a video that has the possibility to embed several other content types at certain time stamps. These include quizzes, prompts, and drag-and-drop tasks like those that were used in the activity *proof*. To complete these tasks, users must move pictures or words to the correct position using their computer mouse.

Course Presentation (H5P Group, 2013). The main content type in the course is the interactive Course Presentation. It consists of several uploaded slides and can also contain other H5P content types, like drag-and-drop tasks or multiple-choice quizzes. It can be navigated using hyperlinks that forward the user to another slide by clicking on a dedicated area of the current one (e.g., the desk in Hasel's office).

Memory Game (H5P Group, 2014). In the Memory Game, users have to find matching pairs of cards by clicking on the cards to turn them around. The cards belonging to a pair do not necessarily have to show the same picture, which is why I used them as an exercise for matching symbols with their verbal or figural representation.

Image Pairing (H5P Group, 2018). Quite similarly to the memory game, this content type also helps create a game in which image pairs have to be found. The main difference is that the cards are all visible from the start and do not need to be flipped; also, they are presented to the user in two piles, each containing one of each pair. I used this content type as a variation of the Memory Game in a more challenging context, namely the chapter on Set Theory, to reduce cognitive load.

Personality Quiz (H5P Group, 2016b). This content type allows the creator to include several closed questions and, at the end, display a message to the user based on their answering behavior. It is used in Chapter 4 for the activity *Hasel guesses your number*. This activity contains three yes/no-questions that represent the ones and zeros of numbers from 0 to 7 in binary format.

3.3.3 Diproche

The course not only includes H5P-elements. The proof collections as well as some of the games are implemented as iframes, embedding content from an external website that our team set up to host an adaptation of Diproche. Diproche (= Didactical Proof Checker) is an online tool that provides written feedback on natural language proofs and was developed at Europa University Flensburg to give first-semester mathematical students more opportunities to practice proof writing (Carl, 2022). To make this tool usable for primary school children, we refined it by rearranging the exercises and replacing some of the propositions with easier ones. This was done to make sure none of the exercises would require knowledge of secondary school mathematics and to sort the exercises by increasing difficulty for motivational reasons. Also, more on-screen buttons were included to make entering solutions easier. A change of tone for the feedback phrases and color changes to fit the HCAP corporate design were meant to make the tasks more child-appropriate in general appearance. Also, users can get a glimpse of the solution texts to the proof exercises by clicking a dedicated button.

In the following sections, each chapter will be described in regard to its content and implementation.

3.3.4 Chapter 0

Chapter 0 familiarizes the children with the self-paced asynchronous learning environment. Table 2 provides an overview of the course activities in this chapter.

Table 2

Structure of Chapter 0.

Element	Title	Objective	Implementation
Preview	Preview Chapter 0	Overview Chapter 0	H5P Video
Exploration	Hasel's Letter	Familiarization with the course handling	H5P interactive presentation
Case	Investigation	Cognitive activation	H5P Image Choice
Proof	The Bracelet	Learning about proof structure	H5P interactive video

In the *preview video*, the children are introduced to the chapter's structure as well as to Detective Hasel who addresses them for the first time. He shows the badges the children can earn in Chapter 0 and gives insight into the following activities. The next course element is the very first *exploration* in which they learn to handle interactive presentations, i.e., how to notice clickable surfaces, how to go full screen, and how to navigate back to the start. The *exploration* starts with a letter from Detective Hasel in which he invites the children to visit his detective's office. The children can find

several clickable things in Hasel's suitcase, which is depicted in Figure 3: First, there is a map that leads to an advanced organizer of the whole course. Secondly, they can find a picture of a badge that leads to an explanation of the badge system. Also, a drag-and-drop puzzle and a small riddle can be found.

Figure 3

Screenshot of the exploration in Chapter 0.

After the *exploration*, the participants are led on to the first *case*. Here, the task is to choose the right animal footprint based on several rules. The next activity is a *proof*. As the children are not yet familiar with any algebraic symbols, it features a child-friendly adaptation of the letter game (Gernes, 1999; Hofstadter, 1999). The children have to arrange colorful beads and into 'bracelets' that Hasel's friends are making for each other. In the video, Hasel presents four rules for transferring one 'bracelet' to another and provides an example of how to apply these rules. After that, the children have to prove that a given 'bracelet' can be transferred to another in a drag-and-drop puzzle. This exercise prepares the learners for structuring their own proofs at a later point during the course.

3.3.5 Chapter 1

In Chapter 1, the children are introduced to the idea of proving through reasoning exercises and activities centered around mathematical proving. All activities are listed in Table 3.

Table 3

Structure of Chapter 1.

Element	Title	Objective	Implementation
Preview	Preview Chapter 1	Overview Chapter 1	H5P Video
Exploration	Hasel's Office	Familiarization with self-paced learning; cognitive activation; practice reasoning	H5P Interactive Presentation
Case	The Old Photo	Switching between different representations	H5P Drag and Drop
Game	Contradiction Day	Concluding by contradiction	H5P Memory Game
Proof	The Hasel Code	Writing first complete proof	H5P Interactive Video
Notebook	Notebook Chapter 1	Practicing looking up rules	Moodle Glossary
Proof Collection	Hasel Codes	Practicing proof stucture and reasoning with rules	Own Game (embedded)

Since the *preview video* has the same structure in each chapter, there will be no further description of the individual videos. The *exploration* of Chapter 1 is set in Hasel's office, where children can click on the furniture and several other objects to reveal what lies behind. There is a total of nine different items to explore: Four objects each lead to a logic puzzle from a different topic, clicking on Hasel himself displays additional information on the character, an armchair reminds the user to take a break every now and then, Hasel's notebook contains some important definitions and examples of terms related to proving, and the desk lamp hides an activity in which the children can put their knowledge from the notebook to the test. Lastly, clicking on the pencil cup leads to the question of how many pencils there are in the whole office. When the solution is entered, the user discovers that it is equal to the number of items to be explored in the office. The *case* in Chapter 1 is a logic puzzle in which the children have to reconstruct a rabbit family's portrait in a drag-and-drop activity based on seven written statements. It is followed by the first *game* of the course, the memory game *Contradiction Day*. In this *game*, the children have to find matching pairs of cards, each pair consisting of one picture of an animal telling a lie and one of Detective Hasel countering the lie with a fact. This is a preparatory exercise for conducting a proof by contradiction. The next activity is the interactive *proof* video *The Hasel Code*. It is very similar to the interactive video in Chapter 0, as it is again an adaptation of *The Letter Game* (Gernes, 1999; Hofstadter, 1999), but with letters instead of beads, just like in the original game. This activity gradually increases the complexity in learning how to structure a proof. After completion of this activity,

the *notebook* of Chapter 1 as well as the first *proof collection* become visible. In the *notebook*, the children can look up the rules of *The Letter Game* in case they need them. In the *proof collection*, they can finish the chapter by solving several Letter Game exercises. All of the following chapters end with this combination of *notebook* and *proof collection*, allowing children to look up words and symbols as the proofs become more formal.

3.3.6 Chapter 2

Chapter 2 is the first chapter in which proofs are presented through the lense of a mathematical topic, i.e. Boolean Logic. Table 4 illustrates the schedule for this chapter.

Table 4

Structure of Chapter 2.

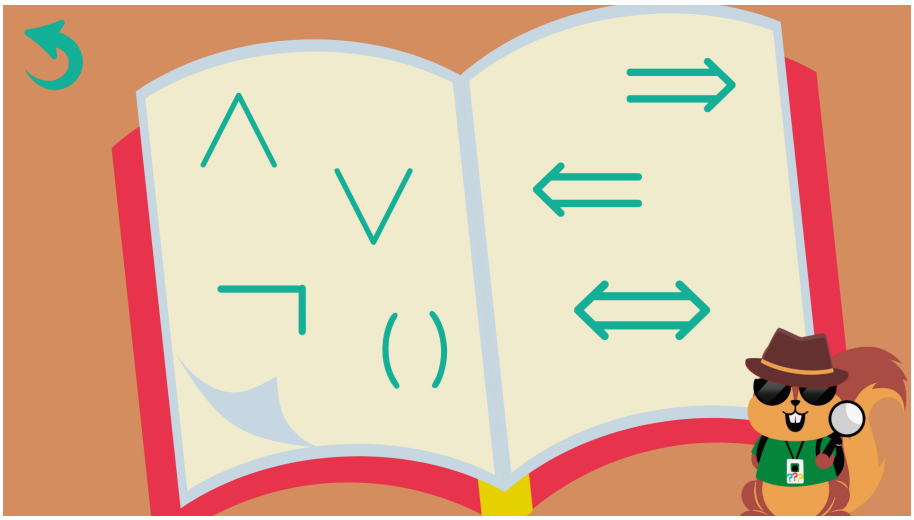
Element	Title	Objective	Implementation
Preview	Preview Chapter 2	Overview Chapter 2	H5P Video
Exploration	Secret Signs I	Familiarization with basic words and symbols of Booelan logic	H5P Interactive Presentation
Game	Secret Sentences I	Translatinh sentences into symbols	Own Game (embedded)
Game	The Truth Table	Learning to identify equivalent statements	H5P Interactive Presentation
Case	Hasel's Nut Search	Finding statements and expressing solutions with symbols	Own Game (embedded)
Proof	If and Only if	Writing first Boolean proof	H5P Interactive Video
Notebook	Notebook Chapter 2	Looking up symbols of Boolean Logic	Moodle Glossary
Proof Collection	Boolean Logic	Writing proofs in Boolean Logic	Own Game (embedded)

In the *exploration Secret Signs I*, the symbols for 'and', 'or' and 'not' as well as conclusion arrows pointing in both directions are presented (see: Figure 4). The children can click on the symbols in Hasel's secret notebook in whatever order they like. When a symbol is clicked on, Detective Hasel gives an example from one of his earlier cases and then presents a quiz question which has to be answered by picking the right symbol. When all symbols are explored, the children can check their

understanding by solving a drag-and-drop puzzle including all the symbols.

Figure 4

Screenshot of the notebook from the exploration in Chapter 2



The next activity, the *game Secret Sentences I*, allows the children to practice their newly acquired knowledge by translating written statements into formal expressions. This is followed by another *game* in which truth tables have to be completed by dragging the correct logical values into the grid. This teaches the children to recognize matching patterns of truth values among equivalent statements. *Hasel's Nut Search* is an embedded exercise adapted from Diproche in which a colorful pattern in a grid must be described by entering a Boolean expression. In the adapted version, the variables are represented by a picture of a squirrel and a nut respectively. After completing this activity, the children encounter an interactive video in which Hasel first explains the meaning of 'if and only if' using the example that he only goes swimming if his friends come along. He then shows a basic proof from Boolean Logic and asks the children to fill in the blanks in two more examples. Once they have done so, they can look up all the symbols in the *notebook* and write their own natural language proofs for the statements from the *proof collection*. While working on these exercises, they will receive written feedback on the logical correctness and coherence of their proof texts.

3.3.7 Chapter 3

Chapter 3 contains an *exploration* and a *pre-exploration* because in Set Theory, a definition can be represented by a symbol as well by a Venn diagram, and both connections are necessary for reasoning in this subject (see: Table 5).

Table 5

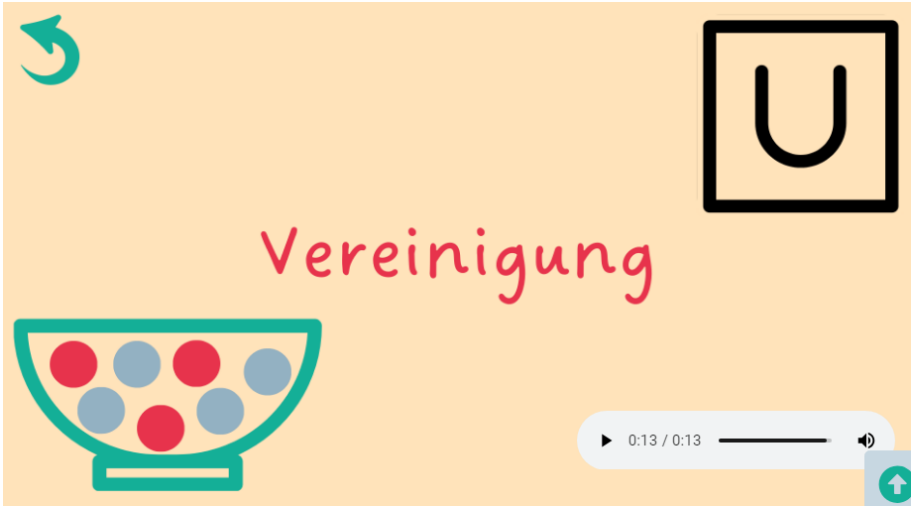
Structure of Chapter 3.

Element	**Title**	**Objective**	**Implementation**
Preview	Preview Chapter 3	Overview Chapter 3	H5P Video
Pre-exploration	Lots of Sets	Learning names and diagrams of set concepts	H5P Interactive presentation
Exploration	Secret Symbols II	Learning all the symbols for the different set concepts	H5P Interactive Presentation
Game	Secret Sentences II	Translate sentences into symbols	Own Game (embedded)
Game	Secret Set-Memory	Connect words and symbols of set concepts	H5P Memory Game
Case	Lots of Files	Connect diagrams and names for set concepts	H5P Image Pairing
Proof	Right Within the Set	Write first Set Theory proof	H5P Interactive Video
Notebook	Notebook Chapter 3	Look up symbols of Set Theory	Moodle Glossary
Proof Collection	Set Theory	Practice writing proofs in Set Theory	Own Game (embedded)

In the *pre-exploration*, the children can explore Hasel's picture wall and click on different photos showing him and his friends in different constellations. When a picture is clicked, it goes full screen and turns into a drag-and-drop exercise in which the animal friends have to be arranged into the right sets to match a given definition. The *exploration* then links these definitions to symbols, using mnemonic illustrations in which the symbol is combined with an everyday item. An example is provided in Figure 5. The children have to find these illustrations by clicking on the tiles of a mosaic.

Figure 5

Screenshot from the exploration in Chapter 3 introducing the symbol for "union"



Like its predecessor in Chapter 2, the *game Secret Sentences II* is an exercise in translating natural language statements into formal ones. It is followed by a memory game in which the learner has to match several pairs of cards by connecting symbols with their respective meanings. The *case Lots of Files* is designed similarly: It consists of a matching game with cards showing a diagram and cards showing a word. This way, the children can practice both the iconic and the symbolic meaning of the words. In the next step, the children can work through the *proof* video of Chapter 3, in which Hasel explains how to prove that two sets are equal, using an analogy about football teams. Then the children can watch and complete formal examples for basic Set Theory proofs. In the *notebook*, they can then recap their knowledge of Set Theory symbols before they move on to the *proof collection* to write several proofs from the field.

3.3.8 Chapter 4

Chapter 4 is a chapter on Number Theory. Its activities are listed in Table 6.

The *exploration* aims at enabling the children to use the mathematical operations needed for conducting Elementary Number Theory proofs: Reasoning with divisibility and number parity, forming squares of numbers, and handling variables. In this *exploration*, the children can find a game in which they have to sort even from odd numbers, a secret notebook that contains some rules for divisibility by monadic numbers, and a game to practice squaring numbers. Additionally, there are two smaller, hidden cases. One case has the children sort several numbers in order to give the result of evaluating a mathematical term with one variable. This activity is depicted in Figure 6.

Table 6

Structure of Chapter 4.

Element	Title	Objective	Implementation
Preview	Preview Chapter 4	Overview Chapter 4	H5P Video
Exploration	On the Way to the Numbers	Computing expressions with brackets, variables and squares; describe number parity	H5P Interactive Presentation
Game	Hasel Guesses Your Number	Compare sets of numbers	H5P Personality Quiz
Case	Numbers on the Run	Discuss number properties	Own Game (embedded)
Proof	The Number Machine	Write first Elementary Number Theory proof	H5P Interactive Video
Notebook	Notebook Chapter 4	Look up words and symbols from Elementary Number Theory	Moodle Glossary
Proof Collection	Number Theory	Practice writing proofs in Elementary Number Theory	Own Game (embedded)

Figure 6

Screenshot of a Case From the exploration in Chapter 4

In the other case, the children have to break a code using the divisibility rules from the notebook. Next in Chapter 4 is a *game* called *Hasel guesses your number*. To play, the user picks a number between 0 to 7 and answers three yes/no-questions about it (e.g., "Is your number one of these: 1,3,5,7? "). The *yes*-answers represent the ones from the binary system and the *no*-answers represent the zeros, so the number can be derived from the given information. Since the children most likely are not familiar with the binary system yet, they are asked to come up with an explanation for Hasel's trick. This aims at teaching them how to spot patterns and form conjectures. The following activity is the *case Numbers on the run* and presents the children with a randomly generated name for a category and three properties that a number in this category possesses (e.g., "contains the digit zero "or "looks the same when read back to front"). The goal of this exercise is to pick all matching

numbers from a list of five. This is to familiarize the children with checking mathematical definitions. In the *proof* video on Number Theory, Detective Hasel presents a machine that takes a number and, in two steps of basic arithmetic operation, turns it into another. This leads over to the first Elementary Number Theory proofs concerning the parity of numbers after certain mathematical operations. Again, there is one example to watch and two to complete. In the *notebook*, the children can look up the concepts they learned in the *exploration* to prepare themselves for conducting their own Number Theory proofs in the *proof collection*.

3.3.9 Changes implemented after Study 1

After the pilot study (see: Section 4), parts of the initial course design were revised based on the feedback by parents and children: I further simplified the user interface of the *proof collections* by switching from keyboard entries to a point-and-click system. Furthermore, little screencasts with worked examples were included before each *proof collection* to reduce the cognitive load caused by the tool. The most significant change revised the order of the course chapters: To introduce the children to natural language proofs by way of a more familiar topic and fewer new symbols, the chapter on Elementary Number Theory was moved before the chapter on Boolean Logic. In the efficacy study, which I describe in Section 6, this revised course design was rolled out. The final course version with screenshots and short descriptions of every activity is presented in the course booklet available at `https://osf.io/qf6by/`.

4 STUDY 1: DEVELOPMENT AND PILOT STUDY OF AN ASYNCHRONOUS ONLINE ENRICHMENT COURSE ON MATHEMATICAL PROVING

The content of this chapter has been accepted for publication in the International Journal of Mathematical Education in Science and Technology (https://doi.org/10.1080/0020739X.2025.2543834). The proportional contributions of the (co-)authors to the manuscript are presented in the subsequent table. This article may not exactly replicate the final version published in the journal. It is not the copy of record.

Author	Author position	Scientific ideas (%)	Data generation (%)	Analsysis & interpretation (%)	Study writing (%)
Xenia Stein	first	70	90	90	85
Katerina Tsarava	second	5	5	5	5
Armin Fabian	third	0	0	5	10
Merlin Carl	fourth	20	0	0	0
Alla Kutkina	fifth	0	5	0	0
Walther Paravicini	sixth	5	0	0	0

Abstract

Supporting the development of mathematically gifted children is a key challenge in today's educational system. Providing enrichment courses that offer extracurricular and challenging content, such as rigorous proof-based activities, is a promising strategy. However, participation in these courses is often hindered by tight schedules, long travel times, and limited availability of course spots. Asynchronous online courses have the potential to mitigate these barriers by offering flexible access. Nonetheless, these courses introduce new challenges for gifted primary school children, including the need for high levels of self-regulation and proficiency with digital devices. This study investigates the feasibility of an asynchronous online course specifically designed to foster the mathematical proof competencies of gifted primary school children. To do so, we rolled out the course within a summer program in the German federal state of Baden-Württemberg and collected open and closed feedback as well as progress data from 304 talented children (mean age = 8.95, SD = .66, 122 female). We found that children and legal guardians perceived the course as very positive. However, not all children completed the course. The need for more instruction, reduced typing, and a wider range of difficulty levels ought to be addressed in future courses.

*Keywords:*primary education; giftedness; mathematics education; mathematical proof; curriculum enrichment; asynchronous learning

4.1 Introduction

Fostering gifted children is crucial to ensure progeny and progress (Subotnik et al., 2011). Given the significant role of mathematical abilities in addressing complex challenges in the world, supporting gifted children in mathematics seems to be specifically important (see: OECD, 2019; Rebholz, 2017). A core approach in fostering gifted children is providing additional learning opportunities outside regular school hours with content that exceeds the school curriculum and deepens the knowledge in certain fields (Worrell et al., 2019). Such opportunities are commonly subsumed under the term extracurricular enrichment courses and have consistently shown to be beneficial in fostering gifted children's domain-specific performance (Herbein et al., 2018; Schiefer et al., 2021), and motivation (Rebholz, 2017). However, on-site enrichment courses do not serve the needs of all children. Costs related to material, time, and commuting can outweigh the benefits, preventing some families from accessing these opportunities (Lohaus & Wild, 2021). Furthermore, limited numbers of available spots in on-site courses make selection processes necessary, something that likely induces judgmental bias in teachers (Golle et al., 2022). To address these challenges, transitioning enrichment opportunities from on-site formats to asynchronous online courses may present a promising approach. Nonetheless, participation in asynchronous online courses can introduce new challenges for learners and teachers, such as dealing with technical issues or preserving one's attention remotely (Weaver et al., 2022). Moreover, there is a lack of rigorous studies investigating online enrichment courses for mathematically gifted children, leaving questions about their feasibility unanswered. Against this background, the present study investigates whether and how an asynchronous online enrichment course can be a feasible option to foster mathematically gifted children. Given the key role of proof in mathematics and the fact that the procedure of proving aligns well with the characteristics of mathematically gifted children (Bardy & Bardy, 2020; Boero, 1999; E. Brunner, 2014; Krutetskii, 1976), we specifically developed and implemented an enrichment course that aimed at fostering children's mathematical proving ability.

4.1.1 Fostering Gifted Children

Helping to unfold the talents of gifted children is an important duty of society, as they may become key players in driving innovation and social progress (Renzulli, 2016; Subotnik et al., 2011; Tannenbaum, 1983). Therefore, these children should be offered adequate and affordable opportunities to further develop their talent (Lohaus & Wild, 2021). But how can we best support gifted children in developing their talent? To date, several models have been proposed to conceptualize talent and its development. While some models emphasize a single general factor

(Terman, 1922), others recognize multiple domains of talent, such as music, mathematics, and spatial ability (Gardner, 1983). These discussions have led to integrative models that consider both individual dispositions and environmental factors as crucial for the development of exceptional performance (Dai, Moon, & Feldhusen, 1998; Feldman & Goldsmith, 1986; Renzulli, 2016). In our study, we follow the 'Talent Development Mega Model' (TDMM) (Subotnik et al., 2011), which describes individual cognitive abilities as the basis of talent and how these develop over time. The personal environment and own psychological strengths are mentioned as key factors that ideally lead to eminence and own contributions to the domain (Subotnik et al., 2011). The TDMM features three stages as proposed by (Renzulli, 2016). For very young children (first stage), a teacher should provide a gifted student with possibilities to explore several (new) fields (*teaching for falling in love*), later (second stage) a teacher should support the student in gaining expertise in their domain of interest and at last (third stage) a teacher needs to encourage innovative contributions to that domain (Subotnik et al., 2011). Gifted primary school children can be located in the first and partly in the second stage of this model, as they are in an early educational state, but have already encountered some subjects and domains. Therefore, enrichment courses should provide them with insights into new fields, as well as the possibility to acquire skills in these fields. In the past, such extracurricular enrichment courses have, in many cases, proved beneficial for gifted primary school children (Herbein et al., 2018; Schiefer et al., 2021; Trautwein et al., 2023).

4.1.2 Challenges for On-Site Enrichment Settings

While the significance of fostering gifted children is indisputable, implementing effective enrichment programs presents several challenges, particularly within traditional on-site settings.

First, to participate in an enrichment course, the perceived benefits need to outweigh the costs (Lohaus & Wild, 2021). Depending on the individual situation, these costs can be very challenging for families of gifted children (e.g., if a child lives in a rural area and needs to be driven to the course site or when the necessary learning materials are expensive).

Second, the limited number of spots in on-site courses makes selection procedures necessary. Therefore, children are often nominated by teachers to qualify for enrichment programs. This can lead to biases against girls and children with a low socio-economic status (SES) or with an immigrant background (Golle et al., 2022). Large-scale ability testing could help for a fair distribution of opportunities, but is harder to administer and more expensive than teacher nomination (Golle et al., 2022). The Advanced Academics Model, however, contradicts the practice of selection fundamentally by stating that all those who do well in a domain should receive enrichment exercises, to ensure that everybody is fostered according to their needs (Peters, Matthews, Mcbee, & Mccoach, 2021).

This is, in many cases, not possible due to the limited capacities mentioned above.

Lastly, the schedules of children as young as primary school age are often already tight due to numerous scholarly extracurricular activities (Fölling–Albers, 2000). Thus, participating in a synchronous enrichment course could mean that they have to give up another activity.

Considering these challenges, an asynchronous online course could serve a broad share of children either as a pull-out program during the school day or as an additional enrichment after school, without costs for extra commuting or tuition fees for individual enrichment. Additionally, digital formats can meet individual needs better than traditional ones (Lohaus & Wild, 2021), and the opportunity to set one's own pace in a learning process may be further beneficial (Tullis & Benjamin, 2011). This is particularly relevant for gifted children, as prior research demonstrated that the ability to work independently develops earlier in (mathematically) talented children compared to their peers (Fuchs, 2006).

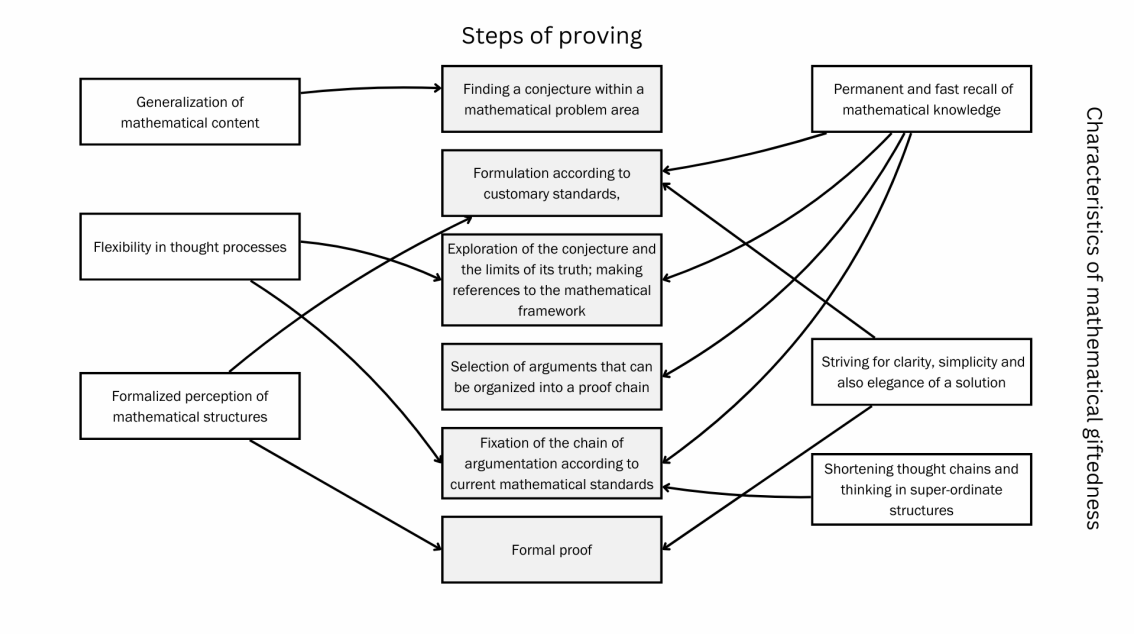
4.1.3 Proving as a New Field of Exploration for Gifted Students

To foster gifted children, not only in mathematics, it is important to choose appropriate content and activities. That is, content that is new and challenging but at the same time not overwhelming (Rebholz, 2017). This demand brought our attention to mathematical proving. "A proof is a conclusive argument that a proposed result follows from accepted theory."(Hersh, 1997, p. 6). As all mathematical knowledge over time is constructed and validated through such arguments, proving has often been referred to as the *heart of mathematics* (Bass, 2011). Still, when students come across rigorous proving for the first time in university, they perceive it as particularly challenging (Carl, 2022; Glosauer, 2019). This is probably why proving has been considered too difficult as a topic to teach in compulsory education (Bass, 2011). Hence, in mathematical enrichment programs, educators often prefer activities on problem-solving (Käpnick, 1998) or inductive thinking (Ko & Song, 2011) to formal mathematical proof learning. The process of problem solving, however, explicitly requires a proof of every single step of a proposed solution, as described in Polyá's prominent model of problem solving (Pólya, 2010). Therefore, it would be a logical consequence to integrate proving as well when problem-solving activities are implemented in enrichment courses.

In 1976, Krutetskii published a framework of characteristics of mathematically gifted children (e.g., flexibility in thought processes, formalized perception). We compared this model to Boero's model from 1999, which illustrates the steps of proving, as we depict in Figure 7, and we observed many ways in which these characteristics could be beneficial for the process of proving:

Figure 7

Characteristics of mathematically talented children linked to the steps of proving (own graphic based on Boero (1999) and Krutetskii (1976))



The ability to *generalize mathematical content* is likely to be an inspiration for *coming up with a conjecture*, which is the first step of a proof. In the second step of Boero's model, the *formulation of the conjecture*, it can be beneficial to *recall mathematical knowledge fast*, to *formalize*, and to *strive for clarity*. For the third step, in which the *conjecture is explored*, again, *mathematical knowledge* is needed, but also *flexibility in thought processes*, so that the conjecture can be considered from every angle. In step number four, when the *arguments are selected*, *mathematical knowledge* needs to be recalled as well. The same applies to the next step when the *arguments are arranged into a chain according to mathematical standards*. This also requires *flexibility in thought processes* to compare different possibilities and *shortening thought chains* to avoid redundancy in the arguments. When finally the *formal proof* is written, *striving for clarity* as well as *formalized perception* are particularly useful. Together, it can be assumed that mathematically gifted children do better than their peers when it comes to proving, while also enjoying the process of proving, as it fits their cognitive need for new and challenging content.

In the curriculum for mathematics at primary schools in Baden-Württemberg (Germany), mathe"-matical proving is not listed as a competence, but mathematical reasoning is (Ministerium für Kultus, 2016a). The curriculum for the academic secondary school includes proofs, but only of geometrical and not algebraic conjectures (Ministerium für Kultus, 2016b). This tendency to keep formalism away from (primary) school children is in line with Piaget's model of cognitive

development, according to which formal perception develops in early adulthood (Piaget & Inhelder, 1977).

On the other hand, studies with gifted primary school children suggest that this target group can develop deductive skills and formal reasoning much earlier than their peers (Bardy & Bardy, 2020; Ko & Song, 2011). German secondary school students do not show high mathematical proof competencies but tend to have a positive attitude toward proving (Bass, 2011). This lack of competency is likely related to the absence of mathematical proof in primary school (Dreyfus, 1999). Therefore, it could be worthwhile to teach proving earlier so that children might develop both a positive attitude and the desired competencies.

Furthermore, the idea of algebraic proving for primary school children is not new: In the 1960s, the New Mathematics movement called for the integration of algebraic proof (e.g., in Set Theory) in primary school (Bass, 2011). The main hurdle for the successful implementation of this curriculum was the lack of formal knowledge among teachers and legal guardians (Kline, 1973). Again, asynchronous online courses may help to overcome this gap as these courses yield plenty of possibilities for individual support, such as automated feedback and self-paced instruction videos.

Because gifted learners are particularly attracted by reasoning tasks and the use of logical language (Bock & Borneleit, 2000), it seems reasonable to develop an enrichment course on mathematical proving. In most courses for novices, Boolean Logic, Set Theory, and Elementary Number Theory form the content to introduce mathematical proof techniques (Carl, 2022; Glosauer, 2019). As none of these topics is part of the current curriculum in both primary and secondary school in the state (Ministerium für Kultus, 2016a, 2016b), they seem particularly suitable for an extracurricular enrichment course. This is why we developed the course *Logical Detectives* based on these considerations in line with current evidence-based practices.

4.2 The Present Study

The goal of this exploratory study was to investigate whether an asynchronous online course on mathematical proving is a feasible strategy to foster gifted primary school children's competencies. To do so, we developed and implemented the asynchronous online course *Logical Detectives* (see: subsection 4.3.3). Our main research question was:

RQ: Is an asynchronous online mathematical proof course a feasible option to fostering gifted children?

To answer this question, we piloted the course *Logical Detectives* and formulated three more

specific research questions:

RQ1: How do gifted primary school children perceive specific aspects of the asynchronous online mathematical proof course?

RQ2: How do legal guardians of gifted primary school children perceive the asynchronous online mathematical proof course?

RQ3: How far did participants progress in the course and are there systematic dropout patterns?

4.3 Materials and Methods

4.3.1 Participants

Participants were recruited within the Hector Children's Academy Program (HCAP), an extracurricular program for gifted primary school children in the German federal state of Baden-Württemberg for which the most promising decile of each class is nominated by their teachers (Trautwein et al., 2023). Since 2020, the HCAP has offered online courses via the Learning Management System (LMS) Moodle for more than 15.000 children (as of September 2024). Children from all of the 69 HCAP locations in Baden-Württemberg could take part, as the entire study was carried out online. The course was designed for primary school children from Year 3 to 4, typically children between 8 and 10 years old. As the course took place during the summer holidays, we decided to address the target group of upcoming third and fourth-graders. Thus, all children registered on the HCAP Moodle platform were invited to take part if they were in Year 2 or 3 at the time of the study. They were informed that the study was voluntary and they could withdraw from it at any time with no disadvantages. The final sample included 304 children (mean age = 8.95, SD = .66, 122 female).

Written consent was obtained from both children and their legal guardians prior to data collection. The study was approved by the ethics committee of the faculty of economics and social sciences at the University of Tübingen (file number A2.5.4-296_vb). The statewide coordination of the HCAP granted permission for this online study.

4.3.2 Design and procedure

To answer the research questions mentioned above, a pilot study with three different measurements was designed (see: 4.3.4): Children's feedback during and after the course (*RQ1*), feedback from their legal guardians after the course (*RQ2*), as well as log-data (*RQ3*). In Table 7, we provide an

overview of this data collection procedure, and after that, describe it in more detail.

Table 7

Data Collection Procedure.

Recruitment & Informed Consent	Log-data Collection					Post Course Feedback (Legal guardians & Children)
	CF 1	CF 2	CF 3	CF 4	CF 5	
	Course Participation					
	6 weeks					

Note. CF = Chapter Feedback (individual time points)

The registered participants could access the course, which consists of five chapters (see: subsection 4.3.3), at any time they wanted within the six-week time period of summer holidays in the German state of Baden-Württemberg. After the end of this period, we asked children and legal guardians to provide feedback on the course design, in both open and closed questions (Post Course Feedback). Additionally, short questionnaires were displayed to the children after each course chapter (Chapter Feedback). In these, the children could rate the course elements and content that they had just worked through. For the short questionnaires, we used a built-in function of the Moodle LMS. For the feedback after the course (see: subsection 4.3.4), we used the online survey system Unipark. Throughout the course, we collected log data via the built-in dashboard of the Moodle platform.

4.3.3 The Course "Logical Detectives"

The Course's Core Components In the development of this course, we followed the procedure proposed by (Nelson et al., 2012) and defined several core components from which the activities of the intervention were then derived. The first two of these core components target the meta-cognitive level and are meant to foster the children's engagement with the online materials, while the other two form the basis of how the mathematical content is delivered in the course. We curated the following core components:

Self-paced learning is a format in which learners allocate their study time to the exercises themselves (Tullis & Benjamin, 2011). This format can be beneficial in online courses and increase performance, especially on difficult tasks (Tullis & Benjamin, 2011).

Automated real-time feedback is an immediate response provided by a computer, which would not be possible for a human to deliver as quickly J. Schneider et al. (2016). Receiving feedback that is instant and task-specific greatly increases its utility for the learner (Mory, 2004). Thus, automated real-time feedback is likely to be beneficial for students in online courses. It can help to improve

the learners' motivation to practice, their ability to review their performance, their self-concept, and their post-training performance (J. Schneider et al., 2016).

Iconic and symbolic logical reasoning are forms of mathematical reasoning that correspond to stages of Bruner's model of representations. The so-called EIS-model describes how cognitive development depends on the interplay of enactive (i.e., touchable) over iconic (i.e., graphical) to sunderline (i.e, formal) representations (Bruner, Olver, & Greenfield, 1971). In the didactics of mathematics, this model is widely used to design instructional material to reach both a thorough understanding and confident use of mathematical symbols (Lambert, 2011). Transitioning from iconic to symbolic logical reasoning can help to develop mathematical proof skills in Boolean Logic and Set Theory, two disciplines that both have their own mathematical symbols (Bronkhorst et al., 2021).

Natural language proof writing is the practice of writing a mathematical proof in a controlled fragment of the language in which people usually read and write. In contrast to formal mathematical language, it is much easier for human readers to make sense of such a proof-text. At the same time, machines can still verify natural language proofs, as these only consist of a limited number of possible words (Carl et al., 2022). In teaching mathematical proof to novices, Carl (2022) showed that natural language proof writing in a drill-and-practice environment can help develop mathematical proof skills in Boolean Logic, Set Theory, and Elementary Number Theory.

Structure of the Course To transform these core components into activities, we designed five chapters with different objectives:

- Chapter 0: Getting to know the technical basics of the course
- Chapter 1: Getting to know the structure of a mathematical proof
- Chapter 2: Learning the basic symbols and finding first proofs in Boolean Logic
- Chapter 3: Learning the basic symbols and finding first proofs in Set Theory
- Chapter 4: Finding first proofs in Elementary Number Theory

Each chapter contains four main elements: A *preview video*, an *exploration*, a *case*, and a *proof*. Additionally, chapters 1-4 contain a *notebook*, a *proof collection*, and one or two *games* each. In the overview of the course content, which is publicly available at `https://osf.io/qf6by/`, we provide an overview of all course elements. For each element, we provide the theoretical background, implementation format, and a screenshot from the respective course element. Additionally, we illustrate which of the aforementioned core components of the course mainly correspond to an

element. To analyze how the target group perceived these elements, we curated feedback measures for legal guardians and children, which will be described in the following subsection.

4.3.4 Measures

Chapter Feedback – Children's Perception Our first research question (*RQ1*) targeted the children's perception of the course elements. Therefore, we asked them to answer a set of 15 questions after each chapter. The questionnaire contained six Likert-scaled items on the learning atmosphere in the online course (e.g., *I was able to work calmly and focused.* (Rebholz, 2017)). Additionally, we included nine items that assessed how much the children liked every single kind of course element (e.g., *Watching a Video, Playing a Game* or *Receiving Feedback*). In chapters 2 and 3 of the course, we added an additional item asking how much the children liked the use of mathematical symbols, as these symbols only appear in the two chapters.

Post Course Feedback – Children's and Legal Guardians' Perception To gain more insights into how the children (*RQ1*) and the legal guardians (*RQ2*) perceived the asynchronous online course, we prepared a feedback questionnaire for each group. The exact item wordings can be found in the appendix (Tables 3 and 4).

After the end of the course, we sent out the post-course questionnaire to all children, and also to those who did not finish the course. They were asked to indicate which course chapters they completed and in which they had either technical or logical support from another person. They could give course feedback by answering ten Likert items (four scale points) on their perception of the course (e.g.,*I had fun during the course.*) as well as in two open questions (*What did you like best about the course? / What was not great about the course?*).

The legal guardians answered a similar questionnaire with eleven items to report how they perceived the course (e.g. *My child had fun during the course.*) as well as three open questions (*What would you change about the course from an organizational/technical/content perspective?*).

Course Log-Data – Children's Progress The last research question addressed the users' progress in the course (*RQ3*). We used the Moodle feature *activity completion* to extract log data from the course. For each user, we harvested the information on which activity they finished at what time. This also allowed us to determine who accomplished which share of the course within the six weeks it was available. This data was automatically stored in the backend of the Moodle platform with the users' names as an identifier. These identifiers had to be deleted before the analyses. Thus, it was not possible to match the log data with the questionnaires, which were submitted under

secret personal codes as demanded by the ethics committee of the faculty.

4.3.5 Analysis

Chapter Feedback – Children's Perception The data from the short questionnaires with the children's feedback on each chapter were downloaded and anonymized. We then recoded the answers to a numeric scale ranging from 0 = *did not like at all* to 3 = *liked very much*. For each variable and chapter, we calculated the mean over the respective sample. We then plotted the progression of these means over the chapters (see: subsection4.4).

Post Course Feedback – Children's and Legal Guardians' Perception For each closed feedback question, we calculated the overall mean and standard deviation for the whole sample. Additionally, we computed inter-item correlations within both the children's and the legal guardians' questionnaires as well as correlations between the items of both scales.

With the written feedback from legal guardians and children, we followed the procedure of qualitative content analyses (Mayring, 2022): First, we went through the feedback on each question and identified categories of reappearing topics. In the second step, we coded each statement of feedback into one of these categories and analyzed the frequency of the different aspects among legal guardians and children. Detailed insights into this procedure are provided at `https://osf.io/p6rwu/`.

Course log-data – Children's Progress We exported the log data from the teacher dashboard in the Moodle course and instantly replaced the usernames with random identifiers. We determined for each activity how many users completed it and then calculated the difference to the succeeding activity to see how many participants stopped the course at that point. We then compared these numbers for activities from different chapters and of different types.

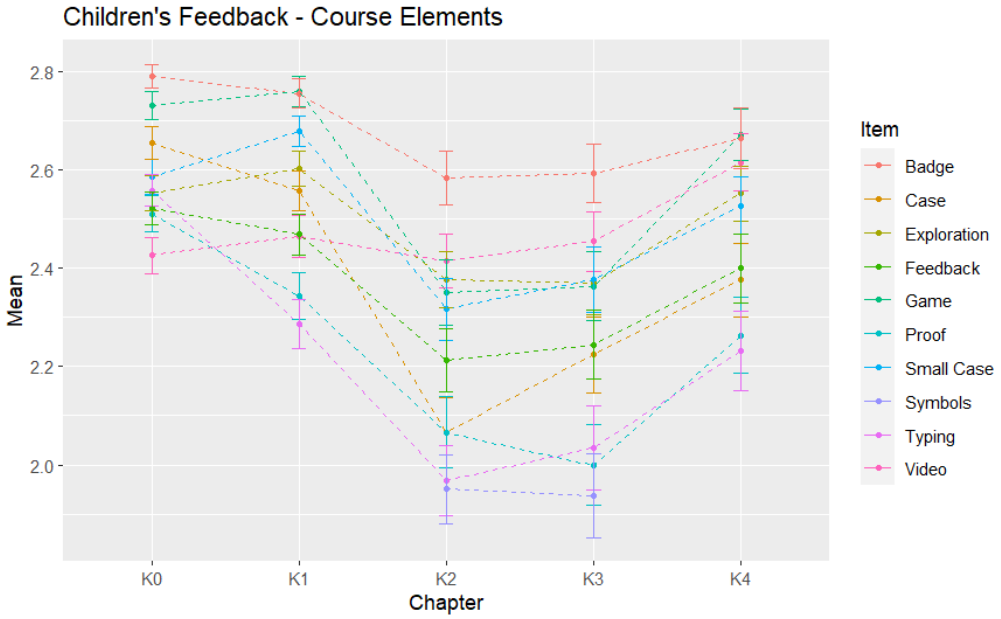
4.4 Results

4.4.1 Chapter Feedback from the Children

In the questionnaires that were displayed to the children after each finished chapter, they could indicate how much they liked different course aspects. Figure 8 shows the mean rating for each course element in each chapter

Figure 8

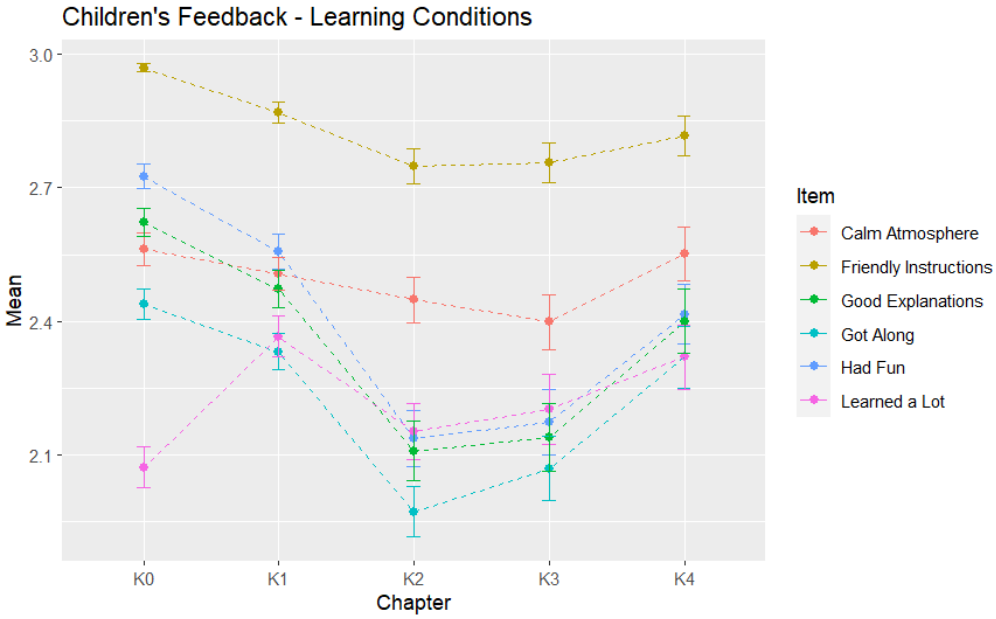
Chapter Feedback: Children's rating of each course element from 0=Not at all to 3=Very much (means per chapter)



When focusing on each chapter as a whole, we can see that the variance per item is the lowest in K0, and also the means of the different course elements only range between the very positive ratings of 2.4 and 2.8. Throughout the chapters, both the intervals of variance and the distance between the different means increase. Remarkably, all categories have a drop in popularity in chapters K2 (Boolean Logic) and K3 (Set Theory). The course element *Symbols* (use of logical symbols) only appears in chapters K2 and K3 and was therefore only mentioned in the questionnaires of these. In both cases, we can see that it was the least popular course element. The course elements *Typing* and *Proof* were among the least popular chapters in each chapter but the introduction chapter (K0). In contrast, receiving *Badges* for accomplished tasks was the most popular course element in all five chapters.

Figure 9

Chapter Feedback: Children's rating of the learning conditions from 0=Do not agree at all to 3=Fully agree (means per chapter)



In the same questionnaires, the children were able to rate different aspects of the learning atmosphere of each chapter. As displayed in Figure 9, on average, the children agreed or fully agreed with the statement that the instructions were friendly in all chapters. Again, we can see a slight decrease in most means from K0 to K1, a bigger drop from K1 to K2, a slight increase to K3, and a bigger increase to K4. In this last chapter, the means are on a similar level as in K1 again. Exceptions from this pattern can be seen for the category *Calm Atmosphere*, which does not vary that much throughout the chapter, as well as for the category *Learned a Lot*, which has a quite low mean in the introductory chapter K0.

4.4.2 Post Course Feedback (Closed Questions)

Both children and legal guardians answered several four-step Likert items addressing their perception of the course quality. Additionally, the children indicated in a binary variable for each chapter if they had help from an adult with handling the computer / solving the riddles ('1') or not ('0'). The following tables show the wording of these items as well as the descriptive statistics for the closed feedback in the post-course questionnaire. After the descriptive statistics, we will present the inter-item correlations and describe all observations with a significance at the p=0.001 level.

Table 8

Children's Feedback Data - Descriptive Statistics

Variable	Description	*N*	*M*	*SD*
K2HT01	technical help in Chapter 0	114	0,21	0,41
K2HT02	technical help in Chapter 1	114	0,36	0,48
K2HT03	technical help in Chapter 2	114	0,52	0,50
K2HT04	technical help in Chapter 3	114	0,49	0,50
K2HT05	technical help in Chapter 4	114	0,41	0,49
K2HT06	technical help in no Chapter	114	0,12	0,33
K2HR01	help with riddles in Chapter 0	114	0,29	0,46
K2HR02	help with riddles in Chapter 1	114	0,28	0,45
K2HR03	help with riddles in Chapter 2	114	0,36	0,48
K2HR04	help with riddles in Chapter 3	114	0,32	0,47
K2HR05	help with riddles in Chapter 4	114	0,26	0,44
K2HR06	help with riddles in no Chapter	114	0,42	0,50
K2CQ01	The online course was fun.	114	3,25	0,80
K2CQ02	I learned a lot in the online course.	114	3,30	0,83
K2CQ03	I'm glad I participated in the course.	114	3,31	0,86
K2CQ04	I did not manage all technical things on my own.	114	2,82	1,02
K2CQ05	I did not manage all riddles on my own.	114	2,80	1,06
K2CQ06	It was hard for me to read and understand everything.	114	2,19	1,05
K2CQ07	The exercises in the course were too hard.	114	2,18	0,94
K2CQ08	I would have liked to interact with the other children.	114	2,18	1,07
K2CQ09	The online course made me tired.	114	1,82	1,01
K2CQ10	The online course bored me.	114	1,47	0,73

Note. Items of categories K2HR and K2HT were answered on a binary scale (*1 = yes, 0 = no*); Items of category K2CQ were answered on a four-step Likert scale (*1 = Do not agree at all* to *4 = Fully agree).*

Feedback From the Children The share of children who had technical help increases from Chapter 0 to Chapter 1 and is the highest in Chapter 2, in which more than half of the children declare that they had help. For the last two chapters, there is a slight decrease. The same pattern can be observed for the items regarding help with the riddles. Still, there is a remarkable difference between the two categories: While the mean for *technical help in no chapter* is at .12, the mean for *help with the riddles in no chapter* is at .42 .

For the items indicating that the children had fun, learned a lot, or were glad they participated, the means were between 3 (agree) and 4 (fully agree). The questionnaire also featured two more items regarding technical and logical help. In both cases, the means were around equally high (M = 2.82 vs. M = 2.80), indicating that on average the children agreed that they did not manage everything on their own. On the other hand, the means for the children stating that the exercises were hard to comprehend, were too difficult, or that social interaction was missing, were lower than 2.5. This implies that the average child disagreed. A stronger disagreement could be observed for the statement that the course was tiring (M = 1.82) or boring (M = 1.47). Still, for all items, there were relatively high standard deviations ranging in the dimension of a whole unit. This suggests a rather big heterogeneity in perception.

Table 9

Parental Feedback Data - Descriptive Statistics

Variable	Description	*N*	*M*	*SD*
E2EE01	My/our child had fun during the online course.	114	3,29	0,66
E2EE02	My/our child learned a lot in the online course.	114	3,16	0,74
E2EE03	My/our child talked about the course content at home.	114	2,59	0,90
E2EE04	I'm glad that my/our child participated in the course.	114	3,17	0,84
E2EE05	I helped my/our child with technical issues.	114	3,04	1,04
E2EE06	I helped my/our child with the riddles.	114	2,11	1,05
E2EE07	My/our child had difficulties with the technique.	114	2,39	0,95
E2EE08	My/our child had difficulties in understanding the riddles.	114	2,11	1,03
E2EE09	My/our child was overwhelmed with the exercises in the course.	114	2,54	0,90
E2EE10	My/our child felt the need for interaction with other children.	114	1,74	0,98
E2EE11	My/our child spent too much time on the computer due to the course.	114	1,72	0,89

Among the legal guardians, there was also much agreement that their child had fun, learned a lot, and they were glad their child participated. To learn more about the legal guardians' perception, we included another positive item saying *My child talked about the course content at home.* Here,

the mean (m = 2.59) implied no overall tendency regarding agreement and disagreement. For the statement that they helped with technical issues, the legal guardians mostly agreed. Interestingly, the mean of the legal guardians here was higher than the mean of the children indicating that they received help. When it asked if they helped with the riddles, the legal guardians disagreed. In this case, the mean is lower than for the respective item from the children's questionnaire. Two similar items asked if the legal guardians think their child had difficulties with the technique or the riddles. For the item on technical difficulties, there was less agreement than for the one on technical help, while for difficulties with the riddles and help with the riddles, the means were the same. When comparing the legal guardian item *My child was overwhelmed with the exercises in the course.* to the child item *The exercises in the course were too hard.*, there is more agreement for the legal guardian item ($M = 2.54$). Regarding the need for interaction, the legal guardians disagreed even more strongly than the children. The last item from the legal guardian questionnaire was *My child spent too much time on the computer due to the course.*, which was on average disagreed.

Table 10

Inter-Item-Correlation: Course Quality Participant Questionnaire

Variable	1	2	3	4	5	6	7	8	9
1. K2CQ01									
2. K2CQ02	.63***								
3. K2CQ03	.57***	.71***							
4. K2CQ04	-.07	-.22*	-.13						
5.K2CQ05	-.16	-.20*	-.16	.38***					
6. K2CQ06	-.33***	-.44***	-.29**	.36***	.57***				
7.K2CQ07	-.34***	-.44***	-.40***	.27**	.59***	.68***			
8.K2CQ08	-.00	-.04	.07	-.00	.13	.25**	.23*		
9. K2CQ09	-.28**	-.31***	-.28**	.25**	.21*	.51***	.49***	.28**	
10. K2CQ10	-.38***	-.35***	-.39***	.19*	.11	.28**	.20*	.17	.47***

Note. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

First, we will list the results from the child questionnaire on course quality (see: Table 10): There was a highly significant positive correlation among the items K2CQ01-K2CQ03 representing the statements *I had fun / I learned a lot / I'm glad I participated [in the course].* A significant negative correlation was found between each of K2CQ01 and K2CQ02 and the two items K2CQ06 and K2CQ07 (*It was hard for me to read and understand. / The exercises were too hard.*). The latter one also showed a significant negative correlation with K2CQ03.

The two variables indicating a technical (K2CQ04) or cognitive (K2CQ05) overload were significantly correlated with K2CQ06 in a positive direction. For K2CQ05 and K2CQ06, the same relation could be observed regarding K2CQ07. Item K2CQ09 indicates the extent to which the children felt tired due to the course. It correlated significantly positively with K2CQ06 and K2CQ07 and negatively with K2CQ02 (*learned a lot*). The children's perceived boredom (K2CQ10) correlated negatively with the first three items (*I had fun / I learned a lot / I'm glad I participated [in the course]*) and positively with K2CQ09 (*made me tired*). One item showed no significant correlation with any of the others: *I would have preferred to collaborate with other course participants.* (K2CQ08)

Table 11

Inter-Item-Correlation: Course Quality Parent Questionnaire

Variable	1	2	3	4	5	6	7	8	9	10
1. E2EE01										
2. E2EE02	.69***									
3. E2EE03	.26**	.38***								
4. E2EE04	.68***	.64***	.34***							
5.E2EE05	-.13	.00	.06	-.16						
6.E2EE06	-.17	-.11	.02	-.19*	.61***					
7.E2EE07	-.35***	-.31***	.04	-.37***	.35***	.50***				
8.E2EE08	-.47***	-.28**	.01	-.42***	.33***	.42***	.77***			
9.E2EE09	-.19*	-.17	.08	-.28**	.41***	.49***	.69***	.65***		
10.E2EE10	-.13	-.03	-.00	-.08	-.01	.20*	.16	.15	.05	
11. E2EE11	-.09	-.05	.12	-.04	.08	.20*	.20*	.30**	.27**	.17

Note. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Feedback From the legal guardians The legal guardian questionnaire containing similar items from the external perspective showed the following internal correlations (see: Table 11): The four positively formulated items *My child had fun. / My child learned a lot. / My child talked about the course content. / I'm glad about my child's participation.* were significantly positively correlated with almost all of each other. The first and the fourth item were significantly negatively correlated with technical difficulties) and cognitive difficulties. Parental help with technology, help from an adult with the riddles, technical difficulties, cognitive difficulties, and cognitive overload all showed a (highly) significant correlation with each other.

Table 12

Correlation: Course Quality Participant Questionnaire vs. Course Quality Parent Questionnaire

	K2CQ01	K2CQ02	K2CQ03	K2CQ04	K2CQ05	K2CQ06	K2CQ07	K2CQ08	K2CQ09	K2CQ10
E2EE01	.55***	.60***	.56***	-.24*	-.23*	-.45***	-.51***	-.25**	-.48***	-.51***
E2EE02	.46***	.60***	.66***	-.17	-.10	-.30**	-.40***	-.08	-.38***	-.39***
E2EE03	.02	.08	.24**	-.08	-.09	-.06	-.12	-.11	-.14	-.25**
E2EE04	.40***	.60***	.62***	-.25**	-.29**	-.40***	-.55***	-.11	-.37***	-.35***
E2EE05	-.13	-.15	-.06	.42***	.25**	.36***	.36***	-.03	.13	.04
E2EE06	-.15	-.17	-.06	.41***	.32***	.42***	.36***	.13	.28**	.15
E2EE07	-.24*	-.35***	-.28**	.22*	.43***	.62***	.61***	.15	.44***	.19*
E2EE08	-.27**	-.40***	-.32***	.19*	.44***	.49***	.67***	.18	.44***	.22*
E2EE09	-.16	-.26**	-.17	.34***	.45***	.49***	.51***	-.00	.23*	.08
E2EE10	-.16	-.06	.08	-.01	-.03	.06	.06	.51***	.13	.08
E2EE11	-.15	-.14	-.19*	.10	.06	.12	.12	.10	.17	.03

Note. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

When we focus on the correlations between the children's answers and their legal guardians' feedback, the following can be observed: The children's statements that they had fun, learned a lot, and were glad about their participation were significantly correlated to the respective statements from their legal guardians. Furthermore, the children's statement that they learned a lot was negatively correlated with legal guardians indicating difficulties with the computer or the exercises.

The legal guardian stating that their child learned a lot was significantly negatively correlated with the child perceiving the exercises as too hard or the course as boring. Other significant correlations were found between help from an adult with technology, while help from an adult with the riddles correlated positively with the most negatively formulated children's statements. Legal guardians indicating that their child had difficulties with the computer correlated positively with most of these as well, and negatively with the child thinking that they learned a lot. Furthermore, there was a significant positive correlation between legal guardians and children stating that the child missed the interaction with others.

Table 13

Inter-Item-Correlation: Help with Riddles Participant Questionnaire

Variable	1	2	3	4	5
1. K2HR01					
2. K2HR02	.68***				
3. K2HR03	.37***	.47***			
4. K2HR04	.15	.21*	.55***		
5. K2HR05	.15	.16	.42***	.71***	
6. K2HR06	-.54***	-.53***	-.64***	-.58***	-.51***

Note. ***$p < .001$, **$p < .01$, *$p < .05$

In the post-course questionnaire, the children could indicate if they received any help with the riddles in Chapters 0 to 4. These variables show significant positive correlations with each other for most of the chapters. The control variable, indicating no help at all, correlated significantly negative with all others.

Table 14

Inter-Item-Correlation: Technical Help Participant Questionnaire

Variable	1	2	3	4	5
1. K2HT01					
2. K2HT02	.46***				
3. K2HT03	.24*	.21*			
4. K2HT04	.10	.03	.28**		
5. K2HT05	.09	.00	.13	.53***	
6. K2HT06	-.19*	-.28**	-.39***	-.37***	-.31***

Note. ***$p < .001$, **$p < .01$, *$p < .05$

Similarly, the children could state in which of the chapters they had technical help. Here we found a significant correlation only between Chapters 0 and 1 and between Chapters 3 and 4. The control variable indicating no technical help showed a negative direction in all correlations, while here, only some of the correlations were significant.

Table 15

Correlation: Technical Help Participant Questionnaire vs. Help With Riddles Participant Questionnaire

	K2HR01	K2HR02	K2HR03	K2HR04	K2HR05	K2HR06
K2HT01	.43***	.25**	.15	.07	.08	-.18
K2HT02	.29**	.35***	.12	-.08	.01	-.16
K2HT03	.23*	.25**	.54***	.24**	.22*	-.28**
K2HT04	.11	.01	.36***	.58***	.45***	-.34***
K2HT05	.05	-.09	.11	.35***	.59***	-.25**
K2HT06	-.18	-.17	-.22*	-.20*	-.16	.38***

Note. ***$p < .001$, **$p < .01$, *$p < .05$

When computing the correlations of the variables on technical help with those on help with riddles, we found significant correlations between both kinds of help in each chapter, as well as between the two control variables. Cross-subsectional significant correlations could be found for the two kinds of help in Chapters 3 and 4. Technical help in Chapter 3 also correlated positively with help with the riddles in Chapter 2, as well as significantly negative with no help with the riddles in any chapter.

Table 16

Correlation: Technical Help and Help With Riddles Participant Questionnaire vs. Course Quality Participant Questionnaire

	K2CQ01	K2CQ02	K2CQ03	K2CQ04	K2CQ05	K2CQ06	K2CQ07	K2CQ08	K2CQ09	K2CQ10
K2HT01	-.02	.07	.04	-.03	.10	.19*	.16	.11	.14	.14
K2HT02	-.16	-.05	-.14	.06	.11	.25**	.25**	.04	.12	.17
K2HT03	-.16	-.27**	-.19*	.22*	.33***	.26**	.39***	-.06	.12	.10
K2HT04	.01	.03	-.02	.30**	.37***	.29**	.36***	-.01	.13	.04
K2HT05	.01	.09	.07	.06	.13	.13	.11	-.04	.10	-.01
K2HT06	.05	.16	.12	-.22*	-.26**	-.30**	-.33***	.04	-.17	-.06
K2HR01	-.07	-.11	-.05	.27**	.09	.21*	.19*	.11	.17	.14
K2HR02	-.02	-.15	-.09	.34***	.12	.16	.20*	-.03	.06	.13
K2HR03	-.05	-.25**	-.14	.33***	.28**	.37***	.37***	-.06	.25**	.17
K2HR04	-.02	-.09	-.02	.24*	.27**	.32***	.30**	.10	.20*	.08
K2HR05	.02	.03	.04	.09	.19*	.23*	.19*	.07	.07	.02
K2HR06	.09	.17	.13	-.34***	-.28**	-.38***	-.35***	.00	-.23*	-.07

Note. ***$p < .001$, **$p < .01$, *$p < .05$

There is no significant correlation with any of the variables regarding technical help; instead, the

item correlates positively with help with the riddles in Chapters 1 and 2 and negatively with no help in any chapter's riddles. On the other hand, *I didn't solve all riddles on my own.* correlated significantly positively with the variables for technical help in Chapters 2 and 3. Difficulties in understanding and reading showed a positive correlation with help with the riddles in Chapters 2 and 3, and a negative correlation with no help with any riddles. Perceiving the exercises as difficult was positively correlated to receiving help in some chapters and negatively correlated to the two variables indicating no help at all.

4.4.3 Post Course Feedback (Open questions)

Feedback From the Children In the written feedback given with the post-course questionnaire, the children were asked to name what they liked best about the course and what they liked least. Some mentioned more than one thing, which led to more data points than participants. The quantitative content analysis yielded 16 categories for preferred course aspects (see: Table 17) and 16 for the suboptimal ones (see: Table 18).

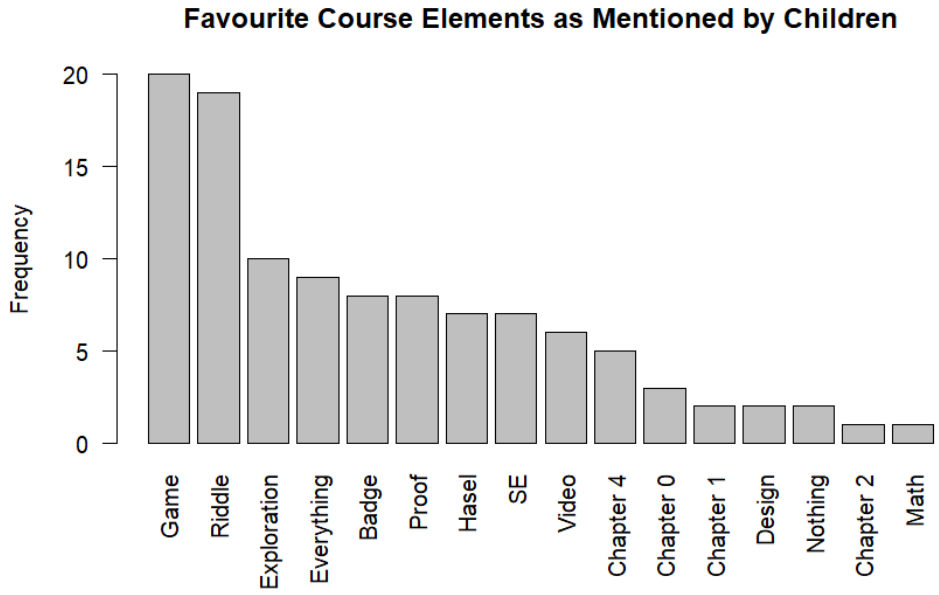
Table 17

Coding Scheme: Written Feedback From Children - Positive Aspects

Categories	Description
Everything	Explicitly refers to the whole course or everything
Design	Refers to the structural design of the course
Hasel	Mentions Hasel or just *the squirrel*
Badge	Mentions badges/nuts/insignia/treats
Game	Mentions the Games in general or a specific game
SE	Refers to self-efficacy (e.g., I learned a lot / I did well)
Exploration	Refers to a specific exploration or the explorations in general
Riddle	Refers to a specific riddle or the riddles in general
Proof	Refers to a specific proof or proving in general
Nothing	States that nothing was great
Video	Mentions a specific video or videos in general
Chapter X	Mentions a specific chapter by name, number, or content
Math	Refers to arithmetics or mathematics in general

Figure 10

Final Written Feedback from Children - Positive Aspects (n=110)



As can be seen in Figure 10, the most popular aspects among the children were the games (20 mentions) and riddles (19) included in the course. About half that many times, children mentioned that they liked the explorations (10) or just everything about the course (9). Then comes earning badges and doing mathematical proofs (each 8), the mascot Hasel and aspects of self-efficacy (each 7), and the videos (6). Chapter 4 was mentioned 5 times, before Chapter 0 (3), Chapter 1 (2), and Chapter 2 (1), while Chapter 3 was not mentioned at all. Other aspects of the feedback were the course design (2) and the mathematical content (1). Two children stated that there was nothing they liked. These numbers are displayed in Table 26.

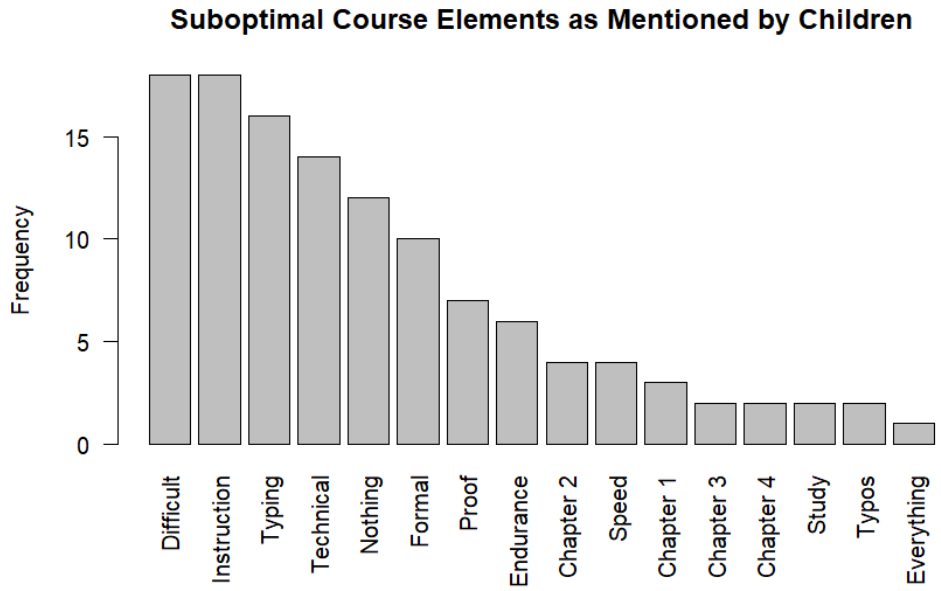
Table 18

Coding Scheme: Written Feedback From Children - Negative Aspects

Categories	Description
Technical	Mentions technical problems specifically or in general
Nothing	*No* or *Nothing*
Instruction	Criticizes lack or quality of instruction
Typing	Mentions difficulties using the keyboard in the respective exercises
Formal	Mentions formal symbols / secret symbols
Difficult	Refers to the difficulty of an exercise or the whole course
Endurance	Mentions that an exercise or questionnaire took longer than expected
Proof	Mentions a specific proof or the topic itself
Speed	Refers to video explanations being synchronized in a fast voice
Everything	*Everything* or *All*
Study	Refers to additional load from the accompanying study
Typos	Refers to spelling errors in exercises or instructions
Chapter X	Mentions a specific chapter by name, number, or content

Figure 11

Final Written Feedback from Children - Negative Aspects (n=121)



Looking at the aspects that the children did not appreciate much (see: Table 18), difficult exercises and lack of instruction were most frequently mentioned (18 times each). Typing with the keyboard

(16) and other technical issues (14) were mentioned almost as often. Twelve children stated that there was nothing to criticize from their perspective. The next frequent aspects were the use of formal expressions (10) and the mathematical proof exercises (7). A total of six children said their endurance was challenged by long exercises. Chapter 2 was mentioned four times, just like the fast speed of the videos, followed by Chapter 1 (3) and Chapters 3 and 4 (2 each). Chapter 0 was not criticized. Furthermore, the children mentioned the additional load from the study (2 and the typos in some texts (2). One child disliked everything.

Feedback From the Legal Guardians When we asked the legal guardians for their open feedback, we did not ask for a positive and a negative aspect, but instead for a comment on technical, content, and organizational matters. The content analysis that we did on these answers yielded three categorical systems that can be found in Tables 19, 20, and 21.

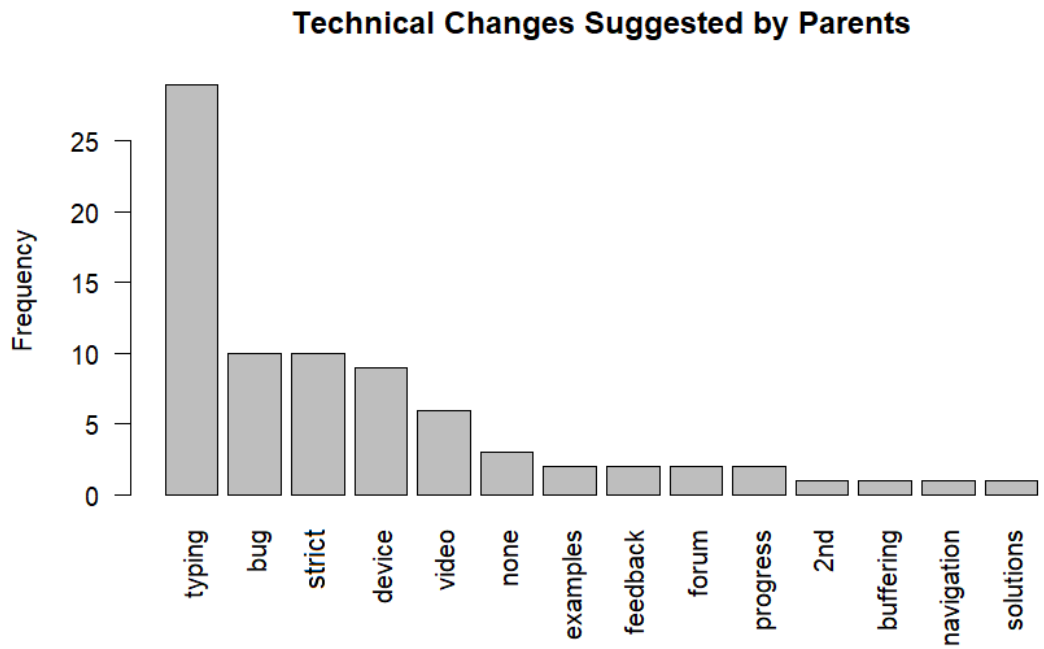
Table 19

Coding Scheme: Written Feedback From Legal Guardians - Technical Aspects

Categories	Explanation
2nd	Thinks that level of technical skill is not feasible for 2nd graders
None	States that there were no technical problems
Buffering	Mentions that the buffering of external elements took too long
Bug	Reports a specific bug or refers to errors in general
Device	Wishes for compatibility with tablets and/or small screens
Examples	Suggests to have more examples
Feedback	Wishes for better feedback in the mathematical proof exercises
Forum	Does not want to receive emails from forum
Strict	Suggests proof checker to accept a broader variety of solutions
Navigation	Wishes for easier navigation between chapters
Progress	Wishes for better monitoring of progress
Solutions	Wants solutions to be displayed
Typing	Suggests buttons instead of typing
Video	Suggests slower talking in videos / better quality /more natural sound

Figure 12

Final Written Feedback from Legal Guardians - Technical Aspects (n=79)



The categorized technical feedback from the legal guardians is depicted in Figure 12. The aspect that the legal guardians mentioned by far the most was typing (29 times), i.e., they criticized that the proving exercises worked via keyboard control and not via mouse. Technical bugs and the proof checker being too strict were mentioned ten times each. Nine legal guardians wished for better compatibility with other devices than computers (e.g., touchscreens, smaller screens). From six legal guardians, we got feedback to adjust the mathematical proof videos regarding speed or sound. No technical changes were suggested by three legal guardians. The wish for more examples, more feedback, fewer notifications from the forum, or more progress monitoring came from two legal guardians each. Finally, there was one comment on the technical level being not suitable for second graders, one on slow buffering, one on the navigation between the chapters, and one on the integration of solutions.

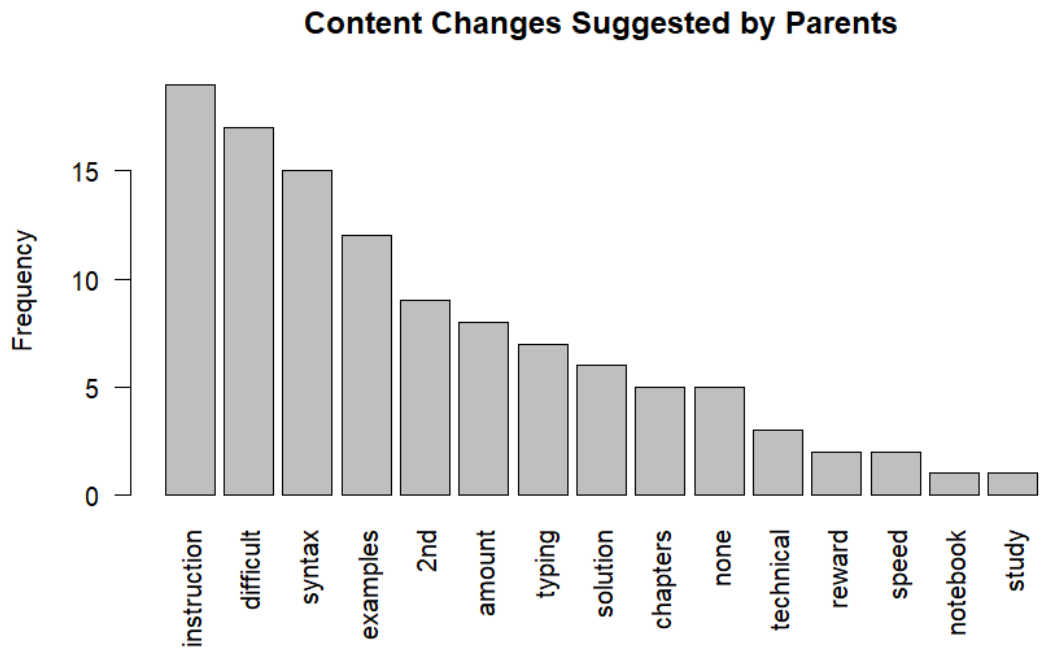
Table 20

Coding Scheme: Final Written Feedback From Legal Guardians - Content

Categories	Explanation
2nd	Thinks that the content is not feasible for 2nd graders
None	No changes in content needed
Amount	Suggests to have fewer exercises or better monitoring of the amount
Chapters	Suggests smaller chapters / other structure
Difficult	Suggests less difficult exercises or differentiation
Examples	Suggests more examples
Instruction	Suggests more instructions or more child appropriate phrasing
Notebook	Suggests to take out the notebook from the course
Reward	Suggests additional rewards (games, certificate)
Solution	Suggests possibility to glimpse at a solution
Speed	Suggests slower talking for Hasel
Study	States that the study was very time-consuming
Syntax	States problems with mathematical symbols
Technical	Reports a technical issue
Typing	Child had problems with typing

Figure 13

Final Written Feedback from Legal Guardians - Content (n=112)



The second categorical system represents the legal guardians' feedback on the course content. The distribution of addressed topics can be found in Figure 13, the exact numbers in Table 29. The most frequently mentioned aspect was the quality and quantity of the instructions (29 times), followed by the perceived difficulty of the exercises (17) and problems with the syntax (15). Twelve legal guardians wished for more examples in the course. The suitability of the course topic for children in Year 2 was doubted by nine legal guardians, while eight legal guardians criticized the workload in the exercises (*amount*). The difficulty of typing with a keyboard was also a topic in this feedback category, as it was mentioned seven times. The need for more solutions was mentioned six times, and five legal guardians wished for smaller chapters, while five comments stated no need for content changes. Other topics addressed were technical issues (3), the wish for more rewards (2), slower video speed (2), as well as the course tool *Notebook* (1), and the scientific study (1).

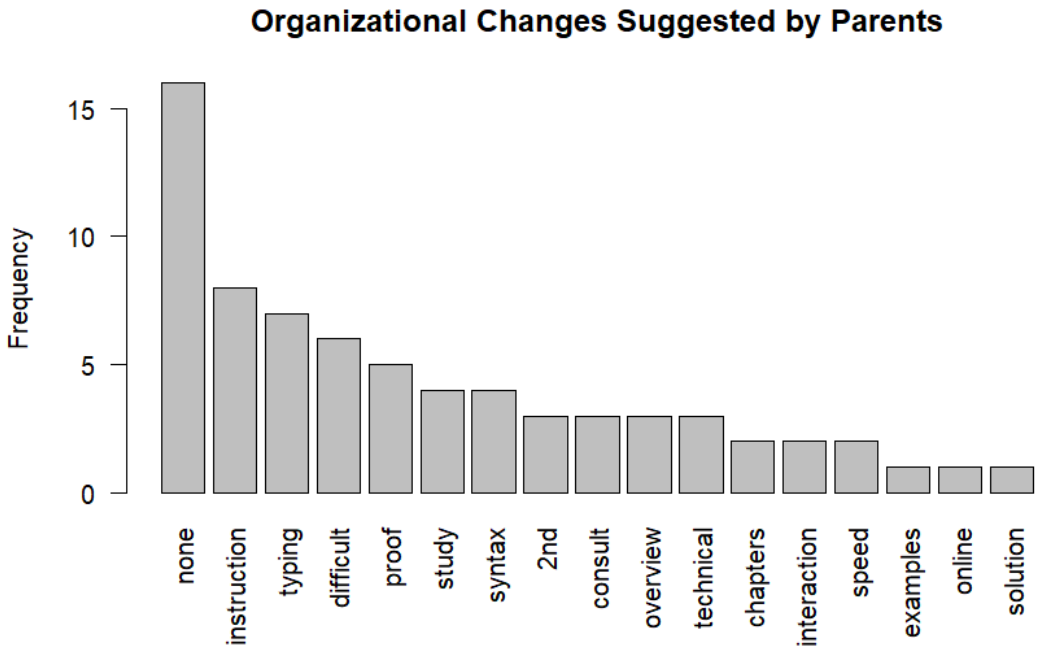
Table 21

Coding Scheme Final Written Feedback From Legal Guardians - Organizational Aspects

Categories	Explanation
None	Course was well organized / no complaints
Chapters	Wishes for shorter chapters
Consult	Wishes for possibility to consult an instructor
Difficult	States that some course elements were too difficult
Examples	Wishes for more examples
Instruction	Wishes for more (explicit) instructions
Interaction	Wishes for exchange with other children
Online	Does not like online courses
Overview	Wishes for a better overview on course progress
Proof	Did not appreciate the mathematical proof parts
Study	Unhappy about the study/time consumption/organization / specific items
Syntax	Child had problems with the logical syntax
Technical	States a specific technical problem
Typing	Child had problems with the typing
2nd	Thinks the course should not be offered for children at the end of grade 2
Solution	Wishes for solutions to be displayed
Speed	Unhappy about video speed

Figure 14

Final Written Feedback from Legal Guardians - Organizational Aspects (n=71)



In Figure 14, we show the distribution of topics addressed in the legal guardians' organizational feedback. 16 legal guardians answered that they would not change anything about the course organization. Other topics that also came up in the qualitative feedback on the two other aspects were the quantity of instruction (8 times), problems with typing (7), the exercise difficulty level (6), and the mathematical proof parts of the course (5). The time consumption and effort of the study (4), as well as the recommendation that the course should only be offered for children from Year 3 (3), also appeared again in this category. Three mentions each go on the wish for more consulting opportunities in the course, a better overview of progress, and technical issues. The wish for smaller chapters, interaction among the children, and slower video speed were mentioned two times each. The number of examples, general unhappiness with online courses, and the visibility of solutions were addressed by one legal guardian each in the organizational category.

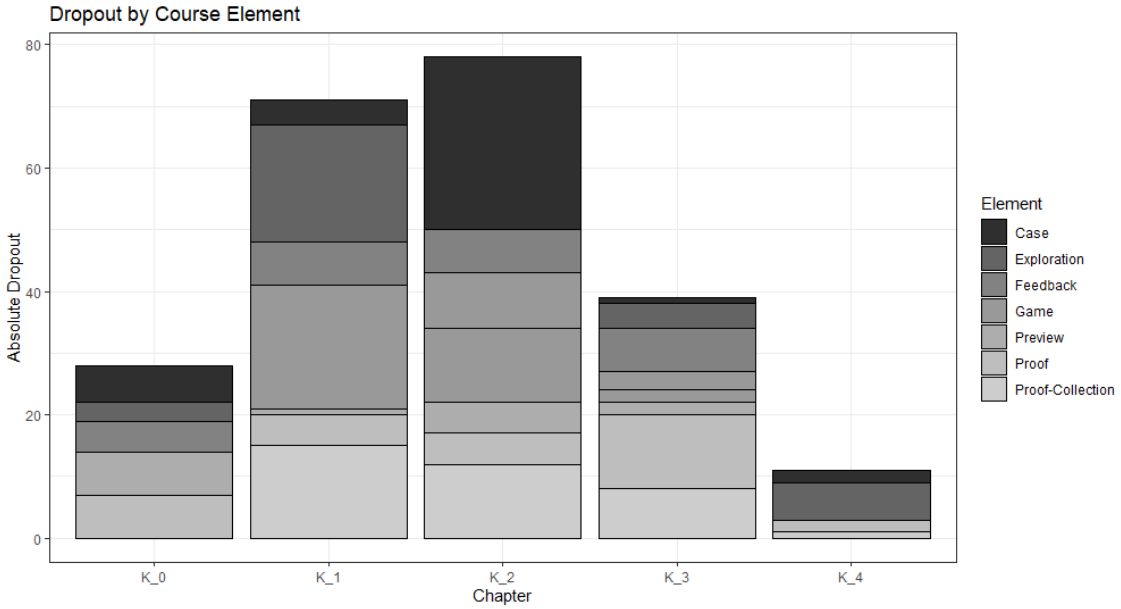
4.4.4 Course log-data

The log data tracked in the course provides a timestamp for each user's activity completion. With this information, we were able to calculate how long it took a user to complete all activities and how much time there was between two consecutive activity completions. We could also see which was the last completed activity of the users who did not finish the course. Figure 15 depicts the

absolute number of children per chapter who did not continue the course. The color indicates which kind of course activity they completed last.

Figure 15

Absolute number of children not finishing, sorted by last chapter visited. Colors indicate the kind of course element last completed.



It can be seen that the highest total dropouts are related to Chapter 1 and Chapter 2, while in Chapter 3 and Chapter 0, there is a mediocre high dropout, and in Chapter 4 a very low dropout. The course elements that many children visited last were the *exploration* in Chapter 1, the *case* in Chapter 2, and the *games* in Chapters 1 and 2. The *proof collections* that are part of the chapters from Chapter 1 on were also an element after which many children did not reappear in Chapters 1-3. In Chapter 3, the interactive video *proof* was also a common last element to complete. It is remarkable that - apart from the *exploration* in Chapter 1 and the *proof* in Chapter 3 - these were all course elements from our adaptation of *Diproche*.

4.5 Discussion

The present study investigated whether an asynchronous online course on mathematical proving is a feasible strategy to foster gifted primary school children's competencies. To do so, we analyzed self-reports by both children and legal guardians and collected log data to gain insights into children's learning behavior.

4.5.1 Key Findings

We found that the children perceived the course to be worthwhile, as they indicated that they were glad to have participated and enjoyed it. Further, the results suggest that they did not miss interaction with other children despite the self-paced nature of the course. Also, the children's responses indicated that they perceived the instructions as friendly and the learning atmosphere as calm. Interestingly, the most popular course aspect to them seemed to have been the digital badges they could receive for finishing activities.

The children's ratings of most course elements (i.e., videos, games, explorations, puzzles) indicate they enjoyed these - games and riddles were by far the most mentioned positive aspects. So it seems that the gifted children greatly appreciated the interactive parts of the asynchronous mathematical proof course. This finding shows the great potential of interactive learning elements, especially in light of the ICAP (Interactive Constructive Active Passive) framework, which suggests that interactive elements are also the most effective way of learning (Chi & Wylie, 2014). This finding is also in line with research on what children value most in self-regulated digital courses: fun elements, appealing design, and rewards (Amaefule et al., 2025).

Still, despite the overall positive feedback, we identified two aspects that afford additional considerations to match the children's demands in the future. Firstly, many children reported that typing entire mathematical proofs on the computer keyboard was difficult for them. The chapters containing such tasks required more technical assistance from their legal guardians and were perceived as more difficult by the children. A possible explanation for this observation can be found in the cognitive load theory. As Hollender et al. (2010) described in their model, technology usage is a source of extraneous cognitive load in digital learning. In our case, the typing might have added to the perceived complexity of the tasks, resulting in less capacity to cognitively engage in the challenging exercises. Secondly, the children perceived the use of formal symbols (like $\subset$ or $\wedge$) particularly challenging. This observation aligns with earlier reports on primary school children having difficulties with mathematical formalism Kline (1973). Together, our findings suggest that the setting was a feasible choice for the enrichment course despite the young age of the target group, and that only a few challenges need to be further addressed.

The feedback from the legal guardians much resembles the children's answers: There was high agreement that the children enjoyed the course and learned a lot. Also, legal guardians reported they appreciated that their child took the course. And, like the children, the legal guardians mostly disagreed that the interaction with other children was missing. Also, the legal guardians did not think that the course caused too much screen time for their child. Thus, the course setting seems

to be generally feasible from the legal guardians' perspective as well. According to Lau, Jian–Bin, and Lee (2021), the legal guardians' satisfaction with online courses for children is moderated by their child's competence in independent learning. Gifted children are more efficient in self-regulated learning than their peers (Risemberg & Zimmerman, 1992). Thus, it is very likely for their legal guardians to be content with the course. This is in line with our findings.

From the log data of the online course, we find that of the 331 children who started the course, a total of 125 children finished all five chapters. This share of roughly 38% is remarkable compared to the less than 10% finishers that are usual for asynchronous online courses (Eriksson, Adawi, & Stoehr, 2016). This comparison suggests that the investigated course is feasible and that the sample was able to engage with the content with above-average endurance.

4.5.2 Theoretical and Practical Implications

From a theoretical perspective, this study can contribute to the discussions on whether the content of mathematical proving is suitable for gifted primary school children.

Challenging, extracurricular content is undoubtedly crucial for the enrichment of talented children (Fuchs, 2006; Käpnick, 1998; Subotnik et al., 2011). Formal mathematical proving is generally known as a challenging mathematical task (Glosauer, 2019). On the other hand, several actors in math education have argued that proving is too complex for children and should not be part of the school curriculum (Bass, 2011). In subsection4.1.3, we argued that Krutetskii's 1976 characteristics of mathematically gifted children, like formalized perception and shortening thought chains, align well with the steps of the proving process from Boero (1999). In our study, we found that the children appreciated tasks in which they aligned and shortened thought chains. However, the rather formal parts of the course were perceived as less appealing by some of the children.

Possible explanations for this finding include the target group's age and composition. First, the characteristics of mathematical giftedness develop gradually over time. Thus, the formalized perception of mathematically gifted primary school children might be precocious, yet not sufficiently developed to understand Boolean symbols rapidly. Second, the domain-specificity of giftedness (see: Subotnik et al., 2011) suggests that not all gifted children are necessarily *mathematically* gifted. Thus, some of the participating children in this study might not display the characteristics described in Krutetskii's model (1976).

From a practical perspective, our findings offer insights into the design of a digital mathematical proof course for gifted primary school children. While the children appreciated working with the interactive presentations and videos, the majority of critical comments addressed our adaptation of the natural language proof checker *Diproche* (originally developed by Carl (2022)). Thus, there

is considerable potential here to derive practical implications. Children and their legal guardians suggested including more examples, instructions, and solutions. Worked examples, in particular, can be effective in drill-and-practice environments. For example, these can be provided in a screen-cast when a new kind of task is established.

Different levels of strategic feedback on the written mathematical proofs can be another measure to face the aforementioned heterogeneity that should be considered when designing such exercises for the age group of primary school children. As discussed in the previous subsection, the children were also challenged by typing on a keyboard, which may have caused frustration in the proof-writing tasks. Therefore, we recommend avoiding tasks with input via text fields. Instead, speech-to-text tools or clickable buttons that insert the required words and symbols could be used to ease cognitive and technical load.

The glossary for logical symbols was also not very popular among the children. According to Elgendi and Shaffer (2020), glossaries in online courses should be replaced with more interactive elements, because alphabetical lists can be unappealing for children. We assent to this view and instead suggest integrating look-up features, such as mouse-over texts, could be integrated directly within the tasks.

Additional findings for the practice can be derived from the aspects that the children appreciated. The popularity of the talking course mascot aligns well with Mayer's personalization principle (2009), which states that a human voice, speaking in a conversational tone, is most effective for online learning. Therefore, we regard this as a feasible way to transfer asynchronous course content to children. Most notably, what the children appreciated most about the course was receiving digital badges for completed activities. This is in line with research on children's demands for self-paced environments, which stresses the importance of gratification (Amaefule et al., 2025). Consequently, we recommend that asynchronous courses for the targeted age group should have a high density of rewards for completed exercises.

Lastly, we reflect on the order in which introductory mathematical proof topics are introduced. Traditionally, in courses for proof novices, Boolean Logic is the first topic, followed by Set Theory and Elementary Number Theory, as the latter is generally perceived as more challenging (Carl et al., 2022; Glosauer, 2019). However, the chapter on Elementary Number Theory was rated as easier than the other two. This probably comes with the necessary didactical reduction of all three topics, which excludes challenging proof techniques (e.g., proof by induction). Based on this, we suggest that proof courses for children could start with simplified Elementary Number Theory exercises before focusing on the topics that require formal symbols.

4.5.3 Limitations and Future Directions

Despite the large sample and multi-perspective approach, there are some constraints to our study.

The first limitation is that in our study, the log data was not gathered in the same system as the self-reports and were consequently anonymized separately. Therefore, it is not possible to link the children's answers to their behavioral traces, and no conclusions can be made about their interplay. To face this issue, we recommend implementing questionnaires directly in the course and deriving the answers as log data alongside the data from all course activities.

The second limitation addresses the format in which the log data was collected from the Moodle system. It allows for insights into which activity was finished at what time, but does not reveal the answers a learner gave or track specific clicks. In contrast, the novel standard experience API (xAPI) makes detailed reports possible that show how a learner interacted with an exercise and what solution they entered (Rotelli, Noël, Lallé, Luengo, & Pesce, 2023). In future studies, this kind of log data could reveal learner profiles and their connection to drop-out, as it was done in self-paced courses for adult learners (e.g. Li & Baker, 2018).

Another aspect to address is that the sample of this study was recruited via a nomination-based enrichment program. Therefore, it could be that not all of the children were gifted, as teachers tend to apply certain biases when nominating for enrichment programs (Golle et al., 2022). This limits the scope of the conclusions we drew about the population of gifted children. However, this pilot study was intended to identify possible difficulties with the material. Therefore, a heterogeneous sample, which includes talented, but not necessarily gifted children, can be useful to uncover instructional gaps, which may have been overlooked otherwise.

Lastly, the results are based on self-reported data and a study design, which did not include a control group, hindering strong inferences about the effectiveness of the course. For future research, this needs to be addressed by adding performance tests and a randomized control group.

4.5.4 Conclusion

Our study demonstrated that enrichment through an asynchronous online mathematical proof course can be a feasible approach even for primary school children. Over three hundred learners from different regions participated in the course at their own pace and provided their data. By administering self-report questionnaires at several points in the course, we were able to identify course elements that were motivating for the children and others that need further refinement. Overall, both legal guardians and children perceived the course as beneficial. Additionally, we got valuable insights into the children's overall dropout behavior through the activity completion data

from the Moodle course.

To enable more children to complete such a course successfully and independently, interventions must fulfill certain additional needs. Specifically, optional scaffolding is required for supporting the heterogeneous target group, typing requirements should be minimized, and algebraic symbols must be introduced carefully and sparingly. Self-paced courses that account for these factors could potentially be a great opportunity for even more children to engage with mathematical proofs at their own pace.

4.5.5 Appendix A

Table 22

Chapter Questionnaire: Learning Conditions

Items (translated)
I enjoyed this chapter.
I got along well in this chapter.
I learned a lot in this chapter.
In the online course, I was able to work calmly and concentrated.
Hasel was able to explain things well.
Hasel was friendly to me.

Note. Options (translated): Completely applies / mostly applies / hardly applies / does not apply.

All items adapted from Rebholz, Golle, Oschatz, and Trautwein (2019).

Table 23

Chapter Questionnaire: Course Elements

Items (translated)
How much did you like it... to see the preview video?
How much did you like it... to click through the exploration?
How much did you like it... to play a found game?
How much did you like it... small cases?
How much did you like it... to solve the big case in this chapter?
How much did you like it... to find the proof in this chapter?
How did you like earning badges?
What did you think of entering your solutions?
How did you find it to get feedback on right and wrong answers?

Note. Options (translated): Very good / good / a bit / not at all.

All items adapted from Rebholz et al. (2019).

Table 24

Post Course Questionnaire - Legal Guardians

Items (translated)
My/our child enjoyed the online course.
My/our child learned a lot in the online course.
My/our child talked about the course content at home.
I am glad that my/our child participated in the online course.
I helped my/child with technical issues.
I helped my/child with the puzzles in the course.
My/our child had difficulties in handling the computer.
My/our child had difficulties in comprehending the course exercises.
My/our child was overwhelmed with the course exercises.
My/our child would have preferred to interact with other children in the course.
My/our child spent too much time on the computer due to the course.

Note. Options (translated): Completely applies / mostly applies / hardly applies / does not apply.

First and second items adapted from Rebholz et al. (2019). All other items are self-generated.

Table 25

Post Course Questionnaire - Children

Items (translated)
I enjoyed the online course.
I learned a lot in the online course.
I am glad that I participated in the online course.
I did manage all technical issues on my own.
I did not solve all the puzzles on my own.
It was hard for me to comprehend.
The exercises in the course were too hard.
I would have preferred to interact with other children in the online course.
The online course made me tired.
The online course bored me.
I had technical help in Chapter X. (X=0,1,2,3,4, none)
I had help with the riddles in Chapter X. (X=0,1,2,3,4,none)

Note. Options (translated): Completely applies / mostly applies / hardly applies / does not apply.

Options for the last two items: Yes/No. First and second items adapted from Rebholz et al. (2019).

All other items are self-generated.

Table 26

Frequency Children's Feedback Positive Aspects

Badge	Chapter 0	Chapter 1	Chapter 2	Chapter 4	Design	Everything	Exploration	Game	Hasel	Math	Nothing	Proof	Riddle	SE	Video	Sum
8	3	2	1	5	2	9	10	20	7	1	2	8	19	7	6	110

Table 27

Frequency Children's Feedback Negative Aspects

Chapter 1	Chapter 2	Chapter 3	Chapter 4	Difficult	Endurance	Everything	Formal	Instruction	Nothing	Proof	Speed	Study	Technical	Typing	Typos	Sum
3	4	2	2	18	6	1	10	18	12	7	4	2	14	16	2	121

Table 28

Frequency Legal Guardian Feedback Technical Aspects

2nd	Buffering	Bug	Device	Examples	Feedback	Forum	Strict	Navigation	None	Progress	Solutions	Typing	Video	Sum
1	1	10	9	2	2	2	10	1	3	2	1	29	6	79

Table 29

Frequency Parent Feedback Content Aspects

2nd	Amount	Chapters	Difficult	Examples	Instruction	None	Notebook	Reward	Solution	Speed	Study	Syntax	Technical	Typing	Sum
9	8	5	17	12	19	5	1	2	6	2	1	15	3	7	112

Table 30

Frequency Parent Feedback Organizational Aspects

2nd	Chapters	Consult	Difficult	Examples	Instruction	Interaction	None	Online	Overview	Proof	Solution	Speed	Study	Syntax	Technical	Typing	Sum
3	2	3	6	1	8	2	16	1	3	5	1	2	4	4	3	7	71

5 STUDY 2: DEVELOPMENT AND VALIDATION OF A PREFORMAL TEST FOR MATHEMATICAL PROOF COMPETENCY

The content of this chapter is currently under review in Thinking Skills & Creativity. The proportional contributions of the (co-)authors to the manuscript are presented in the subsequent table. This article may not exactly replicate the final version published in the journal. It is not the copy of record.

Author	Author position	Scientific ideas (%)	Data generation (%)	Analsysis & interpretation (%)	Study writing (%)
Xenia Stein	first	50	100	50	80
Katerina Tsarava	second	0	0	0	10
Benjamin Goecke	third	50	0	50	10

Abstract

Proving is a core part of mathematics, prompting educators to push for earlier inclusion in curricula. Thus, an appropriate test for proof competency is needed already at primary school level. The Preformal Proving Test (PfPT), introduced here, is an online assessment covering proof in Boolean Logic, Set Theory, and Elementary Number Theory. Designed without formal language and prior knowledge requisites, it ensures accessibility for children. We present the findings from two studies: an online pilot study (n = 409, 42% identifying as female; M_{age}= 9.28, SD = .67) and an on-site validation study (n=180, 36% identifying as female; M_{age} = 9.80, SD= .63) examining internal consistency, factor structure, and validity measures like reasoning and intelligence. Results showed diverse item difficulty and promising psychometrics that merit further investigation. Given the PfPT's ease of administration in online and classroom settings, it is a valuable tool for future primary math research.

Keywords: proving; primary education; proof competency; logic; online; measurement

5.1 Introduction

Mathematical proof has been recognized as one of the main pillars of mathematics (Bass, 2011). However, what is considered a proof may vary depending on the context: Wittmann and Müller (1988) distinguish between three different kinds of proofs, i.e., formal-deductive, substantive-illustrative, and experimental. Formal-deductive proofs consist of logical argumentative chains expressed with formal symbols that hold under professional mathematical standards. Substantive-illustrative proofs also use general mathematical operations, but merely those that can be intuitively recognized as universally applicable, allowing readers to acknowledge the generality of the claim immediately. In contrast, experimental proofs only demonstrate an argument through specific examples and are therefore not as rigorous but particularly suitable for proof novices.

In Germany, formal-deductive proving is taught almost exclusively only at the transition to higher education (Carl, 2022; Glosauer, 2019; Kempen & Biehler, 2014), and the German school curriculum incorporates only limited instructional approaches to fostering proving (Ministerium für Kultus, 2016a, 2016b). However, both researchers and teachers have been advocating for the earlier integration of formal-deductive proof into school mathematics (Ball & Bass, 2003; Bass, 2011; Bleiler, 2009). To bridge this gap, pre-formal activities can be introduced in the mathematics classroom (Gernes, 1999). When implementing such lessons, reliable tests are needed, not only to assess the children's domain-specific prerequisites but also to monitor the learning effects while attending such activities.

To date, several attempts have been made to measure proof competency in adolescents and adults; however, there is a lack of standardized assessment for pre-formal proof competency in primary school children. Existing standardized assessments predominantly target secondary school students and heavily rely on higher arithmetic and basic algebraic skills (e.g., Healy and Hoyles (2000), Kempen and Biehler (2014), Bickerton and Sangwin (2021)), which makes them unsuitable for younger children. Consequently, in the present studies, we aimed to develop and validate a novel assessment of pre-formal proof competency in young children from grades 3 and 4 in the German primary school system (usually age 8 to 10). This assessment consists of child-friendly items designed to assess foundational proof-related concepts in three mathematical disciplines that form the pillars of formal-deductive proving (Carl, 2022; Glosauer, 2019): i. Boolean Logic (BL), that is, formal logic based on binary truth values, ii. Set Theory (ST), which focuses on determining which mathematical elements belong to or are excluded from a set, and iii. Elementary Number Theory (ENT) that consists of universal conjectures involving natural numbers (e.g., divisibility, number parity).

In the present report, we first provide a conceptual framework for mathematical proof competency (PC), describe its nomological network, and present an overview of existing research on the measurement of this concept. Across two studies, we describe the development and validation of the PfPT, a test instrument specifically designed for primary school children (not to be confused with the construct of PC). We provide evidence supporting its applicability in both online and proctored laboratory settings, evaluate its psychometric properties, and provide evidence of convergent and discriminant validity based on correlations with measures of computational thinking, fluid intelligence (i.e., figural, verbal, and numerical reasoning), crystallized intelligence, and domain-specific motivation. Lastly, we explore the correlations of the test score with school grades and demographic aspects (i.e., gender, age, school year, and socio-economic status (SES)).

5.1.1 Mathematical Proof Competency

In defining PC, the literature highlights two core aspects: proof construction and proof comprehension (Waluyo et al., 2019). Proof construction, which is the action of writing a proof for a mathematical conjecture, begins with a cognitive process (i.e., the exploration of a conjecture's scope) and ends in a formal product (i.e., the formal proof) (Boero, 1999). In contrast, proof comprehension describes the process of reading a proof and making sense of it. It starts with the application of formal expressions and rules, followed by their logical evaluation and summary of results. It ends with the development of a cognitive representation of the conjecture's generality and applicability (Yang & Lin, 2008). In other words, proof *construction* moves from cognitive actions to formal expressions, while proof *comprehension* proceeds from formal expressions to cognitive actions.

Since the comprehension of most formal expressions requires algebraic knowledge, which is typically not taught in primary school, we have decided to focus on the pillar of proof construction only. The aim is to explore the underlying cognitive process involved in proof construction even before the development of algebraic knowledge. Thus, in the current report and test development, we will only regard this aspect when referring to PC. The exact definition that we will use in our further deliberations was proposed by (K. Lee, 2016): "[Proving is] the process of constructing mathematical assertions to determine the largest of mathematical objects for which the mathematical proposition is true or false through the search for possible examples and counterexamples."(K. Lee, 2016), p.28. In other words, applying rules to determine for which exact part of the mathematical world a claim holds. Building on this conceptual framework, we now turn to the broader network of cognitive factors that might influence proof competency.

5.1.2 Existing Assessments of Proof Competency

PC assessments can be categorized into the two categories mentioned in the introduction: Assessments targeting proof comprehension and assessments targeting proof construction (Waluyo et al., 2019). Although our research primarily focuses on the proof construction (as detailed previously), we reviewed assessments from both categories to provide a comprehensive overview of existing test instruments in the field. Table 31 gives an overview of the instruments available from the literature.

These assessments were all designed for target groups much older than primary school students: Two tools were developed for university students (Kempen & Biehler, 2014; Mejía–Ramos et al., 2017) and two for middle-school students (Healy & Hoyles, 2000; Senk, 1989). If we focus on our targeted construct of *proof construction*, only one test per age group remains. We considered both these tests as possible candidates for an adaptation for younger students:

The test of Kempen and Biehler (2014) consists of a single, open assignment in which a proof is to be written. To evaluate the answers, the authors developed a categorical system that they based on previous models for proof quality and refined it in a pilot study (Kempen & Biehler, 2014). Still, open questions are very time-consuming for both the test-takers and the evaluators (Chan & Kennedy, 2002). Additionally, they bear a higher risk of coding errors compared to single- and multiple-choice tests (Kennedy & Walstad, 1997). Lastly, the robustness of the test could be improved by having more than one item. For the target group of primary school children, this task might be comprehensible. Still, the mathematical arguments for why this claim holds are most likely not available from their number theory knowledge. Thus, the categorical system developed for the raters will most likely not apply to their age group.

The CDASSG Proof Test by Senk (1989) was developed for students from Year 7 and older. Thus, it seems more likely to be adaptable for younger students. It consists of six items, of which four are complete proofs to be written and two are short-answer questions, all of which were to be completed within 35 minutes (Senk, 1989). In a study with 241 students, they found the instrument to have a good internal consistency (Senk, 1989). However, this assessment requires writing complete proofs. Thus, it seems unrealistic to adapt this instrument for primary school children whose basic writing skills are still developing.

5.1.3 Nomological Network of Proof Competency

Considering the aforementioned definition of proof construction by K. H. Lee (2005), we derived that an instrument measuring this dimension of proof competency should contain items in which several mathematical objects are presented, and the task is to verify or falsify given statements

for the largest possible subset. To cover the preformal aspect and be suitable for the age group of primary school children, these expressions should be provided in figural expressions or with simple arithmetic terms as established during the first school years. Thus, it seemed natural in the item design for the ST and BL sub-scales to use simple geometric figures with different shapes and colors like the New Mathematics movement used in teaching ST to primary school children (Hanna & Knipping, 2020; Sua Flórez, Gutierrez, & Jaime, 2020). This kind of representation is widespread in assessments for other cognitive constructs in the field of reasoning and problem-solving (e.g., in the Cognitive Compiling Test Marinus, Powell, Thornton, McArthur, and Crain (2018)). Thus, it can be assumed that the PfPT will have similarities with assessments for other cognitive constructs, like reasoning or computational thinking tests, when it comes to the item design. Besides, these constructs have relevant overlaps from a cognitive perspective, as we will illustrate in the following.

Reasoning There is much discourse about the question of what *reasoning* is (McHugh & Way, 2018). For example, reasoning can be regarded as the process of evaluating and combining knowledge to a conclusion (Stenning & van Lambalgen, 2010). The corresponding *reasoning ability* according to Kyllonen (2020) is defined as "the power and effectiveness of the processes and strategies used in drawing inferences, reaching conclusions, arriving at solutions, and making decisions based on available evidence. "(Kyllonen (2020) p.1). In psychology, two forms of reasoning with their related abilities are usually distinguished: deductive reasoning, which derives necessarily true conclusions from true premises, and inductive reasoning, which infers information by expanding semantic content from premises to conclusions, while sometimes abductive reasoning, which connects a fact to its cause, is added as a third form (Wilhelm, 2005). In modern models of intelligence, fluid reasoning (gf) is included as one pillar of intelligence, defined as the ability to apply one's attention to solve novel problems (W. Schneider & McGrew, 2012) spontaneously. As an underlying structure, researchers distinguish between numerical (gfn), verbal (gfv), and figural reasoning (gff), each with a specific measurement format (Schroeders, Schipolowski, Zettler, Golle, & Wilhelm, 2016; Wilhelm, 2005).

Within the didactics of mathematics, reasoning and proving are often described as interwoven competencies (Bleiler, 2009; G. Stylianides & Silver, 2007; Thompson et al., 2012). G. Stylianides (2008) even describe reasoning as a prerequisite for proving and established the term *reasoning-and-proving* to refer to activities in the didactical transition from reasoning to formal proving. Similarly, Gutierrez and Jaime (1998) feature *proof* as the highest procedural level in their adaptation of the Van Hiele levels of reasoning. Therefore, we anticipate a significant positive correlation between PC and gf, as well as between PC and gfn, gfv, and gff.

Crystallized Intelligence According to the dichotomous theory of intelligences by Cattell (1957), crystallized intelligence (gc) means knowledge abilities acquired through acculturation. Tests measuring gc assess both the depth and breadth of an individual's cultural knowledge in comparison to their age group, typically using verbal multiple-choice questions drawn from the knowledge domains mentioned above (Schroeders et al., 2016). However, since these tests are based on specific cultural contexts, they may be subject to biases related to individuals' cultural backgrounds (Watrin, Schroeders, & Wilhelm, 2023). Therefore, conclusions about an individual's gc should take these dependencies into account. In the Cattell-Horn-Carroll model (CHC-model) of intelligences, gc is considered the counterpart to fluid intelligence (gf) (W. Schneider & McGrew, 2012). Consequently, we anticipate a low correlation between gc and performance on tasks measuring PC, as we assume that PC is more closely aligned with the reasoning aspects of intelligence.

Computational Thinking Computational thinking is defined as "the conceptual foundation required to solve problems effectively and efficiently (i.e., algorithmically, with or without the assistance of computers) with solutions that are reusable in different contexts. "(Shute, Sun, & Asbell–Clarke, 2017) p.142 and is correlated with problem-solving (Tsarava et al., 2022). Thus, it can be assumed that CT also has intersections with PC, which is likewise related to problem-solving (Pólya, 2010). When comparing the processes of CT (Barr, Harrison, & Conery, 2011) and PC (Boero, 1999), we identify notable similarities, as both processes require similar action: The procedure of CT starts with the formulation of a problem, PC requires the initial formulation of a presumption. Additionally, in each model, a step involves selecting and ordering logical elements, and another step regards the presentation of this chain efficiently and with adherence to the subject's standards. Furthermore, the definition of CT by Shute et al. (2017) locates CT in the field of problem-solving, which itself has conceptual connections to PC: On the one hand, proving in higher mathematics is often regarded as a problem-solving process (Weber, 2005). On the other hand, Pólya's procedure of problem-solving includes the proof of the found solution at the end (Pólya, 2010). Given these similarities, we assume that the cognitive skills underlying CT and proof construction are associated, though still distinguishable. Furthermore, Nurlaelah, Pebrianti, Taqiyuddin, Dahlan, and Usdiyana (2024) showed that in adults, PC can be increased by fostering CT. Therefore, we expect a moderate to strong correlation between PC and CT.

Non-cognitive Constructs For other cognitive STEM-constructs, relations to corresponding non-cognitive variables, like the domain-specific self-concept, have been discovered: For example, Hansford and Hattie found for the general and math self-concept small to moderate positive

correlations with mathematical competence (Hansford & Hattie, 1982). Thus, we consider that there could also be relations between PC and domain-specific non-cognitive constructs. We will explore the respective correlations to expand the nomological network of PC if strong connections are observed. To thoroughly evaluate possible connections, we included the following constructs:

- domain-specific self-concept: Students' confidence in their ability to do well in the subject (here: proving) (Wilkins, 2004); measurement adapted from Arens, Trautwein, and Hasselhorn (2011); Gaspard et al. (2015)

- general self-efficacy: "the conviction that one can successfully execute the behavior required to produce the outcomes"(Bandura (1977), p.79) ; measurement adapted from Beierlein, Kovaleva, Kemper, and Rammstedt (2012)

- domain-specific attainment value: the importance a student sets on the subject (here: proving) (Gaspard, 2015); measurement adapted from Ramm et al. (2006)

- domain-specific interest: positive feelings towards a specific content (here: proofs)(Hidi, 2006; Krapp, 1993); measurement adapted from Stalder (2013)

5.1.4 Summary and Research Gap

In summary, existing assessments of mathematical proof competency are well-suited for learners with prior knowledge of formal reasoning and algebraic symbol use, typically from secondary school or university. To our knowledge, no PC assessment instrument exists for primary school children. This highlights a clear research gap and a need for the development and validation of an assessment instrument that is developmentally appropriate for young learners.

To address this gap, we developed a new test designed specifically for primary school proof novices: The PfPT. Our key design criteria were: i. time efficiency and low cognitive load (i.e., short instructions that do not require elaborate reading skills), and ii. age-appropriateness (i.e., no prior knowledge of algebraic terms necessary, not advanced reading skills).

5.1.5 Aims of the Present Studies

Across two studies, we developed a test on PC suitable for primary school children and validated it. We decided first to carry out an online pilot study (Study 1) to ensure the feasibility of this unsupervised online test. The goal of Study 1 was to determine whether children could complete a reasonable number of items in this test setting and to develop a robust measurement model of PC. Furthermore, we intended to evaluate a broad range of possible items regarding their difficulty and

factor loadings so that the test could be shortened to be both time-saving and statistically sound. The intended test setting was an asynchronous multi-site online setting, which means that there will be no test advisor present. Thus, all instructions and time limits had to be included in the digital test implementation itself. For the final test, we targeted an item difficulty ranging evenly from .25 (guessing threshold for 1-in-4 single-choice items) to .9 to prevent both floor effects in a possible pretest scenario and ceiling effects in a posttest scenario.

In summary, in Study 1, we were interested in the following aspects: 1. Determine the item difficulties for all our developed possible test items for the PfPT. 2. Investigate the feasibility of the digital test setting in an unsupervised online situation. 3. Develop a measurement model of PC with sufficient psychometric properties and sample an item set for the PfPT. 4. Study correlations of figural reasoning (gff) with PC.

In Study 2, we targeted the following aspects: 1. Review the feasibility of the digital test setting in a supervised on-site situation. 2. Refine the developed model of PC from Study 1. 3. Study correlations of CT, gfn, gfv, gff, gc, as well as demographic aspects, with PC.

The study was not preregistered. We provide all newly developed test materials for use by interested researchers, along with all data and code necessary to reproduce our results, at `https://osf.io/tvarh/`.

5.2 Materials and Methods - Study 1

5.2.1 Participants

For our online pilot study, a sample of n = 409 children (M_{age}=9.28, 42% identified as female) was recruited via the Moodle online learning platform of the Hector Children's Academy Program (HCAP). The optional additional gff measures were completed by n=136 of the children (M_{age}=9.22, 41% identified as female). The HCAP is a statewide extracurricular enrichment program for gifted primary school children in the German state of Baden-Württemberg who are nominated by their teachers based on the classroom impression of their talent (Trautwein et al., 2023). The statewide coordination of the HCAP granted permission for this study. Children and parents gave their informed consent for participation. As an incentive for participation, all children who completed all three sub-scales of the initial PfPT version were eligible to participate in a prize game, where ten board games were raffled. The ethics committee of the faculty of economics and social sciences at the University of Tuebingen gave their approval for the study (A2.5.4-296_vb, June 6, 2023).

5.2.2 Study Procedure

The study was carried out via Moodle and Unipark - a digital survey tool. In the first step, children and parents gave their informed consent via an online interface. Then, the children were randomly assigned to one of two initial, parallel test versions (A or B) and could immediately start the test. All children were presented with the items in the same order, sorted by sub-scale and arranged by increasing presumed difficulty (informed by expert rating and a pre-pilot). For each of the three sub-scales (20 items), the time was limited to 20 minutes. The children could take a self-paced break of up to five minutes between the blocks of the different sub-scales. Additionally, all children who wanted to complete more cognitive tasks could voluntarily spend another 15 minutes to complete the gff scale of the BEFKI (Schroeders et al., 2016). This test consists of 16 items in which participants have to continue a figural pattern by selecting the next two tiles, each from a set of three (Schroeders et al., 2016).

5.2.3 Test Design

For the content of our newly developed test, we focused on the mathematical disciplines of BL, ST, and ENT, as these are the basic categories of proof learning, and all proofs in higher mathematics rely on the laws of these (Carl, 2022; Glosauer, 2019). For each of the three disciplines, we identified the crucial mathematical operations and, in a second step, designed several tasks featuring a respective expression: For BL, we constructed items on the logical functions AND, OR, NOT, IF, and combinations of up to two of these functions. For ST, we targeted the cognitive processing of the mathematical operations *union* ($\bigcup$), *intersection* ($\bigcap$), *difference* (A\B), and *complement* (A^C). For ENT, we targeted concepts of *number parity, divisibility,* and calculating with *unknowns*. All designed items, as well as a mapping table showing which items target which construct, are accessible at `https://osf.io/tvarh/`.

In the design of the items, we relied on the definition of proof construction by (K. Lee, 2016) mentioned above. Thus, for the sub-scales on BL and ST, we constructed items in which participants had to verify or falsify a given verbal expression (e.g. *Symbols are red AND square-shaped*) for each element displayed along with the expression, so that they determined the largest set of objects for which the expression is true. To ensure an age-appropriate design, these elements were simple symbols with different shapes and colors, similar to the items in the cognitive compiling test by (Marinus et al., 2018), which was specially developed for young primary school children. For the ENT sub-scale, which inevitably had to contain arithmetical tasks, the items showed a squirrel performing a sequence of arithmetical operations with one variable. The children's task was either

to select which given numbers were valid results of this computation or to decide how the parity of the unknown number would change.

In the first step, we created a total of 80 initial items in line with the framework described above. Example items are provided in figures 16 and 17. To test both the children's ability to verify and to reject statements, we designed tasks in two variants: The first item variant required participants to identify the correct answer. In contrast, the second category tasked them with detecting errors within a provided response, akin to the debugging tasks employed in various programming tests (e.g., the abbreviated Computational Thinking Test (aCTt) (Tsarava et al., 2022)). The allocation of the items to these variants is indicated in the mapping table at `https://osf.io/tvarh/`. To determine content validity, the initial items were reviewed by three experts from the field of mathematics and its didactics, who research proof learning themselves. We then excluded all items that were disapproved by at least one expert and changed the wording according to their suggestions.

After working through the experts' feedback, we continued the test development with a total of 60 items (20 per sub-scale). For the pilot study of the test (i.e, Study 1), we created a parallel version of the item set with five items overlapping in each sub-scale. In advance of this pilot study, each sub-scale was pre-piloted with a child to ensure feasibility. In the following paragraph, we briefly report on the pre-pilot.

Pre-Pilot of the Item Set To ensure that the items were suitable and comprehensible for the targeted age group, a pilot run with semi-structured interviews was conducted and evaluated as part of a master's thesis project. For this purpose, we invited six children (each 8 years old) to work through one sub-scale of either the A- or the B-version of the item set (i.e., 20 items) in 20 minutes as far as they could. After that, in a semi-structured interview, they were asked how they liked the tasks and if there were phrases they did not understand. In general, the children greatly appreciated the illustrative item design. Those children who were given a version of the BL or ST sub-scale were able to work through all 20 items in less than 20 minutes, whereas the children solving the ENT sub-scale needed additional time. Thus, we adapted the instructions of this sub-scale, simplified some items, and added respective examples to make the tasks more comprehensible. Some children also mentioned items with ambiguous wording, which we then changed accordingly.

These aspects show how the items were designed, considering the needs of a very young target group. However, validating a test instrument requires not only qualitative considerations but also quantitative evaluation. In the following section, we therefore present how we assessed the statistical properties of the PfPT across the two studies reported here.

5.2.4 Statistical Analyses

In Study 1, we followed a three-fold analytical strategy: First, we considered all items in versions A and B and examined their item characteristics, including item difficulties (i.e., proportion of participants that solved the item), and, as all items were dichotomous, the tetrachoric intercorrelations of these items. Second, we applied a series of confirmatory factor analyses to a manually reduced item set to determine the underlying factor structure of the sub-scales. This strategy was supposed to inform our third approach. Third, to receive a sufficient item set, we used the method of Ant Colony Optimization (ACO), a meta-heuristic algorithm for short-scale construction in the context of confirmatory factor analysis (Zimny, Schroeders, & Wilhelm, 2024). ACO allows for an automated, data-driven selection of items based on predefined psychometric criteria, such as model fit, internal consistency, item difficulty distribution, and factor correlations. To prevent excessive computing times, we calculated cutoffs for the fit indices in advance so that the ACO would only deliver models obeying these: For the A-Version: RMSEA = .06; CFI = .95; Corr.= .55; ω = .80, and for the B-Version: RMSEA = .08; CFI = .91; Corr. = .65; ω =.80. We applied the ACO procedure separately for the A- and B-versions, generating ten promising item sets per version, each containing 24 items. From these, we selected four sets per version, demonstrating a broad distribution across content categories and sufficient model fit. These selected sets were subjected to additional evaluation of item difficulty and inter-item correlation.

We then assessed the provided models using the thresholds for classical fit indices: For a good model fit, we used: CFI (comparative fit index) $\geq$ .95, RMSEA (root-mean-square error of approximation) $\leq$ .06, and SRMR (standardized root-mean-square residual) $\leq$ .08 (Hu & Bentler, 1999). For acceptable model fit we used: CFI $\geq$ .90, RMSEA $\leq$ .08, and SRMR $\leq$ .10 (Bentler, 1990; Browne & Cudeck, 1992). Additionally, due to recent concerns regarding fixed cutoffs (Groskurth, Bluemke, & Lechner, 2023), we determined dynamic fit indices (Mcneish & Wolf, 2023). Specifically, we evaluated the empirical model fits (CFI and RMSEA) of our data against the suggested dynamic cutoffs for mediocre, fair, and close model fit. In this context, fair fit for CFI and RMSEA is considered the minimum requirement, and close model fit indicates excellent fit of a model (Mcneish & Wolf, 2023; Wolf & McNeish, 2023). In the results section, we will report both traditional and dynamic fit indices but will not reject models based on single fit indices. Furthermore, as an estimate of factor saturation, McDonald's ω was computed (McDonald, 1999). This value indicates the variance accounted for by a latent variable in all indicators to which it is related (M. Brunner, Nagy, & Wilhelm, 2012). In all confirmatory factor analyses, we used WLSMV as an estimator.

After that, we computed the correlations of the children's gff-score with their PC score according

to the latest PfPT version derived from the previous analyses. The same was done for the sub-scales of that model. The package lavaan is used for modeling purposes (Rosseel, 2012), and the package tidyverse is used for basic data transformation (Wickham et al., 2019).

5.3 Results - Study 1

5.3.1 Descriptive Results of the Initial Test Versions

In Study 1, the students were randomly assigned to one of two parallel test versions, each with 60 items in total, as there were 20 items per sub-scale (BL, ST, ENT). The descriptive statistics for the sum scores of both versions are displayed in 32 and 33. Interestingly, in both groups, the children scored lowest in the ENT-subscale and highest in the BL-subscale. This indicates that ENT is the most difficult and BL is the easiest sub-scale.

The skewness and kurtosis values for ST and ENT indicate nothing abnormal. However, for BL, the kurtosis exceeds the acceptable value of +2 in both parallel groups, suggesting a distribution that is too peaked. While in Group A the skew for the BL-subscale was within the acceptable range, in Group B it was less than -2, suggesting ceiling effects on that scale. Thus, in the process of manually selecting items for further test refinement, items that are effortless or very similar should be removed.

5.3.2 Sampling Items From the Initial Item Set

For the usable version of the PfPT, we targeted a lower number of items per subscale and good item statistics. In the initial test versions, we identified several items that failed to meet these demands, including those with extreme difficulty levels that exceeded the expected guessing probability. Consequently, these items were considered for exclusion from the test. In both groups, we sought items with a low item-total correlation ($< .300$) for their respective sub-scales to exclude them as well. In Group A, this was the case for two items in the Boolean Logic sub-scale and five items in the ENT sub-scale. The ST sub-scale of test version A had no such items. In Group B, we found a low item-total correlation for five items in each sub-scale. Therefore, we excluded these items as well.

Furthermore, we discovered excessively high inter-item correlations between several items, so we decided to keep only one or two items of each of these clusters to avoid redundancies. This last aspect allowed us several combinations of keeping and excluding items. Thus, to create candidates for usable PfPT sub-scales, we combined different selections of the remaining items into subsets of seven: For ST, this yielded three possible shorter A-versions and B-versions; for ENT and BL, it

was two possible A- and B-versions each. But not all of these were sufficient when the interplay of the items in each short version was investigated: For each of these subsets, we compared the sum-score distributions to inspect possible floor or ceiling effects visually. As a consequence, we only included one remaining variant per sub-scale and test version in the further considerations, leading to one preliminary shorter item set for the A- and one for the B-version of the PfPT, á 21 items. These item sets already fulfilled our demands for item statistics. Nevertheless, a sound test instrument corresponds with a sufficient measurement model of PC. Therefore, we considered two competing measurement models with each of the item sets: a unidimensional solution and a correlated factors model solution with three factors depicting BL, ST, and ENT. The unidimensional measurement model would indicate that the three subtests (BL, ST, ENT) do not correspond to distinct subordinate constructs of PC. In contrast, the three-factorial model reflects the assumption that each subtest assesses one of three different, but correlated constructs. Importantly, a correlated group factors model with three indicators is statistically equivalent to a higher-order factor model, which reflects the concept of a higher-level construct (PC) explaining the correlations (i.e., common variance) amongst the first-order constructs (i.e., sub-scales).

When looking at the item difficulties in the short version A, we saw a range between .95 and .30. For the B-version, the difficulties span from .91 to .26. To learn about the test's sub-scales, we reviewed the factor correlations in each test version: In version A, the correlation between ST and ENT amounted to r = 0.713 (SE = 0.050), the correlation between ST and BL is 0.859 (SE = 0.056), and the correlation between ENT and BL is r = 0.870 (SE = 0.062), indicating strong positive associations between any two of these factors. In version B, the correlation between ST and ENT is r = 0.806 (SE = 0.108), the correlation between ST and BL is r = 0.868 (SE = 0.112), and the correlation between ENT and BL is r = 0.718 (SE = 0.072). In both versions, all sub-scales correlate highly positively with each other. Still, the correlation between ST and BL is even stronger than any other in both versions. This gave rise to the idea that there is a stronger connection between these two sub-scales. Additionally, when considering the fit indices for the threefold models and the current set of manually derived items, there was no sufficient fit: In version A we obtained CFI = 0.764; RMSEA = 0.105; SRMR = 0.202 with factor saturation values of ω_{BL}= 0.553, ω_{ST}= 0.734 and ω_{ENT}= 0.896. For version B, the analyses yielded CFI = 0.663; RMSEA = 0.085; SRMR = 0.170 with factor saturation values of ω_{BL}= 0.605 , ω_{ST}= 0.426 and ω_{ENT}= 0.776.

Consequently, we revisited our theoretical approach: The mathematical disciplines of BL and ST may differ in their notation and regarding the mathematical objects to which they are applied. However, the logical operations can be seen as parallel (Example: An element from the *intersection* of sets A and B is an element that is in A *and* is in B.). Thus, we decided to combine these factors

into one and further pursue a model with two sub-scales: BL-ST and ENT. Our first approach was to combine the items sampled at this point and test the two-fold model with these. This did not result in sufficient fit statistics. Thus, we started the item sampling procedure from the start, targeting the two-dimensional structure. To compare possible item combinations more effectively, we used a sampling algorithm instead of manual selection this time. We will describe this in detail in the following section.

5.3.3 Ant Colony Optimization

As described in the previous section, the overlap between the BL- and ST-sub-scales speaks for a model with only two factors, while our initial considerations assumed a three-fold structure. Therefore, we decided to start the item sampling procedure afresh for both the A- and the B-version, aiming at one ENT sub-scale and one sub-scale built from the common pool of BL and ST items. We strived to compare many candidates for a small, psychometrically sound item set that represents the two correlated factors well. Thus, we applied an Ant Colony Algorithm (ACO) for fit optimization, a method that efficiently compares the fit indices of many different models and adjusts accordingly (Zimny et al., 2024).

We employed the ACO script twice, once for the A- and once for the B-version of the initial item set, so that for both versions we would receive the most promising twofold correlated factor models. As the final test should not be too long, considering the young age of the target group, we aimed at a model with 24 total items, namely 14 items from the BL-ST sub-scale and 10 items from the ENT sub-scale. The ACO algorithm computed 48.782 different item sets for the A-version of the test and 51.112 item sets for the B-version. We then used the algorithm to determine the ten promising item sets per version, optimizing for CFI, RMSEA, and minimal load. Still, not all of these were able to bear scrutiny. To enable the reader to reproduce these deliberations, the descriptive statistics, correlations, measurement models, and confirmatory factor analysis for all considered item sets are provided in `https://osf.io/tvarh/`.

Eventually, one item set was drawn from version A that we regard as the best solution according to our needs. We will refer to this set as *the resulting item set* from here on.

Firstly, we describe the distribution of item difficulties for the resulting item set 24: The item difficulties ranged from .96 to .25 (guessing threshold) and are almost evenly distributed between. Furthermore, the difficulty ranking of the items shows that all sub-scales contain both more complex and easier items.

The sum scores reached by the participants in the two sub-scales are depicted in Fig.19 and Fig.20. We find that the BL-ST sub-scale is left-skewed, while the ENT sub-scale appears relatively

symmetric but lacks a clear maximum.

We now go on and report on the respective two-dimensional correlated factor model, which we display in Figure 21. We find that the factor loadings of all items, as intended, differ from zero: They range from .41 to .97 for the BL-ST scale and from .36 to .78 for the ENT scale. The model's CFI falls within the acceptable range, and the RMSEA is within the borderline area of .06 to .08. However, the SRMR of .163 in our model does not meet their cutoff value. Computing dynamic cutoffs, we obtained a CFI margin close to .95, and values close to .08 for SRMR and .06 for RMSEA. That indicates that the model fits well in some aspects but not all. Still, we will not discard the model due to a misfit that only concerns one parameter, as we laid out in the theoretical section. After these analyses, we will focus on our last research aim for Study 1: Investigate correlations of figural reasoning (gff) with the PC score. On this, we will report in the following section.

5.3.4 Correlational Analyses with gff

The correlations of the PfPT sub-scales with figural reasoning are depicted in Table 34. As expected, we find a moderate positive correlation for gff with the PC sub-scores. Thus, we can assume that PC and reasoning are related but distinct constructs. Interestingly, the correlation with the ENT sub-scale is higher than with the BL-ST sub-scale, even though the representation of the BL-ST items is closer to the design of the figural items used in the BEFKI (Schroeders et al., 2016).

5.4 Discussion - Study 1

In Study 1, a large theory-driven item set (two sets of 60 items) was developed and piloted. However, a test for primary school children should not be as long as that, as lengthy test sessions tend to tire the children (Jones, Pritchard, Jacobson, Mahone, & Zabel, 2021), and cognitive fatigue can decrease performance (Sievertsen, Gino, & Piovesan, 2016). Therefore, our first research aim of this study was to determine the item difficulties for the initial item set as a basis for sampling from this pool in the next step. We observed a broad and smooth distribution with neither floor nor ceiling effects in both test versions (A and B).

Second, we wanted to pilot the digital test design to determine its feasibility. We found that all children completed the online test. Thus, it is convenient for asynchronous assessments in children. Furthermore, we were able to shorten the test to a duration of 20 minutes in total. This is in line with many authors reporting that online assessments with children as young as that are feasible if researchers pay respect to a child-friendly design, possibility for breaks, and length of the session (Krach, Paskiewicz, & Monk, 2020; McBride et al., 2025).

Our third research aim was to develop a sound measurement model of PC and a corresponding

shorter item set for the PfPT. From a theoretical perspective, we started with items from the categories of BL, ST, and ENT, basic mathematical fields to teach proving (Carl, 2022). Thus, our first model had three pillars of seven items each. However, when applying this model to the pilot data, the correlations between the items in the BL and ST sub-scales were remarkably high. This can be explained by the familiarity of BL and ST in a mathematical sense: All operations with sets can be defined via Boolean expressions. For example, the intersection of sets can be expressed using the Boolean *or*: $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$ (Stoll (2012), p. 12). Thus, we decided to treat BL and ST as one common factor in our model and redo the item sampling procedure. This time, pursuing a two-factor model of 24 items (14 BL-ST, 10 ENT), we applied the Ant Colony Algorithm, a modern item sampling approach for data-driven short scales (Zimny et al., 2024). The resulting test showed good psychometric properties while keeping the advantages of the initial item set.

To investigate the relation between PC and gff, our last research aim targeted the familiarity between the PC sub-scales and the gff scale. We found a weak correlation (r = .30) for the new BL-ST sub-scale, as well as mediocre correlations for the ENT sub-scale (r = .44) and the total score (r = .40). This supports our assumption that PC is related to reasoning. Still, as the correlation did not exceed a medium level, other forms of reasoning (numerical, verbal) should be investigated as well in follow-up studies.

For replication in a more real-world setting, we decided to conduct this validation study (Study 2) in a proctored on-site situation while still using the digital test version. In the following, we will lay out the implementation and outcomes of Study 2, before we sum up the findings of both studies in one, comprehensive discussion.

5.5 Materials and Methods - Study 2

5.5.1 Participants

For Study 2, we aimed to recruit $n = 200$ participants from the age group of 9- to 10-year-old primary school children. A total of $N = 198$ children registered for the study, but only 180 of these children attended a test session and provided their data (N = 180, 36% identified as female, 1% identified as non-binary; M_{age}=9.80, SD= .63). The children were recruited via the HCAP. The statewide coordination of the HCAP granted permission for this study. Children and parents gave their informed consent for participation. As an incentive for participation, all children got a coloring picture and the opportunity to see their test scores. The ethics committee of the faculty of economics and social sciences at the University of Tuebingen gave their approval for the study

(A2.5.4-340_hb, March 20, 2024).

5.5.2 Study Procedure

To recruit participants for the study, we sent out an information sheet to the managing directors of all 69 HCAP sites in Baden-Württemberg, inviting them to contact the study coordinator if they were willing to host one of the test sessions for 15-20 children in their academy. Additionally, we individually contacted managing directors who had previously expressed interest in supporting scientific studies. The academies could then advertise a test session for all interested children from their academy that matched the age group. As soon as an academy announced a test session, all children from the respective place and age group could register. Therefore, their parents received the participant information and an online form to give their informed consent. On the day of the test, the children would meet at their local academy with an instructor supervising the 70-minute test session. Every child sat down at a digital device to complete the first part of the test on it, and after the break, returned to that same device for the second part. The test data was stored immediately and fully anonymously on a server of the University of Tuebingen. At the end of the test, the children could opt in to see their PC and CT scores. They then received a little giveaway and were picked up by their parents.

5.5.3 Measurement Instruments

Computational Thinking For measuring CT, we included the Abbreviated Computational Thinking Test (ACTt) that contains 21 short graphical CT problems in a single-choice setting (Tsarava, 2024). The test is evaluated by awarding one point for every correct item and calculating a sum score.

Fluid Intelligence To evaluate the children's reasoning skills (gfn, gfv, gff), we included the BEFKI, a single-choice fluid intelligence test for third-graders with short scales for gfn and gfv (18 items) as well as gff (8 items) developed by Schroeders et al. (2016). The gff items ask the children to continue a pattern of three given tiles by selecting the next two tiles, each from a stack of three. One point is awarded if the whole pattern is correct. For gfn and gfv, written single-choice tasks are presented. For each test, a sum score is calculated. The value for gf results from the sum of these scores.

Crystallized Intelligence For discriminant validity measures, we applied the test for gc from the BEFKI (Schroeders et al., 2016). It features 16 single-choice questions on declarative knowledge,

evaluated with a sum-score including one point for each question answered correctly (Schroeders et al., 2016).

Demographic Self-Reports Additionally, we followed an exploratory approach to possible demographic dependencies. We had the children report on their gender, age, and latest school grades in German, Mathematics, and Science. To measure their SES, they were asked to indicate (yes = 1, no = 0) which of nine listed things existed at their home (e.g., dishwasher, piano), which then led to a sum-score of up to nine points Leifheit et al. (2020).

Non-Cognitive Constructs For another exploratory approach that regarded non-cognitive constructs, we included 4-step Likert scales on proof attainment value (Ramm et al., 2006), proof interest (Stalder, 2013), general self-efficacy (Beierlein et al., 2012), and proof self-concept (Arens et al., 2011), (Gaspard et al., 2015). All non-cognitive tests are evaluated by calculating means after reversing negatively phrased items. Definitions of these constructs are provided in Section 5.1.3.

5.5.4 Statistical Analyses

In Study 2, we aimed to review, refine, and validate our previous model of PC after rolling out the PfPT in an on-site situation. To do so, we first calculated the item difficulties and inter-item correlations. We visualized the distribution of sum scores among the participants for the PfPT item set emerging from Study 1. We then reapplied the 2-dimensional measurement model suggested by the ACO algorithm in Study 1 and calculated the dynamic fit indices (see: 5.2.4). However, after considering the outcomes of these calculations, we found that the data from Study 2 did not support the two-dimensional model (see: 5.7). Thus, we revisited our initial theoretical approach of a three-fold model. We excluded redundant and overly difficult items and reviewed different item combinations by hand to receive a new three-dimensional model with 15 items. With this model of PC, we evaluated our hypotheses regarding divergent and convergent validity: We calculated the correlations of the test's sub-scales with the constructs of CT, gfn, gfv, gff, gf, and gc. Additionally, we explored the correlations with gender, age, school year, school grades, German spoken at home, and SES sum-score. We also reviewed the test for possible gender biases by considering box plots of the data grouped by gender as well as Welch Two-Sample t-tests for the sub-scales. Lastly, we aimed to explore how PC is related to domain-specific non-cognitive constructs. Therefore, we computed correlations of the PC sub-scores with domain-specific attainment value, interest, self-concept, and general self-concept.

All data, code, and items used in this study are available in a repository of the open science

framework: `https://osf.io/tvarh/`. All calculations and statistical analyses were done using the statistical language R (R Core Team, 2023). We used the package psych for descriptive statistics (William Revelle, 2023), the package lavaan for modeling purposes (Rosseel, 2012), and the package tidyverse for basic data transformation (Wickham et al., 2019).

5.6 Results - Study 2

In the following, we will provide descriptive statistics, item difficulty, and fit indices for the resulting item set from Study 1, answered by the sample from Study 2.

5.6.1 Descriptive Results - Two-Dimensional Model

In Table 35 we deliver the descriptive statistics for all test-scales and self-reports from Study 2. These statistics reveal no obtrusive findings for the scales used. To go into more detail regarding the item set, we show in Figure 22 the distribution of item difficulties for the item set resulting from Study 1 applied to the sample of Study 2.

The item difficulties range from .91 to .08 smoothly, apart from a gap between the difficulties of .4 and .2. This is in line with our expectations for an even distribution. Still, some items have a difficulty that exceeds the guessing probability of .25 for 1-in-4 single-choice items. For the two BL-ST items (ST_dbt_A3, St_dbt_A4), this is not concerning as these are matrix items that have a considerably lower guessing threshold (0.5^9), which they do not undergo. However, the ENT items beyond the threshold can be considered too complex and, therefore, need to be revisited.

To investigate not only the items but also our model of PC, we applied the two-dimensional measurement model from Study 1 to the data from Study 2 to determine if it still holds. We found the factor loadings ranging from .26 to .77 for the BL-ST scale and from .24 to .84 for the ENT scale, matching our expectations. The fit indices, however, did not fully support this model (CFI = 0.735, SRMR = 0.161, RMSEA = 0.055). Therefore, we reconsidered the initial theoretical approach and applied a three-dimensional model to the data from Study 2. In this process, we excluded nine items due to their suboptimal statistical properties. In Table 36, we list these items alongside the data-driven argument for their exclusion and a comment on the possible explanation from a content-oriented perspective. The resulting three-fold model and its statistical properties are described in the next section.

5.6.2 Descriptive Results - Three-Dimensional Model

After further refining the item set and returning to our initial approach with the three sub-scales BL, ST, and ENT, we redid the descriptive calculations for the new set of 15 items. In Table 37,

we provide the statistical properties for these newly composed sub-scales. Again, we do not find any excessive values for skew and kurtosis. Now that we work with a considerably shorter test, we cannot be sure how this aspect influences the former smooth distribution of item difficulties. We therefore depict this distribution for the updated item set in Figure 24.

The item difficulties in the shorter, three-dimensional test range from .91 to .14. Three items lie beyond the single-choice guessing threshold of .25. Again, one of these (ST_dbt_A3) is a matrix item for which the guessing probability is much lower. Remarkably, the overall distribution is less smooth than in the previous test version, but still covers all areas of the spectrum.

As all properties of the item set are in line with our expectations, we will go on and review the three-fold correlated factor model that corresponds with this item set (see: Figure 23).

In the measurement model, the factor loadings range between .27 and. .83. The dynamic cutoff of .042 for RMSEA implies close fit, while the cutoff of .911 for CFI at least implies fair fit. Thus, we decided to keep this three-fold model and use the sampled item set for the PfPT. We used this test version for the correlational analyses, which we will report in the following sections.

5.6.3 Convergent and Divergent Correlations

We computed the correlations between PC and several cognitive constructs to validate our hypotheses regarding convergent and discriminant validity. The results are depicted in Table 38.

5.6.4 Exploratory Correlations

First, we explored the correlations between the PC sub-scores and several demographic aspects, and listed these in Table 39. For the ENT sub-scale, there is a weak positive correlation to the school year, the math grade, and the German grade. The BL sub-scale correlates significantly with math and German grades and native language background, whereas the ST sub-scale does not correlate significantly with any of these aspects. None of the sub-scales correlated significantly with the participants' gender (biserial correlation). Having a closer look at the gender aspect, we also do not find a significant difference in any of the sub-scales (BL: t = -0.391, p = 0.697; ST: t = 0.213, p = 0.832; ENT: t = -1.861, p = 0.065). Box-plots supporting this observation can be found in the supplement (`https://osf.io/tvarh/`). Next, we will turn to the motivational aspects and the correlations regarding these, as depicted in Table 40. We found no significant correlations. Still, among the motivational variables, there are weak to moderate positive correlations. This indicates that the motivational aspects are interwoven, yet not connected to the cognitive construct of PC.

5.7 Discussion - Study 2

In Study 2, the resulting item set from Study 1 was rolled out in a multivariate proctored on-site study. We were able to endorse our results from Study 1 regarding the feasibility of the design.

However, the two-dimensional measurement model resulting from the ACO in Study 1 did not hold. Thus, based on the data from Study 2, we decided to return to our first, theory-driven assumption of a three-factorial structure (BL, ST, ENT). Additionally, we identified several items in Study 2 that were either too hard or redundant. Excluding these items, we settled for a set of 15 items divided into three sub-scales, and computed the descriptive statistics as well as correlations for this. The final test instrument will thus be even less exhausting for the young target group, as the test time can be reduced from over 20 to 15 minutes. This is in line with the demands from research on online assessment (McBride et al., 2025).

Considering the item difficulty for this final item set, we still find a broad distribution. Two items had a difficulty beyond the guessing threshold. However, the distribution of sum scores was close to a normal distribution and did not indicate floor effects, so no further adjustments were made to the item set. The three-dimensional correlated factor model corresponding to the test showed sufficient factor loadings and psychometric properties.

To validate this final PfPT version, we administered the test together with tests on CT (Tsarava, 2024), reasoning (gff, gfn, gfv), and gc (Schroeders et al., 2016). Due to the affinity of reasoning and proving (Bleiler, 2009; Gutierrez & Jaime, 1998; G. Stylianides & Silver, 2007), we expected positive correlations between PC and gff,gfn, and gfv. Indeed, a moderate positive correlation could be found for all three. Furthermore, CT correlated moderately positive with PC. This also matches our expectations, as CT is a form of problem-solving (Shute et al., 2017) and problem-solving is connected to proving (Pólya, 2010). However, the test that we chose for divergent validity - gc - correlated with the PfPT at a moderately positive level as well. This can be explained by the ceiling effects that were observed in the gc test. These are likely to be caused by our sample composition of talented children.

As an exploratory approach, we screened the PfPT for possible correlations with demographic variables. Fortunately, the PfPT does not show signs of discrimination against children with low SES or from a particular age group or gender, matching our demands for a fair test. Further studies could investigate if this assumption holds under consideration of measurement variance. However, the ENT scale shows a weak positive correlation with the school year. This could be caused by the more practice in elementary arithmetics that Year 4 children have compared to Year 3 children (Ministerium für Kultus, 2016a).

Interestingly, we found no association of the PfPT and its sub-scales with the self-reports on domain-specific motivational traits (attainment value, interest, self-concept) or their general self-concept. This contradicts findings from other disciplines in which correlations between outcomes and domain-specific non-cognitive variables were discovered (Hansford & Hattie, 1982).

5.8 Comprehensive Discussion

5.8.1 Key Findings

How can we assess PC in primary school children? To measure PC in primary school children, we developed a test instrument called PfPT. Therefore, we created a large theory-driven set of items on Boolean logic, set theory, and elementary number theory. In Study 1, this initial item set showed a smooth and broad distribution of difficulty. Guided by our theoretical considerations, we assumed a threefold structure of PC and aimed to sample items from this set for a 3-dimensional measurement model. However, all variants of a shorter test with three sub-scales corresponded to a measurement model with extraordinarily high correlations between the BL and ST sub-scales, which can be explained by similarities in the mathematical structure (Stoll, 2012). Therefore, we repeated the item-sampling procedure, this time seeking a two-fold model. Applying the item-sampling algorithm ACO (Zimny et al., 2024), we yielded a two-fold model corresponding to a short test with 24 items and sound psychometric properties.

This version was then rolled out in Study 2 for refinement of the test and measurement model. Surprisingly, in this second study, the model's good psychometric properties were not reproduced. Thus, we could not approve the two-dimensional structure. Instead, the data led us back to our initial approach of a threefold structure. As a consequence, we revised the test once more, resulting in a short version with three sub-scales and 15 items. The threefold model of PC satisfies the thresholds for close fit (RMSEA) and fair fit (CFI) according to McNeish (2018). Therefore, we find that the PfPT in this last version, which corresponds to the threefold model, can be used as a measurement instrument for PC.

Still, future studies are needed to reconsider the psychometric structure of the PfPT and similar tests, as our results from the two studies did not align in all aspects.

How feasible is the digital test setting? In Study 1, we rolled out a quite long initial item set (60 items) in an unsupervised online setting. In this section, we review the feasibility of the test setting. Feasibility, i.e., the question of whether the test can be conducted as planned in this setting, is not to be confused with psychometric adequacy, which is discussed in the previous section. Despite possible concerns regarding the online setting, we found all 409 children finishing the test

session. In addition, the broad difficulty ranking of the items implies that the children answered the test on their own without content-wise parental help. The short test version that we created from the data of Study 1 and implemented in Study 2 had an administration time of 20 minutes. After Study 2, we further refined the test so that the final version can now be administered in only 15 minutes. Considering this aspect as well as the colorful and appealing design, the PfPT is very suitable for our target group of primary school children (Krach et al., 2020; McBride et al., 2025). In Study 2, which was conducted in a proctored on-site situation, again, all children finished the test session. Also, the items used still showed a broad difficulty ranking. This implies that the test is equally feasible in both online and classroom contexts. Yet, there were differences in the outcomes of both studies. Camos, Mariz Elsig, Öncü, Wohlhauser, and Belletier (2025) report that the presence of a proctor can alter the performance of the participants, and McBride et al. (2025) assume that the choice of device makes a difference in online studies. Therefore, future studies should investigate if there are differences between online and on-site administration of the PfPT when the same item set is used.

What are the cognitive and non-cognitive correlates of PC? From a theoretical perspective, the current research offers interesting insights into the nomological network of PC. In Study 1, we found evidence of a positive connection between gff and PC. Therefore, in Study 2, we added additional scales related to reasoning and problem-solving to explore potential connections further. Our results support the findings from Study 1 regarding the overlap of PC and gff. Additionally, we observed high correlations of PC with gfn and gfv. This is in line with our expectations from the literature about the familiarity between proving and reasoning (G. Stylianides, 2008; G. Stylianides & Silver, 2007; Thompson et al., 2012). Also, the overlap between PC and CT supports the hypotheses regarding a connection between proving and problem solving that we derived from Pólya (2010). In follow-up studies, one could go on to measure the correlates if an actual problem-solving test was given instead of one for CT, which is strongly correlated and structurally similar to problem-solving Tsarava et al. (2019).

The correlational results suggest that PC is a distinct cognitive skill, separate from CT, gf, and gc. Contemporary intelligence structure models (like the CHC-model by W. Schneider and McGrew (2012)) strive to depict the full scope of human cognitive ability, to locate all cognitive skills in the model relative to each other, and understand the dependencies. Therefore, further research is needed to determine how PC interacts with different skills and how familiar it is with other forms of reasoning and problem-solving.

Furthermore, we expected an overlap between PC and the domain-specific non-cognitive variables

(self-concept, utility value, intrinsic value), as Hansford and Hattie (1982) found it for the field of mathematics in general and its corresponding beliefs. This was not the case in our investigation. One possible explanation is that the children in our sample had never been introduced to proving before and therefore held quite uninformed and somewhat high beliefs about what it means to them and how good they are at it. Another possibility is that PC is not connected to the self-concept and other domain-specific beliefs.

Finally, we want to address our findings related to demographic factors: Our research showed no gender difference in the PfPT, which is very positive, as usually we find a gender gap in mathematical ability tests (Cheema & Galluzzo, 2013; Reis & Park, 2001). A possible explanation is reported by (Cheema & Galluzzo, 2013): They discovered that the math gender gap disappears when self-concept is taken into account. As noted in the previous section, we found no correlation between PC and self-concept. This could explain why there was no imbalance found regarding PC and gender. Still, when searching for other demographic biases, our results suggest that Year 4 children scored higher in the PfPT than Year 3 children. As discussed in Section 5.7, we believe this is due to their advantages in primary school arithmetic.

Altogether, these findings suggest that the PfPT—particularly in its final, theory-driven threefold structure—offers a promising tool for assessing PC in primary school children. At the same time, the inconsistencies between Study 1 and Study 2 highlight the need for further research with larger and more diverse samples to examine and critically consolidate the underlying psychometric structure.

5.8.2 Limitations

As all studies, the present study is not without limitations, and these need to be considered when interpreting the current findings. The most salient limitation concerns the sampling procedure: All participating children were recruited through an extracurricular enrichment program for which they were nominated by their teachers (Trautwein et al., 2023). As such, the sample likely comprises a disproportionate number of students with above-average cognitive abilities and a heightened interest in academic tasks. This limits the external validity of the results, as the findings may not generalize to the broader population of primary school children, particularly those with average or below-average achievement levels. In future work, it would therefore be valuable to replicate the present analyses with a more heterogeneous and representative sample.

A second limitation relates to the psychometric properties of the developed subscales. The internal consistencies and factor saturations observed for the three intended subscales were low, suggesting that a considerable proportion of variance is attributable to item-specific noise rather than the targeted constructs. Consequently, any substantive interpretation of the subscale-level findings

should be made with caution (e.g.,Taber (2017)), and future revisions of the instrument should prioritize improving the reliability of the dimensions. Finally, it is worth noting that the data were collected in a specific educational and cultural context, which may limit the applicability of the findings to other settings.

Taken together, these limitations highlight the need for continued refinement of the instrument and further validation work based on larger and more diverse samples. At the same time, it is essential to acknowledge that the study's primary contribution mitigates these limitations. To date, no validated instrument has been available for assessing proof-related problem-solving at the primary school level. In this light, the present work represents a valuable initial step by providing a theoretically grounded and transparently documented measurement approach that can serve as a basis for subsequent development and validation efforts.

5.8.3 Desiderata for Future Research

Given the demands from educational practice for an earlier integration of proof in the school mathematics curriculum (Ball & Bass, 2003; Bass, 2011; Bleiler, 2009), one of the main motivations behind this work was the open question of how effective proof interventions are for primary school children. Future studies can now utilize the PfPT for pre- and post-assessments in intervention studies aimed at promoting PC among primary school children.

We developed the PfPT optimizing for time efficiency, low cognitive load, and age-appropriateness, yielding a test that not only satisfies these demands but is also feasible for administering online or in a supervised on-site setting. Future research may explore whether test outcomes differ across these two settings, potentially shedding light on the influence of environmental factors on PC assessment. Furthermore, it would be valuable to explore whether the results remain consistent when the stimuli in the items are replaced with more child-contextualized materials, as demonstrated by Goecke et al. (2024) for an assessment of working memory capacity.

As our two studies used different item sets, the question of test-retest reliability remains subject to future research. Furthermore, given the observed correlations between PC and self-reported math grade, as well as the higher school year, administering the PfPT alongside an age-appropriate standardized mathematics achievement test (e.g., DEMAT 3+ (Roick, Gölitz, & Hasselhorn, 2018)) could provide deeper insights into the relation between PC and general mathematical ability. Investigating additional characteristics connected to reasoning and mathematics, like the need for cognition, could further expand the nomological net of PC: Considering the results of Jonsson, Mossegård, Lithner, and Karlsson Wirebring (2022) that the need for cognition pertains to creative mathematical reasoning but not algorithmic mathematical reasoning, it would be intriguing to find

where PC fits within this framework.

5.8.4 Conclusion

In conclusion, the PfPT represents a feasible, age-appropriate, and theoretically grounded instrument for assessing proof-related problem solving in primary school children. The present study provides an important first step by developing and documenting a short, child-friendly, digital test that can be administered efficiently in both online and on-site settings. At the same time, the current version of the instrument shows some psychometric limitations, particularly concerning the reliability and distinctiveness of the subscales. These limitations should be considered when interpreting results, and further work is needed to improve the measurement precision and to validate the test in more diverse and representative samples. Nevertheless, the PfPT offers a valuable foundation for advancing research on proof-related competencies at an earlier stage of education, and it can serve as a starting point for continued refinement and broader application in educational research.

5.9 Appendix B

5.9.1 Tables

Table 31

Existing tests for PC

Authors	Construct	α	Target Group	Question Type	Content
Mejía–Ramos et al. (2017)	proof comprehension	.70	university students	multiple choice	3 proofs
Healy and Hoyles (2000)	proof comprehension	NA	pupils (age 14-15)	mixed	2 conjectures
Kempen and Biehler (2014)	proof construction	NA	university students	open	1 proof
Senk (1989)	proof construction	.85	pupils (Year 7-12)	open	6 items

Table 32

Group A, M_{age}=9.26, SD = .65; N=210, 41% identified as female

sub-scale	M	SD	Median	Min	Max	Skew	Kurtosis
Set Theory (ST)	11,39	3,99	12,00	0,00	20,00	-0,36	-0,06
Boolean Logic (BL)	14,73	3,07	15,00	1,00	19,00	-1,35	2,85
El. Number Theory (ENT)	10,94	3,80	11,00	3,00	19,00	0,15	-0,62

Table 33

Group B , M_{age}=9.29, SD = .69; N=199, 44% identified as female

sub-scale	M	SD	Median	Min	Max	Skew	Kurtosis
Set Theory (ST)	12,24	4,66	12,00	0,00	20,00	-0,32	-0,46
Boolean Logic (BL)	15,50	3,15	16,00	0,00	19,00	-2,13	6,18
El. Number Theory (ENT)	8,78	4,50	9,00	0,00	19,00	0,29	-0,64

Table 34

Correlations of the PfPT and its sub-scales with gff

Variable	*M*	*SD*	1	2
1. BL-ST sub-scale	9.16	2.54		
2. ENT sub-scale	5.32	2.42	.66** [.57, .73]	
3. gff	11.42	3.27	.30** [.09, .48]	.44** [.25, .60]

Table 35

Descriptive statistics for demographics, motivational variables, and test scores from Study 2.

	N	M	SD	Median	Min	Max	Skew	Kurtosis
Gender	178	1,63	0,48	2,00	1,00	2,00	-0,56	-1,70
Age	180	9,80	0,63	9,86	8,29	10,87	-0,37	-0,88
School Year	180	3,56	0,50	4,00	3,00	4,00	-0,22	-1,96
Math Grade	177	1,27	0,49	1,00	1,00	3,00	1,57	1,52
German Grade	178	1,51	0,59	1,00	1,00	4,00	0,86	0,58
Science Grade	159	1,48	0,57	1,00	1,00	3,00	0,69	-0,55
Native Speakers	180	1,48	0,82	1,00	1,00	4,00	1,50	1,10
SES	180	5,48	1,23	6,00	2,00	9,00	0,16	0,06
Attainment Value	180	3,39	0,57	3,33	2,00	4,00	-0,86	-0,03
Domain Specific Interest	180	3,19	0,72	3,33	0,00	4,00	-1,08	1,70
General Self Concept	180	3,41	0,52	3,33	0,00	4,00	-1,54	7,96
Self Concept Domain	180	3,49	0,55	3,67	1,17	4,00	-1,38	2,09
Computational Thinking	180	13,68	4,29	14,00	0,00	22,00	-0,28	-0,21
Numerical Reasoning	180	6,83	1,82	7,00	0,00	9,00	-0,86	0,34
Verbal Reasoning	180	6,97	1,73	7,00	0,00	10,00	-0,93	0,96
Figural Reasoning	180	10,11	3,28	11,00	0,00	15,00	-0,68	-0,39
Fluid Intelligence	180	7,97	1,82	8,33	0,00	10,67	-0,84	1,07
Crystallized Intelligence	180	11,80	2,73	12,00	0,00	16,00	-1,08	1,94
Proof sub-scale BLST	180	8,61	2,48	9,00	2,00	14,00	-0,33	-0,35
Proof sub-scale ENT	180	2,71	2,00	3,00	0,00	10,00	0,78	0,71

Note. Gender dummy coded: 1= identified as female, 2=identified as male.

Table 36

Excluded items with exclusion criteria

Item	Exclusion Criterion	Comment on Content
ENT_bas_A_4	low rit, very difficult, no significant factor loading	contains more text than other items
ENT_bas_A_8	content redundancy	similar to ENT_bas_AB_4
ENT_bas_AB_4	content redundancy	similar to ENT_bas_AB_8
ENT_tri_A_4	low item discrimination	requires careful reading
ENT_bas_AB_1	content redundancy, model fit deterioration	similar to ENT_bas_AB_2
BL_veb_A_1	content redundancy	similar to BL_veb_A2 and BL_veb_A3
BL_has_A_2	content redundancy	similar to BL_has_A1
ST_dbt_A2	item difficulty	combination of several concepts
ST_gle_A_2	content redundancy	similar to ST_gle_A1

Table 37

Descriptive statistics for PC scores from Study 2.

	N	M	SD	Median	Min	Max	Skew	Kurtosis
Proof sub-scale ST	180	2,49	1,38	2,00	0,00	5,00	0,02	-0,80
Proof sub-scale BL	180	3,51	1,15	4,00	0,00	5,00	-0,80	0,32
Proof sub-scale ENT	180	1,57	1,30	1,00	0,00	5,00	0,44	-0,64

Note. For demographics, motivational variables, and other test scores, see Table 35.

Table 38

Correlations: Test scores and PC scores from Study 2

Variable	M	SD	1	2	3	4	5	6	7	8	
1. Computational Thinking	13.68	4.29	$\alpha = .76$								
2. Numerical Reasoning	6.83	1.82	.51** [.40, .61]	$\alpha = .59$							
3. Verbal Reasoning	6.97	1.73	.53** [.41, .62]	.60** [.50, .69]	$\alpha = .55$						
4. Figural Reasoning	10.11	3.28	.57** [.46, .66]	.37** [.24, .49]	.40** [.27, .52]	$\alpha = .66$					
5. Fluid Intelligence	7.97	1.82	.68** [.59, .75]	.75** [.68, .81]	.76** [.69, .81]	.85** [.81, .89]	$\alpha = .64$				
6. Crystallized Intelligence	11.80	2.73	.35** [.22, .47]	.46** [.34, .57]	.39** [.26, .50]	.20** [.06, .34]	.40** [.27, .51]	$\alpha = .66$			
7. Proof sub-scale ST	2.49	1.38	.35** [.22, .47]	.29** [.15, .42]	.38** [.24, .49]	.41** [.28, .52]	.46** [.33, .57]	.21** [.07, .35]	$\alpha = .53$		
8. Proof sub-scale BL	3.51	1.14	.39** [.26, .51]	.32** [.18, .45]	.35** [.22, .47]	.30** [.17, .43]	.40** [.27, .52]	.30** [.17, .43]	.22** [.08, .36]	$\alpha = .40$	
9. Proof sub-scale ENT	1.57	1.30	.37** [.24, .49]	.39** [.25, .50]	.23** [.09, .37]	.13 [-.02, .27]	.28** [.14, .41]	.32** [.18, .45]	.13 [-.01, .27]	.28** [.14, .41]	$\alpha = .52$

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates p <.05. ** indicates p <.01. Cronbach's α indicates internal consistency.

Table 39

Correlations: Demographic aspects and PC scores from Study 2.

Variable	M	SD	1	2	3	4	5	6	7	8	9	10
1. Gender	1.63	0.48	SI									
2. Age	9.80	0.63	.05 [-.10, .19]	SI								
3. Schoolyear	3.56	0.50	.05 [-.10, .20]	.77** [.70, .82]	SI							
4. Grade Math	1.27	0.49	-.03 [-.18, .12]	.02 [-.12, .17]	.02 [-.13, .17]	SI						
5. Grade German	1.51	0.59	.24** [.09, .37]	.02 [-.13, .17]	.05 [-.10, .19]	.31** [.17, .43]	SI					
6. Grade Science	1.48	0.57	.07 [-.09, .23]	-.10 [-.26, .05]	-.12 [-.27, .04]	.07 [-.08, .23]	.30** [.15, .44]	SI				
7. Native Speaker	1.48	0.82	.15* [.00, .29]	-.03 [-.17, .12]	.06 [-.08, .21]	-.11 [-.25, .04]	.08 [-.06, .23]	.02 [-.14, .17]	SI			
8. SES	5.48	1.23	.02 [-.13, .16]	-.01 [-.16, .13]	.05 [-.10, .19]	-.04 [-.19, .11]	-.03 [-.18, .11]	.02 [-.14, .17]	.14 [-.01, .28]	SI		
9. Proof sub-scale ENT	1.57	1.30	.13 [-.01, .28]	.03 [-.12, .17]	.23** [.09, .36]	-.26** [-.39, -.11]	-.25** [-.38, -.10]	-.15 [-.30, .01]	.02 [-.12, .17]	.05 [-.10, .19]	$\alpha = .52$	
10. Proof sub-scale BL	3.51	1.14	.03 [-.12, .18]	.09 [-.06, .23]	.11 [-.03, .25]	-.16* [-.30, -.02]	-.23** [-.36, -.08]	-.06 [-.21, .10]	.18* [.03, .32]	.03 [-.12, .17]	.28** [.14, .41]	$\alpha = .40$
11. Proof sub-scale ST	2.49	1.38	-.02 [-.16, .13]	.06 [-.09, .20]	.06 [-.09, .20]	-.12 [-.26, .03]	-.11 [-.25, .04]	-.06 [-.22, .09]	-.00 [-.15, .14]	-.01 [-.16, .13]	.13 [-.01, .27]	.22** [.08, .36] $\alpha = .53$

Note. German school grades range from 1=very good to 6=insufficient. Therefore, positive correlations have a negative sign. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates p <.05. ** indicates p <.01. Cronbach's α indicates internal consistency. SI=single items, Gender dummy coded: 1=identified as female, 2=identified as male.

Table 40

Correlations: Motivational aspects and PC scores from Study 2

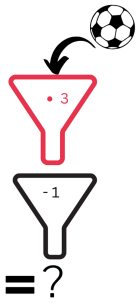
Variable	*M*	*SD*	1	2	3	4	5	6
1. Attainment Value	3.39	0.57	$\alpha = .67$					
2. Domain-Specific Interest	3.19	0.72	.41** [.28, .53]	$\alpha = .93$				
3.General Self-Concept	3.41	0.52	.25** [.11, .38]	.25** [.11, .38]	$\alpha = .71$			
4. Domain-Specific Self-Concept	3.49	0.55	.37** [.24, .49]	.40** [.27, .52]	.32** [.19, .45]	$\alpha = .85$		
5. Proof sub-scale ENT	1.57	1.30	.04 [-.11, .19]	.14 [-.01, .28]	.11 [-.04, .25]	.09 [-.06, .23]	$\alpha = .52$	
6. Proof sub-scale BL	3.51	1.14	-.07 [-.22, .07]	-.06 [-.21, .08]	-.01 [-.16, .13]	-.02 [-.17, .12]	.28** [.14, .41]	$\alpha = .40$
7. Proof sub-scale ST	2.49	1.38	-.03 [-.17, .12]	.01 [-.14, .15]	.02 [-.13, .17]	-.03 [-.17, .12]	.13 [-.01, .27]	.22** [.08, .36] $\alpha = .53$

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates p <.05. ** indicates p <.01. Cronbach's α indicates internal consistency.

5.9.2 Figures

Figure 16

Example Item ENT sub-scale



A	The funnels turn **even** numbers into **odd** numbers.
B	The funnels turn **odd** numbers into **odd** numbers.
C	The funnels turn **even** numbers into **even** numbers.
D	None of the above.

Figure 17

Example Item BL sub-scale

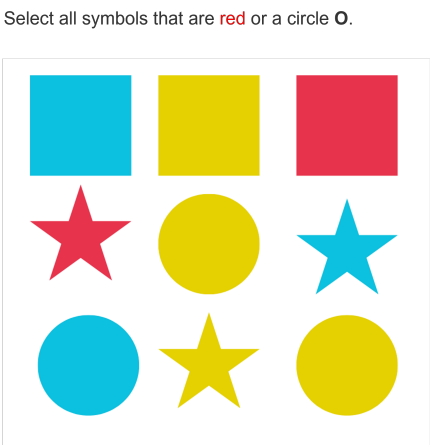


Figure 18

Item difficulty for the resulting item set

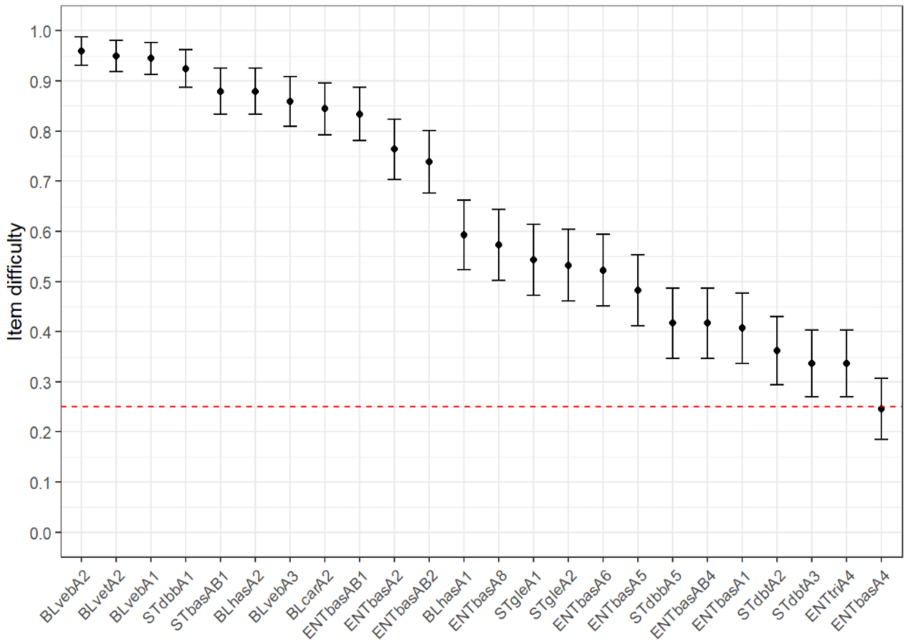


Figure 19

Distribution of scores BL-ST sub-scale

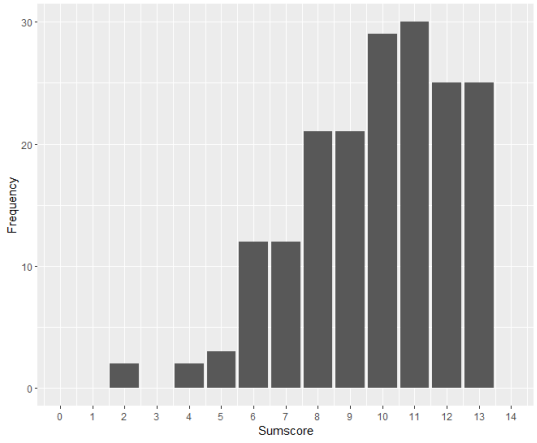


Figure 20

Distribution of scores ENT sub-scale

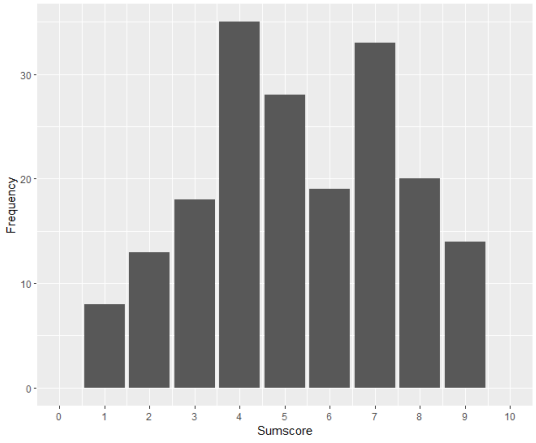


Figure 21

Correlated Factor Model of the two PC sub-scales. n = 199; $\chi^2(250) = 498.07$, CFI = .968, RMSEA = .071, close dynamic fit, SRMR = .163. $\omega_{BL-ST} = .786$, $\omega_{ENT} = .838$.

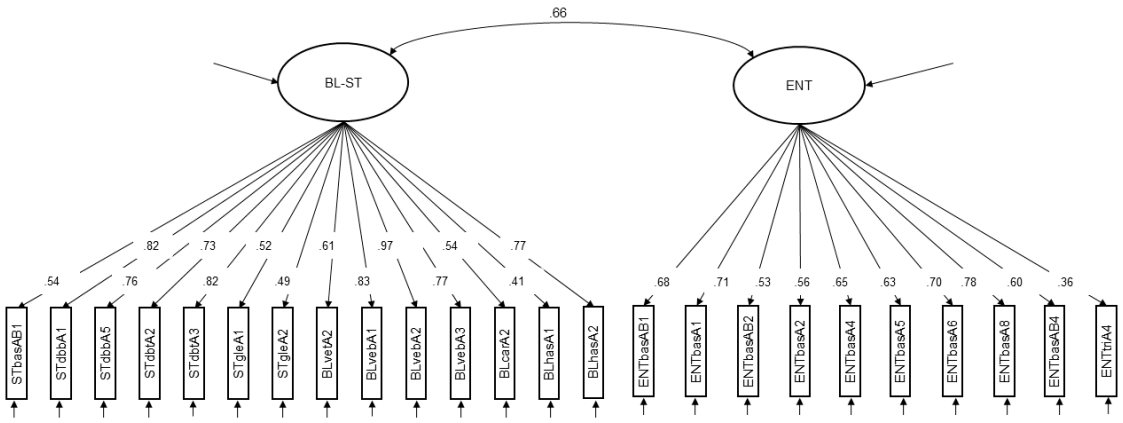


Figure 22

Item difficulty Study 2. red line: guessing threshold for single choice items. Blue line: guessing threshold for matrix items.

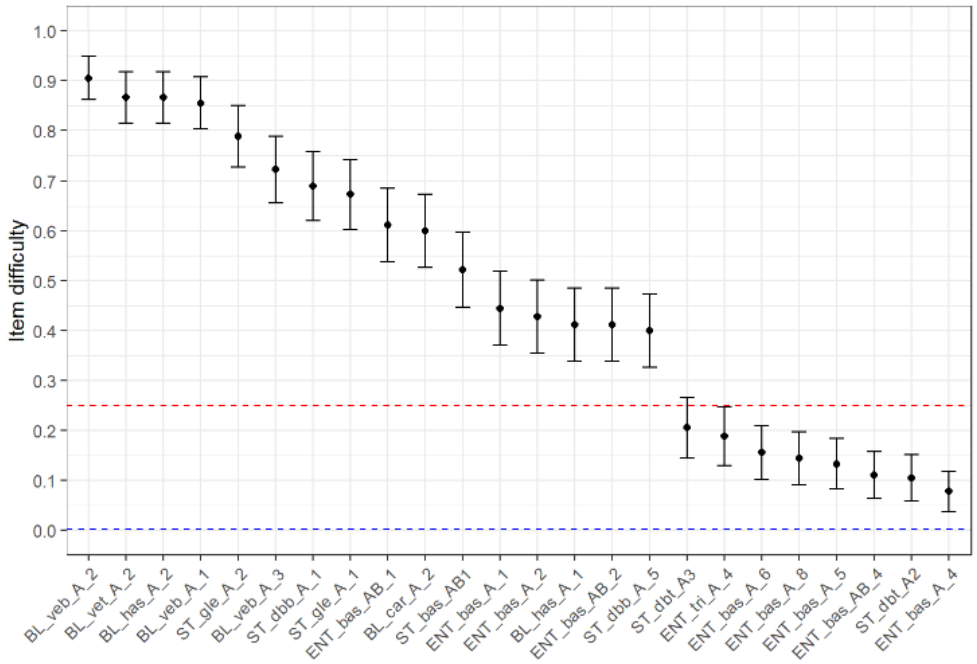


Figure 23

3-dimensional measurement model from Study 2. n = 180 $\chi^2(87)$*=99.01, CFI = .94, RMSEA = .028, SRMR = .11,* $\omega_{ST} = .529, \omega_{BL} = .421, \omega_{ENT} = .520$.

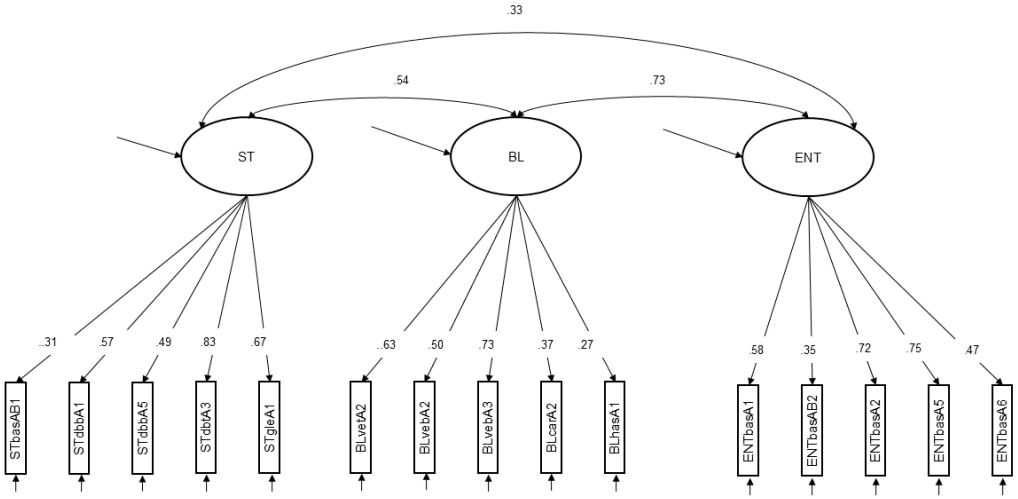
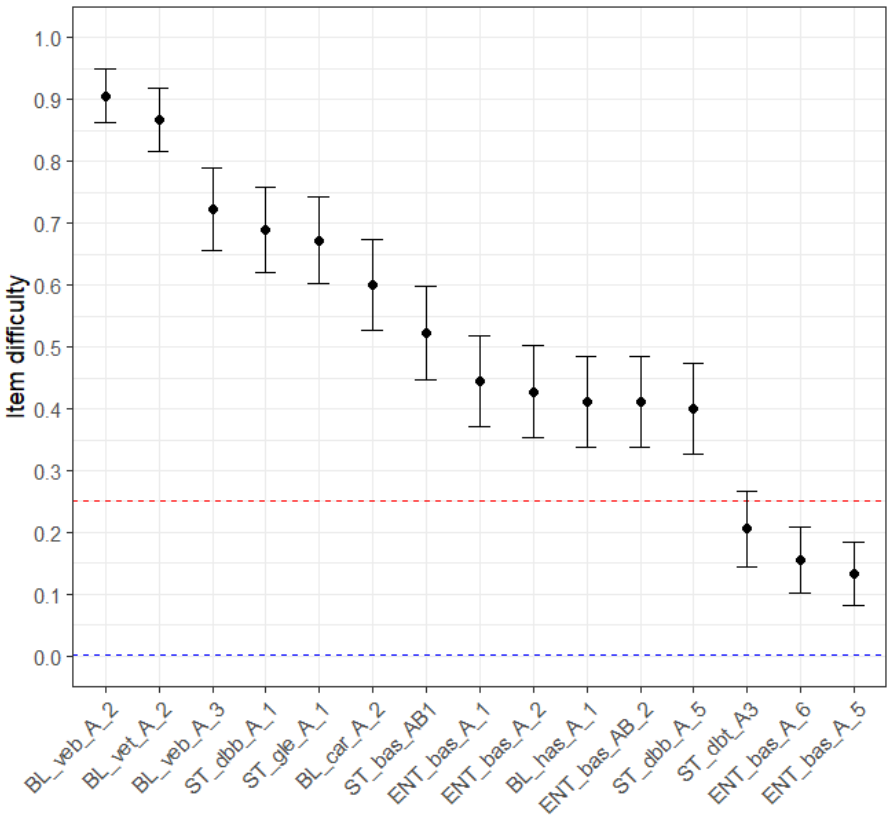


Figure 24

Item difficulty Study 2 (three-dimensional model). Red line: guessing threshold for single-choice items. Blue line: guessing threshold for matrix items.



6 STUDY 3: A SELF-PACED ONLINE COURSE TO INTRODUCE MATHEMATICAL PROOF TO TALENTED PRIMARY SCHOOL CHILDREN: A RANDOMIZED CONTROLLED TRIAL

The content of this chapter is in preparation to be submitted to Computers and Education Open. The proportional contributions of the (co-)authors to the manuscript are presented in the subsequent table. This article may not exactly replicate the final version published in the journal. It is not the copy of record.

Author	Author position	Scientific ideas (%)	Data generation (%)	Analsysis & interpretation (%)	Study writing (%)
Xenia Stein	first	80	100	70	70
Armin Fabian	second	10	0	20	20
Jessika Golle	third	10	0	10	10

Abstract

Gifted children need special cognitive challenges. Enrichment courses can provide such challenges by teaching extracurricular content, such as mathematical proving. Therefore, we designed and evaluated an asynchronous online proof course *Logical Detectives (LD)* for gifted primary school children. It contains lessons on Boolean Logic, Set Theory, and Elementary Number Theory, as well as exercises in reasoning and logic. For the present report, we measured the effects of this intervention in a randomized controlled field trial (N=269 children). We found no intervention effects on proof competency, nor proof interest, intrinsic value, or attainment value for proving. Our results indicate adverse intervention effects on students' proof self-concept, suggesting that proof learning in an asynchronous online course may pose too many challenges for primary school students, even those who are intellectually gifted. The results will be discussed in light of current theories on gifted education and digital learning.

Keywords: mathematical proof, primary school, self-paced course, motivation, domain-specific self-concept, RCT

6.1 Introduction

To solve contemporary global and individual problems, mathematical abilities are a crucial tool (OECD, 2019; Winter, 1995). One of the most central mathematical abilities is mathematical proving (Dawkins & Weber, 2017; Rav, 1999; Schoenfeld, 2009). However, proof education only plays a minor role in most school curricula around the world (e.g., in German federal states Ministerium für Kultus, 2016a, 2016b). This is a possible reason why students transitioning from high school to university often lack rigorous proof techniques and experience it as particularly challenging (Carl, 2022; Glosauer, 2019). Therefore, some practitioners argue for an early introduction of mathematical reasoning and proving (Cervantes-Barraza, Hernandez Moreno, & Rumsey, 2020; Nunes et al., 2007; A. Stylianides, 2007). Still, others regard it as too challenging for young students- especially in primary school - to present their arguments in an accurate formal language, arguing that the failure of the New Mathematics shows the unsuitability of rigorous proving in primary school (Dreyfus, 1999; Kline, 1973). However, for gifted and talented students, challenging content is necessary to delve deeper into unfamiliar fields and explore them (Renzulli, 2016; Subotnik et al., 2011).

Mathematically gifted children seem to be specifically noteworthy in the context of mathematical proving, given that their mathematical reasoning skills on average develop earlier than those of their peers (Moya Pérez, Gutierrez, & Jaime, 2015) and that the cognitive and mathematical processes involved in a formal proof align well with the traits of mathematical giftedness (Stein, Tsarava, Fabian, et al., 2025). Studying the extent to which mathematical proof can be effectively taught to gifted primary school children can serve as a test run for a broader integration of the topic in primary school: If learning effects are achieved, this can serve as a proof of concept. If real-world studies with gifted primary school children reveal that the content is too challenging, this can serve as a starting point for revising teaching materials, as these will likely also be too challenging for average students. Various studies from South Korea on mathematical proving with gifted primary school students led to promising results (Chang et al., 2006; Ko & Song, 2011; K. H. Lee, 2005; Na, 2011). However, these studies only involved small samples and no standardized test instruments or control groups.

Asynchronous online courses have great potential for serving larger groups of children. Furthermore, these courses have several advantages for gifted education, such as flexibility and cost-effectiveness (Weaver et al., 2022), that can outweigh challenges like organizational matters, tutoring fees, or unavailability in rural areas, which would otherwise hinder many children from participating in fostering programs (Lohaus & Wild, 2021). Additionally, these courses enable differentiation and self-paced learning, which can be particularly beneficial for gifted children, but are often challenging

to implement in traditional physical classrooms (Leikin, 2021). Lastly, asynchronous online learning is especially suitable for challenging tasks (Tullis & Benjamin, 2011) and therefore seems a suitable environment for a proof course. Thus, we decided to conduct a large-scale intervention study using the asynchronous online proof course *Logical Detectives* for gifted primary school students, which had previously been piloted (Stein, Tsarava, Fabian, et al., 2025), to investigate the effects on the proof competency of gifted primary school students. In line with the state of the art in intervention studies (Lendrum & Wigelsworth, 2013), we carried out the research in a randomized controlled trial (RCT). For additional insights, we explored motivational effects and included demographic variables as covariates and in additional interaction models.

Taking these interdisciplinary perspectives into account, we decided to investigate the efficacy of an asynchronous online proof intervention for gifted primary school students. Therefore, we conducted a large-scale study, carried out as a randomized controlled trial, which is discussed in this report.

6.2 Theoretical Background

6.2.1 Mathematical Proving and Proof Competency

Mathematical proof has been described as the *soul of mathematics* (Schoenfeld, 2009), stressing its central role in the scientific discipline of mathematics. According to Rota (1997), proof can be defined as

> ... a sequence of steps which leads to the desired conclusion. The rules to be followed by such a sequence of steps were made explicit when logic was formalized early in this [the 20th] century, and they have not changed since. [p.34]

This definition can be regarded as a scientific proof definition. For a classroom definition of proof, independent of the class's formal mathematics experience, A. Stylianides (2007) suggested the following:

> 1. It [a proof] uses statements accepted by the classroom community (set of accepted statements) that are true and available without further justification;
> 2. It employs forms of reasoning (modes of argumentation) that are valid and known to, or within the conceptual reach of, the classroom community;
> 3. It is communicated with forms of expression (modes of argument representation) that are appropriate and known to, or within the conceptual reach of, the classroom community. [p.291]

As these definitions allow for much space in between, approaches to teaching mathematical proving differ in the extent to which they reduce the formalism and diverge from a professional mathematical proof. To classify these differences, Biehler and Kempen (2016) developed a system based on four aspects in question: 1. Are formal symbols replaced? By what kind of representation? 2. If formalism is reduced, how is the generality of a proof ensured? 3. How does the approach discuss differences compared to a formal mathematical proof? 4. What is the target group and learning goal? This system will be relevant for describing and comparing existing courses in Section 6.2.5.

To be effective, a proof intervention should foster commensurable competences related to proving. Waluyo et al. (2019) suggests that there are two main competences related to dealing with mathematical proofs: proof comprehension and proof construction (Waluyo et al., 2019). In the present investigation, we will only refer to proof construction with the term proof competency (PC). K. Lee (2016) summarizes PC as "[...] the process of constructing mathematical assertions to determine the largest of mathematical objects for which the mathematical proposition is true or false through the search for possible examples and counterexamples"(p.28). This definition of PC is what we have in mind when we speak of fostering children's proof competency in the intervention reported here.

To measure this competence, several researchers have presented test instruments. For example, the CDASSG Proof Test by Senk (1989), which contains six items (four complete proofs and two short-answer questions) and has an administration time of 35 minutes ($\alpha = .85$). As these tests require that one can already write full proofs, they do not seem suitable to describe the developing proof skills of a primary school student. Therefore, in the current study, we use the Preformal Proving Test for primary school children, which we developed and validated in previous studies (Stein, Tsarava, & Goecke, 2025).

6.2.2 Proof Integration in School Curricula

In the second half of the previous century, the New Mathematics curriculum was introduced in many Western countries to teach proof skills and algebraic reasoning in primary and secondary schools (Moon, 1986). As this curriculum did not produce the desired results and was very unpopular among parents and teachers (Kline, 1973), proving was excluded from the curricula in most countries and still is (Hanna & Knipping, 2020).

For example, the German primary school curriculum emphasizes the competence of mathematical reasoning, but not the ability to prove (Ministerium für Kultus, 2016a). In the academic school track, children encounter basic geometrical proofs, but not how to construct and write rigorous algebraic proofs (Ministerium für Kultus, 2016b). In addition, the curriculum does not require that

these proofs be completed at a specific formal level (as distinguished by, e.g., Wittmann & Müller, 1988). However, PC is one of the first competencies necessary to acquire in university studies of any STEM subject (Glosauer, 2019). This transition can be particularly challenging, given the absence of prior proof practice (Kempen, 2019). Therefore, A. Stylianides (2007) argues that children should be exposed to mathematical proving from primary school on.

Nevertheless, many educators consider proving to be a very challenging task and argue against implementing this topic in primary schools (Bass, 2011). Gifted children, however, often lack challenging mathematical tasks in school (Rotigel & Fello, 2016). Thus, for these children, it may be beneficial to learn formal proof at the primary education level. Furthermore, due to their advanced capabilities, this target group could act as a vanguard and help identify aspects of proof learning that are generally too challenging for their age group. These insights could support the development of proof-related exercises for all primary school children.

6.2.3 Gifted Primary School Children and Mathematical Proving

Looking at models for the process of proving (Boero, 1999) and the attributes of mathematically gifted students (Krutetskii, 1976), we find several intersections (for a thorough discussion, see: Stein, Tsarava, Fabian, et al., 2025). For example, *formalized perception* is presumed to be one possible characteristic of mathematically gifted students (Krutetskii, 1976). It supports *formulation in line with mathematical standards*, which is necessary for the proof-writing process (Boero, 1999). Thus, these children may have an advantage in acquiring PC. Findings from the practice support this assumption: Moya Pérez et al. (2015) found that mathematically talented secondary school students in their study were likely to write more correct and more deductive proofs than their peers. One can speculate that this assumption also holds for gifted primary school students. Furthermore, enrichment courses can be a great possibility for (mathematically) gifted children (Käpnick, 1998; Subotnik et al., 2011). Such courses offer additional, extracurricular content for students to develop and follow their interests (Subotnik et al., 2011; Ziegler, 2008). For mathematical enrichment, topics such as mathematical reasoning and problem-solving are very common (Bardy & Bardy, 2020; Käpnick, 1998). Proving is mentioned as an enrichment content in some models (e.g., Bardy & Bardy, 2020), but is less prominent than problem-solving and reasoning. Given the familiarity of reasoning and proving (Bleiler, 2009; G. Stylianides, 2008; G. Stylianides & Silver, 2007) as well as the necessity of proving for problem-solving (Pólya, 2010), the idea of proving as a designated enrichment content is strongly supported.

In summary, the process of proving could be more doable and rewarding for mathematically talented learners than for their peers, as it allows for the application of their skills (Leikin,

2010). In a study with gifted primary school students, Ko and Song (2011) implemented an intervention including pre-algebraic proof activities, which introduce proof-writing playfully. They made promising observations related to the children's PC. Similar results were found in other studies with gifted primary school classes in South Korea (Chang et al., 2006; K. H. Lee, 2005; Na, 2011): For example, K. H. Lee (2005) taught a gifted class geometrical proof through proof-like activities.

Nevertheless, the aforementioned studies have primarily focused on geometrical proofs and have not analyzed learning effects using large samples and RCT designs. However, such approaches are needed to determine the efficacy of an intervention (Humphrey et al., 2016; Nelson et al., 2012) and, in our case, to determine whether PC can be promoted in primary school children. To do so, we decided to conduct a study with a scalable proof intervention for gifted primary school children.

6.2.4 Advantages of Asynchronous Online Courses

Asynchronous online courses come with many advantages, which can be beneficial for conducting large-scale interventions, gifted students, and the process of proof learning:

First, scalability, geographical freedom, and cost-effectiveness that come with the absence of a physical classroom (Weaver et al., 2022), allow for a large number of participants. Such large sample sizes not only strengthen an intervention study but also promote educational fairness. A more diverse audience in terms of socioeconomic status and location can participate in the course (Weaver et al., 2022). As a consequence, enrichment can become accessible to a more diverse group of gifted children. This provides an opportunity for children with low SES, who are often overlooked in enrichment programs (Rothenbusch, Zettler, Voss, Lösch, & Trautwein, 2016).

Second, flexibility and convenience allow for the integration of asynchronous online courses into almost every child's schedule (Lin & Gao, 2020; Weaver et al., 2022). This aspect contributes to the aforementioned aspect of scalability and also helps children choose courses based on content rather than time preferences. Exploring new content and pursuing resulting passions are key parts of enrichment for gifted children (Renzulli, 2016).

Furthermore, asynchronous online courses enable self-paced learning (Lin & Gao, 2020). In this setting, learners can decide what to review and how much time to spend on each task, leading to more effective learning sessions and being especially helpful for challenging tasks (Tullis & Benjamin, 2011). Since proving is generally seen as a difficult topic, self-paced learning could be an ideal context to acquire PC. Self-paced learning also enhances learning autonomy, which is a key need for gifted students during the learning process (Brink, 2025; Leikin, 2021).

Lastly, in asynchronous online courses, educators can smoothly embed automatic real-time

feedback. Such feedback can enhance practice motivation, improve the ability to review one's performance, enhance self-concept, and lead to better post-training performance (J. Schneider et al., 2016). In the context of mathematical proof, this may help overcome common formal mistakes and misconceptions in one's written proofs (Carl, 2022).

6.2.5 Asynchronous Online Tools for Proof Learning

To identify possible intervention content for the current study, we reviewed existing asynchronous online tools that teach skills related to mathematical proving. Our main focus was their suitability for primary school children with no prior knowledge of formal mathematics.

In the Early Algebra jump-and-run game *Curse Reverse*, the player can steer an archaeologist through temples and caves and help her overcome obstacles by working with algebraic terms (Morales & Torres, 2020). This way, children can get used to working with unknowns in a playful environment. However, it does not teach the learner how to arrange arguments into a proof.

In contrast, the tool *Toyproofs*, developed by Monks and Carter (2014), teaches children the structure of a proof. In several colorful mini-games, the learner transforms a line of symbols by applying rules until it matches the displayed pattern. These exercises can be regarded as an excellent preparation for proof-writing, but do not contain a transfer of the proof-structure to actual algebraic content.

Other tools like *The Incredible Proof Machine* (Breitner, 2016) or QED (Tao, 2018) offer the possibility to practice linking arguments with algebraic symbols. Nevertheless, the algebraic symbols used in the exercises are not introduced before. In the case of QED, the user interface is likely designed for an older audience, as it features extensive text and complex navigation.

In summary, we found no online tool that provides a comprehensive introduction to mathematical proving, suitable for young proof novices. Furthermore, all these approaches differ from a full mathematical proof, either by iconic instead of formal representation or by the means of generality. According to (Biehler & Kempen, 2016), an approach that teaches proving should also contain a comparison between the content of this approach and actual formal proving. Therefore, we developed a novel course that takes these aspects into account to teach PC to gifted primary school children on an iconic and formal level, including an introduction to formal mathematical symbols from Boolean Logic, Set Theory, and Elementary Number Theory. This intervention was piloted in a previous study (Stein, Tsarava, Fabian, et al., 2025) and subsequently refined. In the following Section, we will report on the content and design of this course.

6.2.6 The Course 'Logical Detectives'

The asynchronous online course 'Logical Detectives' is a self-paced online training for promoting PC in gifted children of grades 3 and 4. The course promotes formal notation and argumentation, which are crucial for proof-writing (Boero, 1999) and are particularly suitable for mathematically gifted children (Krutetskii, 1976).

The Course's Core Components In line with Nelson et al.'s research on intervention implementation, the activities for the course were derived from a number of core components, specifically selected for the target group and intervention goals. Two core components address the act of proving:

Iconic and symbolic logical reasoning. Transitioning from iconic to symbolic logical reasoning can help to develop mathematical proof skills in Boolean Logic and Set Theory, two disciplines that both have their own mathematical symbols (Bronkhorst et al., 2021). This transition corresponds well to the actions of formalization and reasoning mentioned as relevant steps in Boero's (1999) model of proof-writing.

Natural language proof writing. This term refers to writing mathematical proofs in a controlled fragment of the language that humans use to read and write (Carl, 2022). In contrast to formal mathematical language, it is easier for human readers to make sense of such a proof-text. At the same time, machines can still verify natural language proofs, as these only consist of a limited number of possible words (Carl et al., 2022). Natural language proof writing is closely connected to the remaining two core components, which regard the asynchronous online setting:

Automated real-time feedback. It refers to immediate responses that a computer provides right after the learner enters an answer. This kind of feedback is beneficial, as receiving instant feedback greatly increases its utility for the learner (Mory, 2004). Thus, automated real-time feedback is likely to improve online courses. It can help to increase the learners' motivation to practice, their self-reflection, and their performance (J. Schneider et al., 2016).

Self-paced learning. Here, learners decide on their own study time and repeat tasks as they prefer (Tullis & Benjamin, 2011). This format can be beneficial in online courses and increase performance, especially on difficult tasks (Tullis & Benjamin, 2011). Furthermore, gifted students are more likely to engage in self-paced learning due to better self-regulation strategies (Risemberg & Zimmerman, 1992).

Course Content On a content level, we chose the mathematical context of Elementary Number Theory, Boolean logic, and Set Theory, as these are typical topics for novice proof learning that

are usually not taught in school (Glosauer, 2019; Grieser, 2017). As a consequence, we devoted one chapter of the self-paced course to each of these pillars. Additionally, there was one chapter introducing the children to proving itself and one for making them familiar with the asynchronous learning format to reduce extraneous cognitive load, which can be an obstacle in online courses (Hollender et al., 2010). The initial structure of the course was as follows: Chapter 0 - Introduction to Asynchronous Online Learning; Chapter 1 - Introduction to Mathematical Proving; Chapter 2 - Boolean Logic Proofs; Chapter 3 - Set Theory Proofs; Chapter 4 - Elementary Number Theory Proofs. Each chapter followed the structure of a lesson as suggested by Meyer and Junghans (2021), starting with an introduction to an idea or problem, followed by a working phase in which learners acquired skills or knowledge, and closing with documentation to secure these skills and exercises for repetition. Table 41 depicts the structure of the course chapters and how they correspond to the phases of proof learning.

Table 41

Structure of the Chapters

Element	**Objective**	**Phase**
Preview	Advanced organizer of the chapter	Introduction
Exploration	Getting familiar with words and symbols of the chapter	Working on formal knowledge
Case	Find statements and express solutions with symbols	Working on reasoning skills
Proof	Complete a given proof with blanks	Documentation
Notebook	Look up symbols of the chapter	Scaffolding for the exercises
Proof Collection	Practice writing natural-language proofs	Exercises

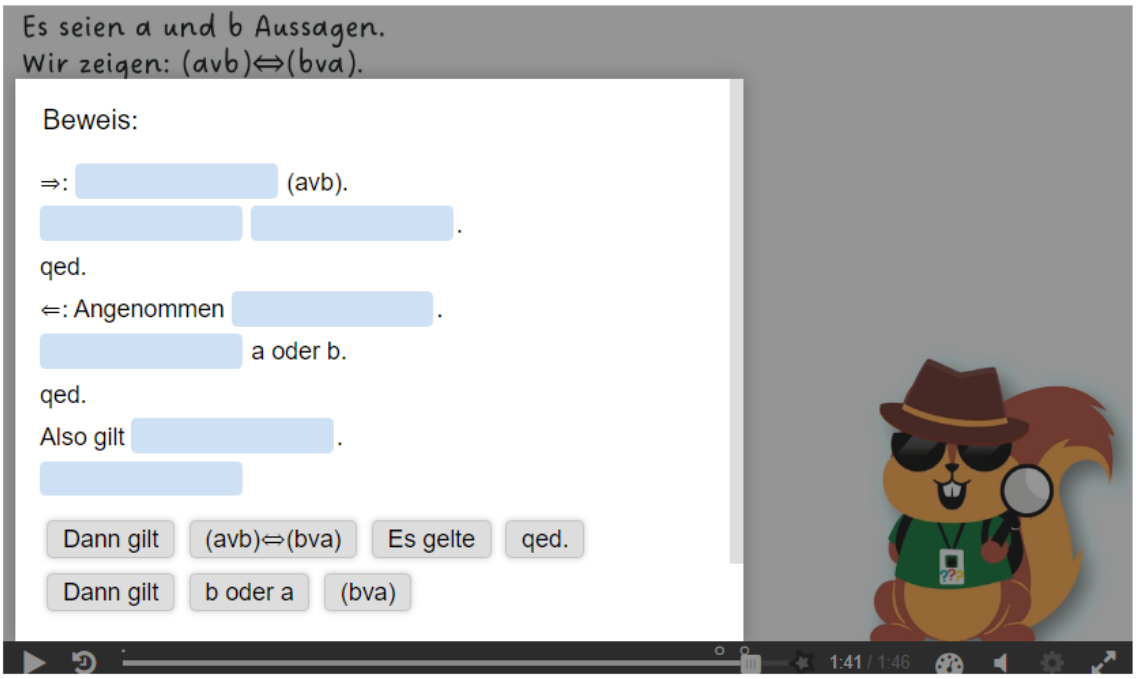
Additionally, we included games, riddles, and problem-solving tasks in the chapters to maintain the children's motivation, as these challenging activities are likely to be rewarding for mathematically gifted children (Bardy & Bardy, 2020; Käpnick, 1998).

To provide the children with automated real-time feedback in every exercise, we implemented the course elements using two technical features: Most course elements were delivered as H5P content within a Moodle LMS. H5P is a plugin that allows creating interactive videos, presentations, and other learning media with instant evaluation and feedback prompts that are displayed to the learner. We utilized this infrastructure to create a self-paced learning landscape featuring pre-defined hints and feedback prompts tailored to common errors in proof learning. For the natural language proof exercises, we adapted the university math drill-and-practice system *Diproche* (Carl, 2022). This system allows for instant individual feedback on natural language proofs composed by the learner. Proof-novices in university perceived this as motivating and beneficial Carl et al. (2022). We altered

the exercises to make them solvable without secondary school mathematical knowledge. We also revised the design and feedback tone to be more child-appropriate and added several help buttons and instruction pages. Screenshots of exemplary tasks are provided in Figures 25 and 26.

Figure 25

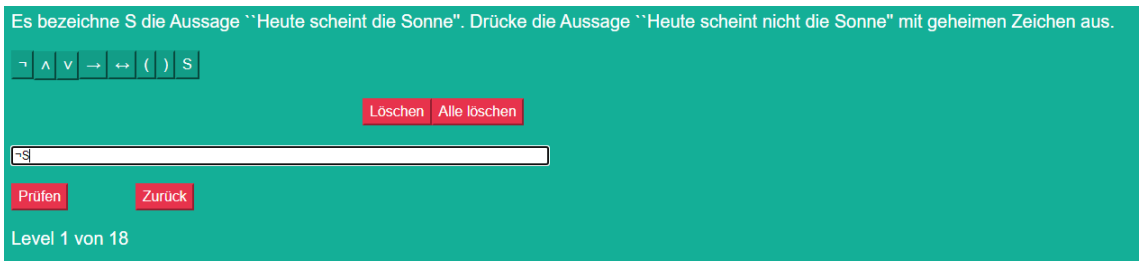
Screenshot of the Proof Video in Chapter 2



Translation: *Let a and b be statements. We are proving: (a∨b) ⇔ (b∨a)* Stein, Tsarava, Fabian, et al. (2025)

Figure 26

Screenshot of the Game Secret Symbols in Chapter 2



Translation: *Let S be the statement "The sun is shining today". Express the statement "The sun is not shining today" with secret symbols.* Stein, Tsarava, Fabian, et al. (2025)

We depict all activities and their correspondence to the core components in the course booklet available at `https://osf.io/qf6by`.

Pilot Study and Revision In a six-week pilot study with N=304 children from grades 3 and 4, the setting and topic of 'Logical Detectives' were investigated for their feasibility (Stein, Tsarava,

Fabian, et al., 2025): The course turned out to be feasible and appealing for most of the children, and especially the reasoning tasks and the interactive environment were very popular. Also, the children reported no lack of social interaction. Nevertheless, the study found that the course participants would appreciate more feedback, additional instructions, and scaffolding, as well as changes to the user interface, allowing them to complete all tasks using a computer mouse and no keyboard.

As a consequence, the course was revised after the pilot study to increase the match with the target group: Most changes address our adaptation of the natural language proof checker *Diproche*. We replaced the typing input fields in the proof exercises with on-click buttons that print the necessary words and symbols. Additionally, we provided more constructive feedback on the use of logic and the structure of the proof. Furthermore, we replaced some exercises with easier ones. As a broad share of participants wished for more examples, instructions, and solutions, we implemented two help buttons: one to print a strategic help statement based on the current progress and one to glimpse at the solution. In front of every task, we integrated a screen-cast showing the completion of an example task. At the end of each series of proof exercises, the children now receive another digital reward in the Moodle course. Regarding the Moodle course, the structure was adapted: To make navigation easier, we developed a graphical icon for each activity and a more compact layout. Chapter 4 (Elementary Number Theory) was perceived as more feasible and simpler by the participants; thus, we decided to place it between Chapter 1 (Introduction to Proof) and the former Chapter 2 (Boolean Logic). To make the logical symbols more appealing to children, we included more illustrations and short quizzes in the *Explorations* that accompany them. The activity *Notebook*, a glossary for looking up logical symbols, was removed due to concerns over feasibility and popularity. Instead, we included all definitions as hovering messages in the on-click buttons in our adaptation of *Diproche*.

Still, we do not know if this revised course has a positive effect on PC and how it influences domain-specific motivational variables. This encouraged us to investigate the course effects in a controlled setting, as we describe in the following section.

6.2.7 The Present Study

The present study aimed to investigate whether an asynchronous online course can effectively foster proof competency in gifted primary school children. To this end, we preregistered one confirmatory research question in the Registry of Efficacy and Effectiveness Studies (#20667.1v1):

RQ1 (preregistered): What is the effect of attending the course compared to not attending the course on children's proof competency? We hypothesized a positive treatment effect, expecting that participation in the course would improve proof competency relative to a control condition.

In addition, we investigated the following exploratory research questions:

RQ2: Does course participation influence children's domain-specific motivational beliefs, including self-concept, interest, attainment value, utility value, and persistence?

RQ3: Do effects on proof competency or motivational outcomes differ by age, gender, or baseline performance (only for PC)?

6.3 Methods

6.3.1 Sample

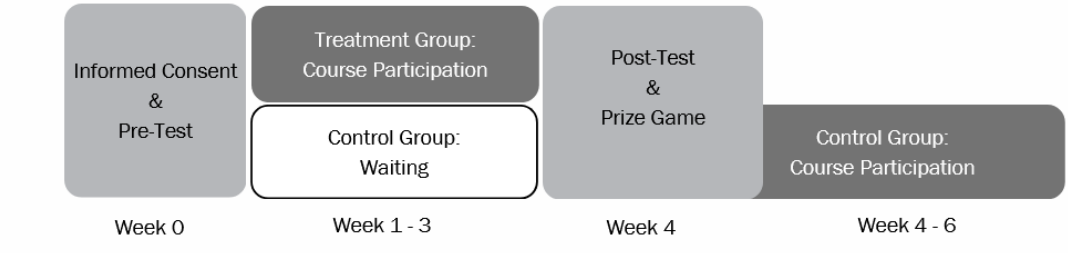
Data for this study were collected from 269 children (M_{age} = 9.59, SD_{age} = .59; 107 female) participating in the Hector Children's Academy Program (HCAP). The HCAP is a statewide STEM enrichment program for talented primary school children with 70 sites across Baden-Württemberg (Germany) and an additional online course program based on Moodle. To participate in the enrichment, children have to be nominated by their teachers. After that, they can freely choose from the course program. A detailed description of the HCAP course program and its scientific monitoring is provided in Trautwein et al. (2023).

6.3.2 Procedure

To draw sound conclusions about the efficacy of *Logical Detectives*, we implemented the present study using a randomized waiting-control-group design and monitored the outcomes with the Preformal Proving Test (PfPT), which had been developed and validated in two previous studies by Stein, Tsarava, and Goecke (2025). The exact procedure is depicted in Figure 27.

Figure 27

Illustration of the study procedure

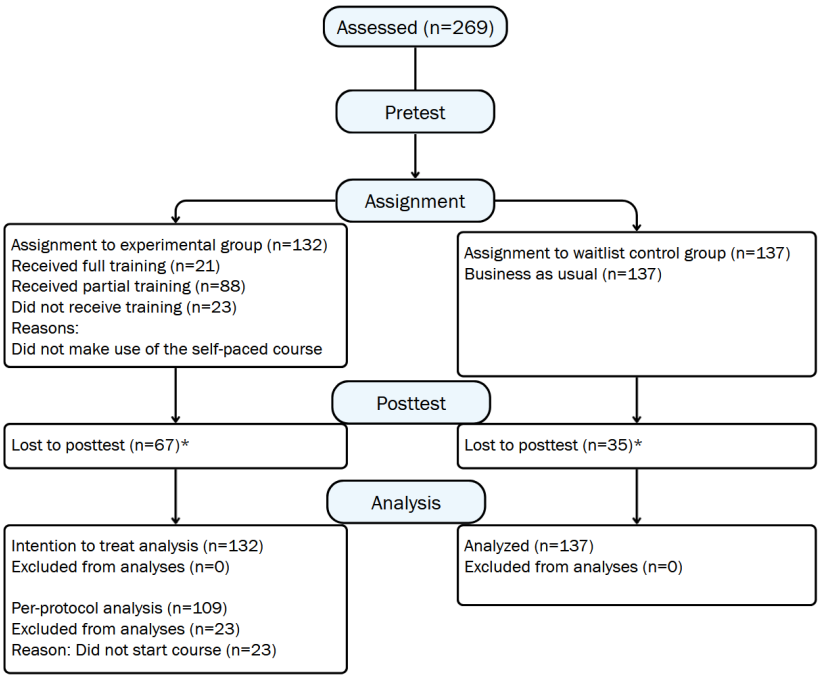


Study participants were recruited via the online summer holiday program of the HCAP. Here, children and parents were informed about the procedure in written form and could then provide

their informed consent for participation. Additionally, HCAP mascot Hasel explained the procedure to the children in age-appropriate videos with easy sentences and supporting illustrations. After they had registered for the study, the children completed an online pre-test via the survey tool unipark to measure their proof competency and several motivational constructs (see: Section 6.3.3). At the end of the test, the children were assigned to either the treatment group or the wait-list control group using a random variable generated within the survey tool. The treatment group was immediately given access to the asynchronous online course on mathematical proving (see: Section 6.2.6) and could practice with it as often as they liked within the following three weeks. At the end of this period, the children from both groups were asked to retake the online test. After completing this post-test, the children from the wait-list control group were given access to the course. As an incentive for both groups, all children who had completed both tests were eligible to participate in a raffle for bookstore vouchers. In Figure 28, the randomization and treatment participation are displayed.

Figure 28

Overview of participation in the data-collection.



Note. *According to our preregistration, we did not exclude participants with missing post-tests.

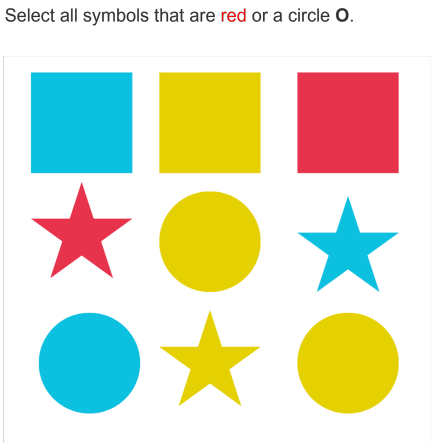
6.3.3 Measures

Proof Competency For the assessment of proof competency in the pre- and post-tests, we used the Preformal Proving Test (PfPT), which we developed and validated in two previous studies in

2024 (Stein, Tsarava, & Goecke, 2025). The test is made up of 24 items, administered remotely on a computer. To complete it, the children were given a 20-minute time limit. Items that the participant did not reach were coded as 'incorrect'. In Figure 29, an example item is depicted.

Figure 29

PfPT Example Item Stein, Tsarava, and Goecke (2025)



Motivational Constructs To assess motivation before and after the course, we asked the children to report on their domain-specific academic self-concept (3 items, adapted from Arens et al. (2011) and Gaspard et al. (2015) - example item: *Everything regarding mathematical reasoning comes easy to me.*), interest (3 items, adapted from Stalder (2013) - example item: *Everything regarding mathematical reasoning is interesting to me.*), attainment value (3 items, adapted from Ramm et al. (2006) - example item: *I do not care about mathematical reasoning.*) and utility value belief (1 item, self-generated - example item: *Mathematical reasoning will help me with math in school.*). Furthermore, we included scales of general self-efficacy (3 items, adapted from Beierlein et al. (2012) - example item: *In difficult situations I can rely on my skills.*) and persistence (3 items, self-generated - example item: *When I have the choice, I always pick the difficult math exercises.*). All items were answered on a 4-step Likert scale ranging from *I do not agree at all*(=1) to *I fully agree*(=4).

Demographic Self-Reports In addition to the cognitive and motivational measures, we asked the participants to report their gender (0 = female, 1 = male), age (years, months), grade (3 or 4), as well as school grades in math, German, and science (1 = very good, 6 = insufficient).

6.3.4 Analyses

The primary outcome variable was proof competency. Secondary outcomes were students' domain-specific motivational beliefs, namely self-concept, interest, attainment value, utility value, and persistence. All pre-test and outcome variables were z-standardized before analysis. The primary predictor was treatment condition (0 = control, 1 = intervention). According to our preregistration (Registry of Efficacy and Effectiveness Studies, #20667.1v1), we controlled for the participants' reported gender and pre-test score. We decided to alter the initial approach and include the participants' age as a predictor instead of their school grade, as we did not survey how many of the talented children in our sample had been accelerated in their school career (earlier enrollment, grade skipping). Thus, the age variable seemed more reliable to us.

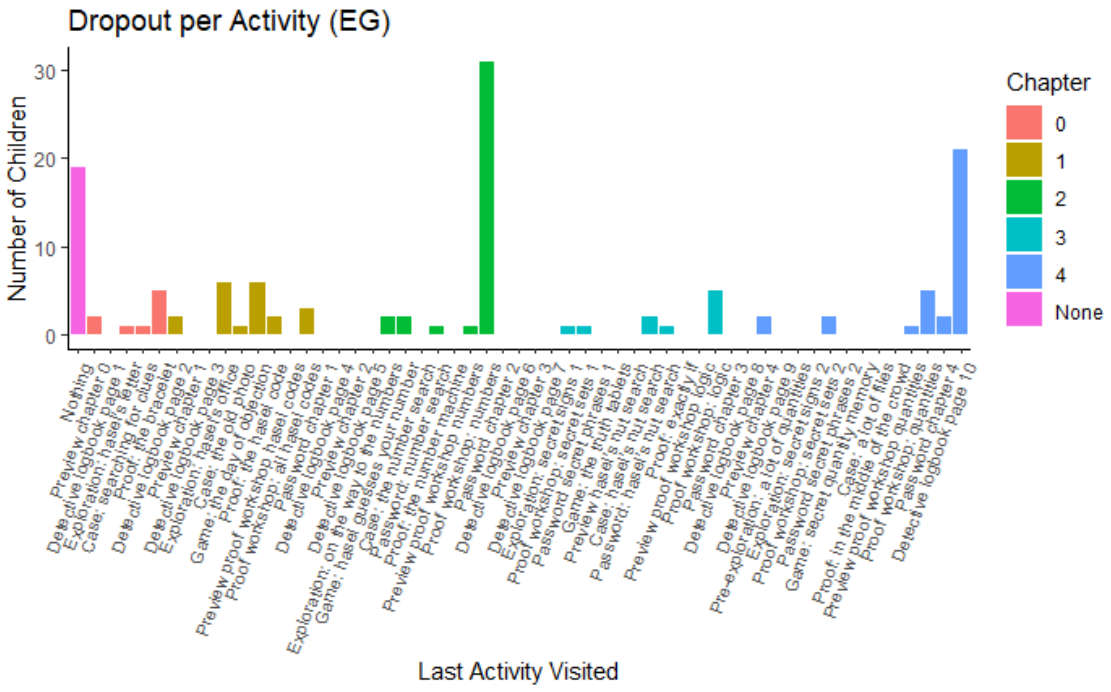
In the first step, we computed descriptive statistics for all study variables. Second, we evaluated intervention effects using structural equation modeling (SEM) in R, employing the lavaan package (Rosseel, 2012). In these models, each post-test outcome was regressed on treatment condition (primary predictor), the respective pre-test score (baseline control), age, and gender (further control variables). In the next step, we calculated two additional models for each dependent variable, regarding the interaction effects of treatment with age and gender. Additionally, for our primary outcome, PC, we investigated a model that included the interaction between pre-test score and treatment. These calculations, which examined effects and interactions, were first carried out, including all participants (intention-to-treat analyses). They were then repeated, excluding participants who did not start the course (per-protocol analyses). As a robustness check, we reran all calculations a second time, excluding participants who did not complete the course.

This approach allowed us to test the preregistered hypothesis that the intervention increased the participants' PC and to explore potential effects on motivational outcomes and possible interaction effects. Finally, as another robustness check, we conducted a censored variable analysis in Mplus to predict the motivational outcomes, accounting for ceiling effects observed in the children's motivational self-reports.

Missing Data As the children were only assigned to a group at the end of the pre-test, only children who completed the pre-test were assessed for the study. Additionally, some children did not start the course after completing the pre-test (non-compliers), and some did not follow the course to the end. This resulted in missing data in the log data and the post-test. Figure 28 shows how many children were missing from which test. Figures 30 illustrate the progress made by the children in the course. Following the preregistration, we treated missing data with the full-information-maximum-likelihood method (FIML) when computing the SEM.

Figure 30

Number of Children Stopping per Activity - Experimental Group



6.4 Results

6.4.1 Descriptive Results

Histograms depicting the distribution of PC scores and motivational self-reports at both time points are provided in the appendix (Section 6.7.1). Tables 42 and 43 show the descriptive statistics of all variables assessed in the pre- and post-test.

6.4.2 Effects on Proof Competency

As depicted in Figure 28, some children who had been assigned to the treatment group did not participate in the training in the first place. Therefore, we report on the analyses two times: Once for the whole group (intention-to-treat analysis, $n_{treat} = 132, n_{cont} = 137$) and once with the non-compliers excluded (per-protocol analysis, $n_{treat} = 109, n_{cont} = 137$, see: Table 44 and Table 45).

Table 42

Descriptive Statistics for Pre-test Values of Domain Specific Motivational Constructs and Proof Competency

Construct		T1				
		N	**M**	**SD**	**Min**	**Max**
Gender	CG	137	0,58	0,49	0,00	1,00
	EG	131	0,62	0,49	0,00	1,00
Age	CG	137	9,56	0,57	7,92	11,09
	EG	132	9,62	0,61	7,84	11,34
School Year	CG	137	3,27	0,45	3,00	4,00
	EG	132	3,23	0,42	3,00	4,00
Math Grade	CG	136	1,29	0,47	1,00	3,00
	EG	130	1,28	0,49	1,00	3,00
German Grade	CG	136	1,53	0,54	1,00	3,00
	EG	130	1,5	0,56	1,00	3,00
Science Grade	CG	127	1,31	0,56	1,00	3,00
	EG	121	1,4	0,57	1,00	4,00
Proof competency	CG	137	9,25	2,83	1,00	15,00
	TG	132	9,45	2,51	3,00	15,00
Utility value	CG	137	3,61	0,70	1,00	4,00
	TG	132	3,51	0,73	1,00	4,00
Attainment Value	CG	137	3,42	0,58	1,67	4,00
	TG	132	3,29	0,66	1,33	4,00
Self concept	CG	137	3,55	0,45	2,00	4,00
	TG	132	3,48	0,46	2,00	4,00
Domain specific interest	CG	137	3,41	0,59	1,00	4,00
	TG	132	3,24	0,70	1,00	4,00
Persistence	CG	137	3,61	0,43	2,33	4,00
	TG	132	3,46	0,52	1,33	4,00

Note. TG = Training group, CG = Control group; gender was dummy-coded: 0= female, 1= male

Table 44

Average Causal Effects (intention-to-treat) on proof competency

Effect	β	SE	p
Intervention	-0.042	0.122	0.734
Gender	0.062	0.123	0.612
Age	0.112	0.067	0.095
Pretest	0.597	0.061	<0.001
R^2		0.403	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 43

Descriptive Statistics for Pre-test Values of Domain Specific Motivational Constructs and Proof Competency

Construct		T2				
		N	**M**	**SD**	**Min**	**Max**
Proof competency	CG	102	10,45	2,59	3,00	15,00
	EG	64	10,45	2,78	4,00	15,00
Utility value	CG	102	3,58	0,64	1,00	4,00
	EG	65	3,48	0,75	1,00	4,00
Attainment Value	CG	102	3,35	0,57	1,33	4,00
	EG	65	3,35	0,64	2,00	4,00
Self concept	CG	103	3,49	0,38	2,40	4,00
	EG	65	3,36	0,37	2,60	4,00
Domain specific interest	CG	102	3,3	0,62	1,00	4,00
	EG	65	3,14	0,71	1,33	4,00
Persistence	CG	102	3,57	0,48	2,00	4,00
	EG	65	3,51	0,48	2,33	4,00

Note. TG = Training group, CG = Control group; gender was dummy-coded: 0= female, 1= male; demographics and school grades were only assessed at T1.

Table 45

Average Causal Effects (per-protocol) on proof competency

Effect	β	SE	p
Intervention	0.008	0.128	0.948
Gender	0.051	0.126	0.684
Age	0.090	0.069	0.191
Pretest	0.609	0.063	<0.001
R^2		0.408	

Note. Pre-test and dependent variable were standardized before the analyses.

A third model, limited to children who completed the course, is presented in Table 48 in the appendix. The pattern of results in the three evaluations was comparable: For proof competency, all three analyses showed no significant treatment effect.

6.4.3 Motivational Effects

As we did in the previous section for PC, we now present intention-to-treat analyses and per-protocol analyses for all assessed motivational variables (see Tables 46 and 47).

Table 46

Average Causal Effects (intention-to-treat) on domain-specific motivational outcomes

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention	-0.292	0.140	0.037	-0.145	0.127	0.252	-0.032	0.126	0.801	-0.149	0.130	0.249	-0.010	0.115	0.934
Gender	-0.274	0.140	0.051	-0.065	0.128	0.613	0.069	0.128	0.587	0.230	0.130	0.078	-0.069	0.115	0.553
Age	-0.084	0.074	0.257	0.008	0.070	0.912	-0.039	0.068	0.563	0.028	0.069	0.688	-0.023	0.062	0.715
Pre-test	0.404	0.067	<0.001	0.661	0.072	<0.001	0.623	0.066	<0.001	0.613	0.069	<0.001	0.712	0.060	<0.001
R^2		.216			.415			.386			.366			.494	

Note. Pre-test and dependent variables were standardized before the analyses.

Table 47

Average Causal Effects (per-protocol) on domain-specific motivational outcomes

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention	-0.345	0.145	0.018	-0.170	0.129	0.188	-0.134	0.130	0.303	-0.213	0.137	0.120	-0.034	0.123	0.783
Gender	-0.242	0.143	0.090	-0.012	0.128	0.928	0.035	0.130	0.785	0.239	0.136	0.078	-0.080	0.121	0.508
Age	-0.101	0.075	0.182	<0.001	0.069	0.996	-0.031	0.069	0.654	0.028	0.072	0.692	-0.012	0.065	0.858
Pre-test	0.402	0.068	<0.001	0.642	0.071	<0.001	0.638	0.067	<0.001	0.597	0.071	<0.001	0.711	0.062	<0.001
R^2		.226			.418			.399			.350			.484	

Note. Pre-test and dependent variables were standardized before the analyses.

The respective analysis, limited to the finishing children, can be found in the appendix (Table 49). For the domain-specific self-concept, we found a significant adverse treatment effect when all participants were considered ($\beta_{\text{Intervention}} = -0.292, p = .037$, see Table 46). These effects increased and remained significant in the per-protocol analysis ($\beta_{\text{Intervention}} = -0.345, p = .018$, see Table 47). In the analyses for only the finishing children (see: Table 49) a similar effect was found for self-concept ($\beta_{\text{Intervention}} = -0.610, p = .006$) and additionally for attainment value ($\beta_{\text{Intervention}} = -0.557, p = .002$) and utility value ($\beta_{\text{Intervention}} = -0.539, p = .006$). No intervention effects were observed in the analyses of domain-specific interest or persistence.

6.4.4 Interaction Effects

To explore possible interaction effects related to our main outcome, PC, we investigated the interaction between gender or age and the intervention. For both of these independent variables, the interaction models are provided in the appendix (Tables 50 to 52 and 56 to 58). However, none of these analyses yielded any significant interaction effects. For PC, we also investigated the interaction of its pre-test score with the intervention, as shown in Tables 62 to 64 in the appendix. The pattern of results was similar for all three versions of the calculations, and no significant interaction effect was found.

To investigate interaction effects on motivational beliefs, we also calculated models examining

the interactions of gender and age with the treatment, with the results presented in the appendix (Tables 53 to 55 and Tables 59 to 61). For the gender variable, no significant interaction effect was observed.

The pattern of results regarding the interaction of age and intervention was comparable to the analyses regarding the treatment effects: In the intention-to-treat analysis, the interaction effect was not significant (see Table 59), but in the per-protocol analysis, the interaction of intervention and age showed a significantly negative effect on the students' proof self-concept ($\beta_{\text{Intervention:Age}} = -0.332, p = .036$) and interest ($\beta_{\text{Intervention:Age}} = -0.288, p = .034$, see: Table 60). However, in the analysis, which only included the course finishers, no significant interaction effect regarding the age was observed (see Table 61).

6.4.5 Predictors of Dropout

Considering the slightly different results in the previous analyses for children who did and did not finish the course, we decided to take a closer look at possible predictors of course dropout. Therefore, we calculated a logistic regression model to predict the likelihood of dropping out, using all pre-test variables as independent variables and including gender and age as control variables. This model is presented in Table 65. Additionally, a variant of this model with group assignment as another independent variable was calculated (Table 66). Still, no significant predictors of dropout could be determined with either model.

6.5 Discussion

In this study, we examined the efficacy of an asynchronous online course on mathematical proving for primary school children in grades 3 and 4. Therefore, a randomized controlled trial with a waiting control group was implemented to monitor the effects of the intervention on mathematical proof competency and domain-specific motivational variables.

6.5.1 Key findings

Our first research question was whether attending the course had a significant effect on the children's proof competency. We hypothesized a positive treatment effect, expecting that participation in the course would improve proof competency relative to a control condition. However, we did not detect significant training effects regarding the participants' proof competency compared to the control group.

The second research question addressed the children's domain-specific motivational beliefs, namely, proof self-concept, interest, attainment value, utility value, and persistence. In the intention-to-treat

and per-protocol analyses, we found significant negative effects on the training group's proof self-concept, while for all other motivational variables, no significant effects were detected. The finishers-only analysis supported these findings and suggested that such effects also pertained for attainment value and utility value.

Lastly, we asked whether the effects on proof competency or motivational beliefs significantly differed by age, gender, or, in the case of PC, baseline performance. For PC, neither age, gender, nor baseline performance was found to interact significantly with course effects. Regarding the motivational effects, no significant differences were found by gender; however, the per-protocol analysis revealed a negative interaction effect between age and treatment on students' proof self-concept and interest. For the remaining motivational variables, no significant interaction effects were observed.

6.5.2 Implications and Future Directions

Previous research suggested that mathematical proving can be taught in primary school (Cervantes-Barraza et al., 2020; Zaslavsky et al., 2011) and that the act of proving (Boero, 1999) is well compatible with the characteristics of mathematically gifted children (Krutetskii, 1976). Still, based on the results of the present study, we were not able to show that the investigated proof course significantly increased the children's proof skills. We will present two considerations to explain this dissent: Firstly, mathematical proof can be viewed as a process with two parts: one related to reasoning and one related to formalism (Selden, 2013). Although the reasoning part was very popular among the children in the pilot study (Stein, Tsarava, Fabian, et al., 2025) and reasoning is generally considered suitable for enrichment at all ages (Käpnick, 1998), the formal part could still be too difficult to understand for most talented primary school children. Secondly, talented children may have more potential than their peers to excel in proof tasks, but not necessarily more interest in the related activities, at least not at the primary school age. Further studies could explore these hypotheses in more detail to reveal which parts of mathematical proving are suitable for enriching gifted primary school students.

To explain the negative effect on the children's proof self-concept, various aspects can be found in the literature: The experience of encountering the new subject of proving might have reduced their self-concept as the unusual tasks challenged them. At the same time, the course might also have provided a clearer picture of what proving is, which then led to a more realistic self-report (Dunning-Kruger-Effect, Kruger & Dunning, 1999). Especially, for gifted children in enrichment settings, the Big-Fish-Little-Pond Effect (H. W. Marsh, 1987) can be additionally demotivating, when they suddenly do not experience themselves as protruding or outclassing their peers (Niederer,

2017; Zeidner & Schleyer, 1999). Furthermore, Jasmin and Ongcoy (2024) identified several aspects in mathematics online courses that can additionally decrease the mathematical self-concept, such as feelings of isolation or domestic distractions. Further studies could help to address the decrease in the children's self-concept by identifying which of these aspects occur most frequently in the context of the LD course.

The aspects of age, gender, and pre-test scores, which we considered possible predictors of course effects and dropout, did not show a significant influence on these. Thus, further characteristics of the participants need to be considered to explain why many children do not finish the course, and the sample does not significantly increase their PC. Recent learning analytics approaches have revealed valuable instruments for predicting dropout (Arizmendi et al., 2022) and outcomes (Li & Baker, 2018) based on digital behavioral traces. Therefore, future research could take a closer look at the children's learning behavior and how this interacts with measurable learning effects and the risk of dropout in the course.

From a practical perspective, our results suggest that both proof learning and asynchronous online courses should be accompanied by rich scaffolding when implemented with primary school children. Self-regulation training could be helpful when integrated into such courses, as it has been found to likely promote gifted students' academic achievement (Risemberg & Zimmerman, 1992). Furthermore, the negative feelings that children experience in challenging learning situations should be addressed, for example, with child-appropriate explanations of the Big-Fish-Little-Pond Effect (H. W. Marsh, 1987) and the Dunning-Kruger Effect (Kruger & Dunning, 1999). Furthermore, we would like to emphasize the importance of learning plans and reminders in ensuring retention in asynchronous online courses, particularly for children (Amaefule et al., 2025). For proof learning, we recommend providing the children with motivating experiences through reasoning exercises and carefully introducing formal elements, transitioning from iconic to symbolic depictions (see: Lambert, 2011, for the EIS-model of depicting mathematical content).

6.5.3 Strengths and Limitations

One significant strength of this study is its use of a randomized controlled trial design with two measurement points and a waiting control group. This robust methodology has been well-established for evaluating similar enrichment courses (Rebholz, 2017; Stark, 2025; Trautwein et al., 2023) and is recognized as a solid approach from a methodological standpoint (National Research Council, 2004). Additionally, the initial sample included 269 children. Such large-scale studies have several advantages compared to smaller studies, including the reliability and generalizability of the findings (Ertl, Hartmann, & Heine, 2020). Another strength is the broad range of outcomes measured in the

study, including proof competency, proof interest, intrinsic value, and attainment value, offering a comprehensive view of the intervention's impact on students. Finally, the study introduced an innovative intervention to the field of talent development. The asynchronous online proof course, Logical Detectives, employed a novel approach to teaching complex topics, including Boolean Logic and Set Theory. Developed using a multi-step method as recommended by Nelson et al. (2012), it addressed the needs of the target group, incorporating insights from existing research in proof education. This course has the potential to enrich the educational landscape for gifted students, particularly after further revisions.

When interpreting our research results, the study's limitations should also be considered. A primary limitation emerges from the ceiling effects observed in the motivational scales of the pre-test, which diminish the study's informative value. This issue likely arose from the fact that the children within the HCAP chose their courses based on personal interest (Trautwein et al., 2023), suggesting that our sample already possessed high motivation and self-concept in this area. To mitigate this in future studies, using questionnaires with more levels and reversed items is recommended. Additionally, a person-centered approach, as suggested by (Staus, O'connell, & Storksdieck, 2021), could yield valuable insights into the motivational effects of the course. Moreover, many children did not start the course after the pre-test, and only a small percentage completed it. These factors reduced the explanatory power associated with the large sample size. One possible reason for dropout may relate to the high domain-specific self-concept of the participants; children might have avoided the course when tasks became more challenging to protect their self-concept (see: Urdan & Midgley, 2001, for more on academic self-handicapping). Ultimately, Logical Detectives does not yet fully meet the needs of primary school children. The gap between the children's computer skills and the course requirements may have affected its feasibility. Additionally, despite gifted children generally having greater potential for self-regulated learning (Risemberg & Zimmerman, 1992), they still require scaffolding (e.g., plans, reminders) and incentives to engage effectively in self-regulated learning (Amaefule et al., 2025). The course may not have provided sufficient support in this regard.

6.6 Conclusion

The present study measured the effects of an asynchronous online proof course. It did not demonstrate a significant intervention effect on the proof competency of talented primary school children. Therefore, it remains unclear whether asynchronous learning can help children improve their proof skills. Additionally, we observed adverse effects on proof self-concept and no significant effects on other motivation variables, such as attainment value and utility value. Thus, future research should involve a revised course design that further enhances proof learning and self-regulation. Moreover,

analyzing log data will offer better insights into which aspects of asynchronous proof learning are effective and which may be overwhelming for talented primary school students.

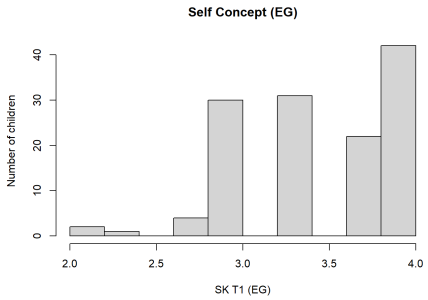
6.7 Appendix C

6.7.1 Histograms

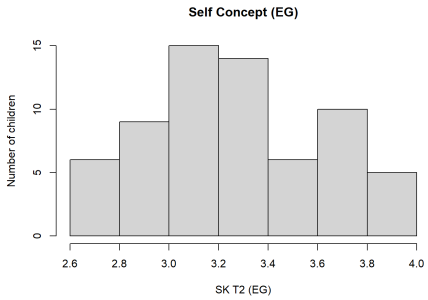
Figure 31

Distribution of self-reports on the domain-specific self-concept

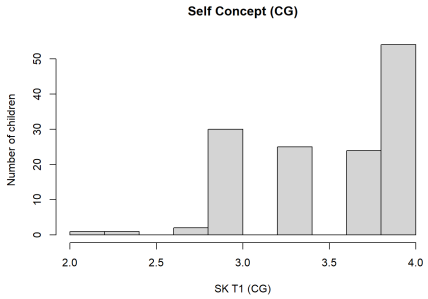
(a) *Pretest Experimental Group*



(b) *Posttest Experimental Group*



(c) *Pretest Control Group*



(d) *Posttest Control Group*

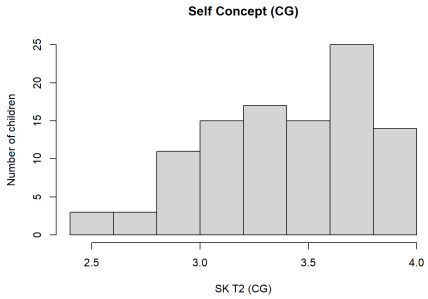
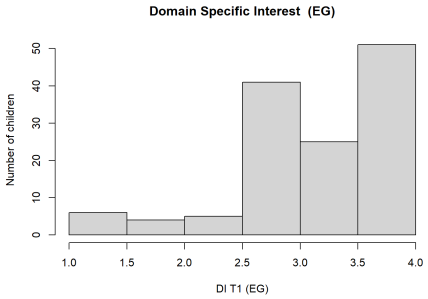


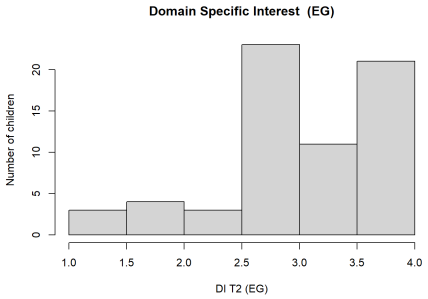
Figure 32

Distribution of self-reports on the domain-specific interest

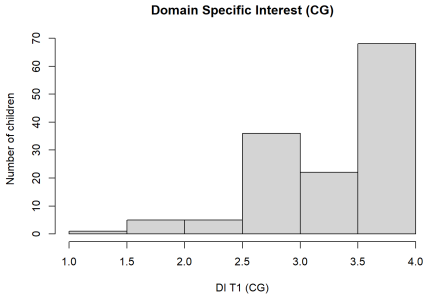
(a) *Pretest Experimental Group*



(b) *Posttest Experimental Group*



(c) *Pretest Control Group*



(d) *Posttest Control Group*

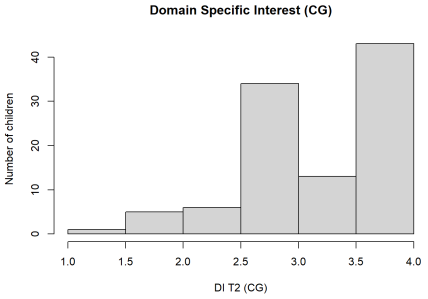
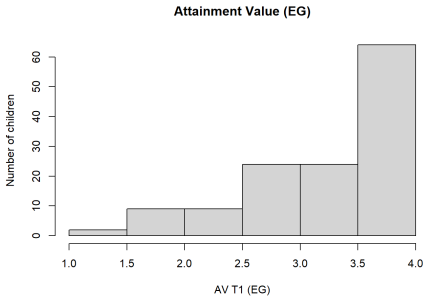


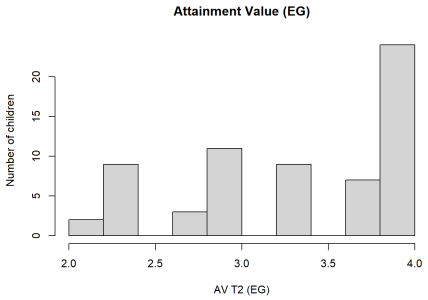
Figure 33

Distribution of self-reports on the domain-specific attainment value

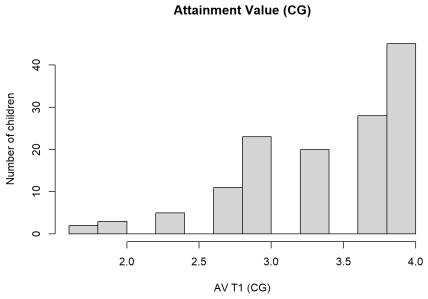
(a) *Pretest Experimental Group*



(b) *Posttest Experimental Group*



(c) *Pretest Control Group*



(d) *Posttest Control Group*

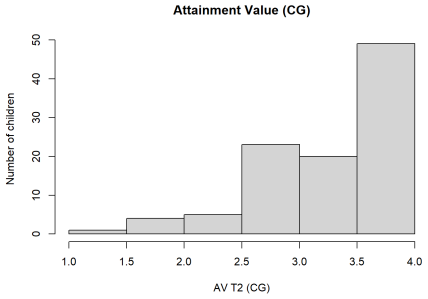
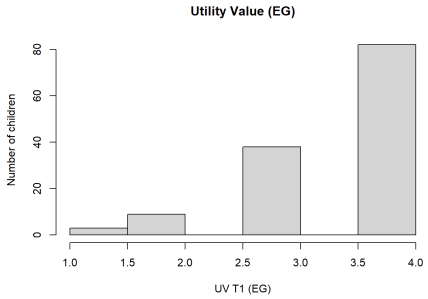


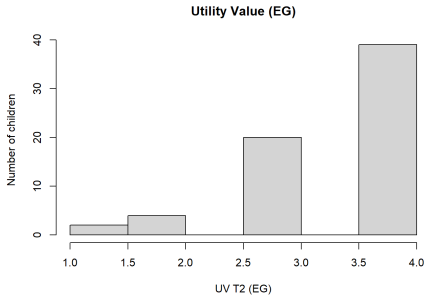
Figure 34

Distribution of self-reports on the domain-specific utility value

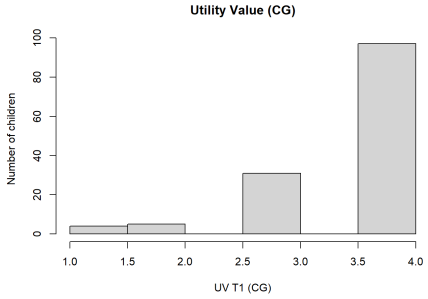
(a) *Pretest Experimental Group*



(b) *Posttest Experimental Group*



(c) *Pretest Control Group*



(d) *Posttest Control Group*

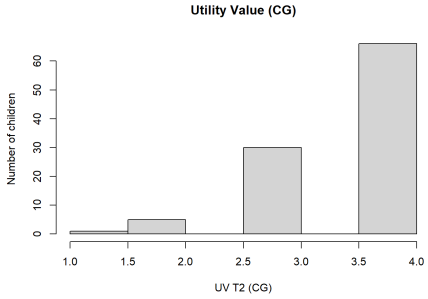
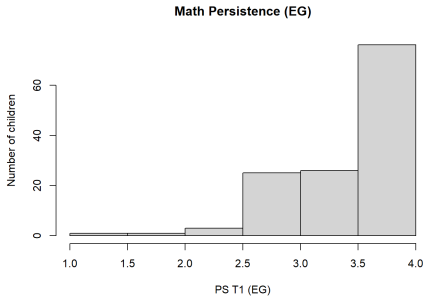


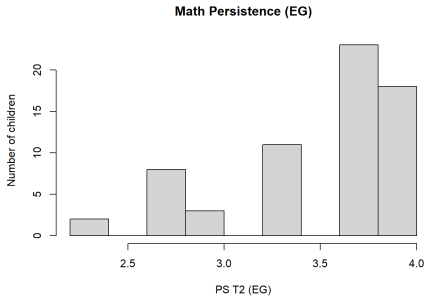
Figure 35

Distribution of self-reports on the domain-specific persistence

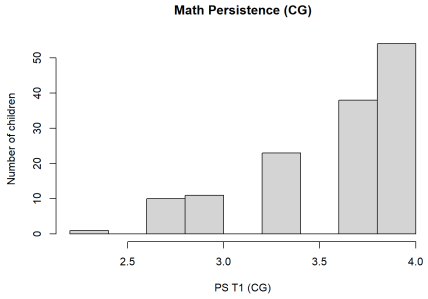
(a) *Pretest Experimental Group*



(b) *Posttest Experimental Group*



(c) *Pretest Control Group*



(d) *Posttest Control Group*

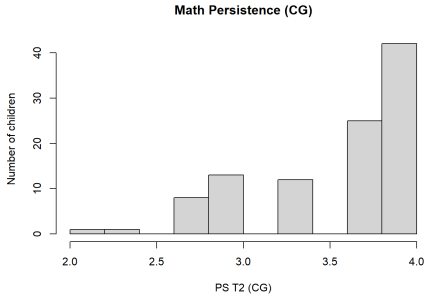
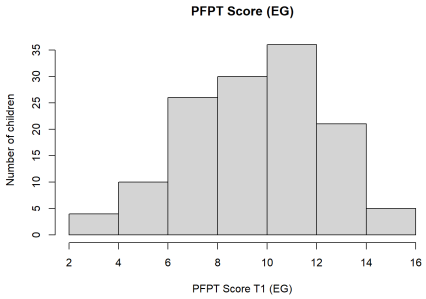


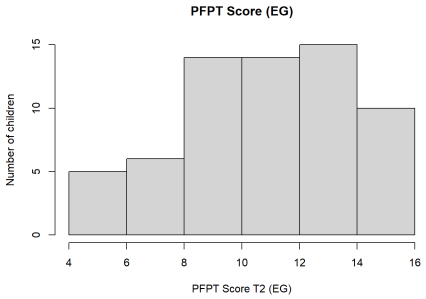
Figure 36

Distribution of proof competency scores

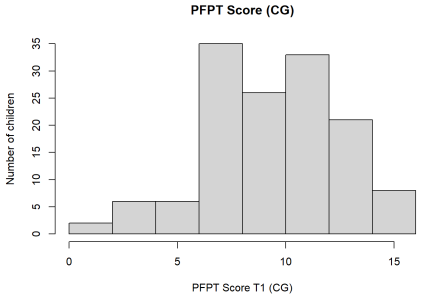
(a) *Pretest Experimental Group*



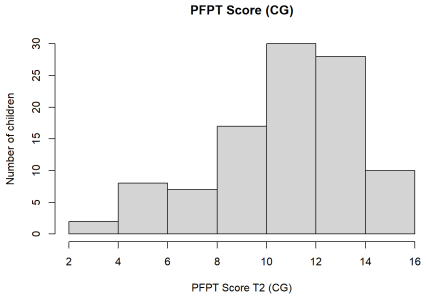
(b) *Posttest Experimental Group*



(c) *Pretest Control Group*



(d) *Posttest Control Group*



6.7.2 Effects on Finishing Children

Table 48

Effects on proof competency (finishers only)

Effect	β	SE	p
Intervention	0.101	0.191	0.599
Gender	-0.030	0.145	0.837
Age	0.083	0.080	0.298
Pretest	0.644	0.069	<0.001
R^2		0.452	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 49

Average Causal Effects (finishers only) on domain-specific motivational outcomes.

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention	-0.610	0.221	0.006	-0.345	0.189	0.067	-0.557	0.180	0.002	-0.539	0.195	0.006	0.189	0.177	0.285
Gender	-0.282	0.168	0.092	-0.024	0.144	0.868	0.169	0.138	0.222	0.344	0.149	0.021	-0.063	0.132	0.632
Age	-0.030	0.090	0.734	0.075	0.081	0.355	0.046	0.074	0.538	0.154	0.080	0.054	0.032	0.072	0.659
Pre-test	0.442	0.082	<0.001	0.646	0.080	<0.001	0.676	0.069	<0.001	0.642	0.074	<0.001	0.772	0.069	<0.001
R^2		.225			.389			.465			.424			.512	

Note. Pre-test and dependent variables were standardized before the analyses.

6.7.3 Interaction Analyses - Interaction of Treatment and Gender

Table 50

Gender Interaction Model for proof competency (intention-to-treat)

Effect	β	SE	p
Intervention:Gender	0.139	0.252	0.582
Intervention	-0.127	0.198	0.521
Gender	0.009	0.157	0.957
Age	0.110	0.067	0.101
Pretest	0.600	0.061	<0.001
R^2		0.404	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 51

Gender Interaction Model for proof competency (per-protocol)

Effect	β	SE	p
Intervention:Gender	0.125	0.264	0.636
Intervention	-0.069	0.207	0.740
Gender	0.007	0.158	0.966
Age	0.087	0.069	0.205
Pretest	0.612	0.063	<0.001
R^2		0.409	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 52

Gender Interaction Model for proof competency (finishers only)

Effect	β	SE	p
Intervention:Gender	-0.278	0.428	0.516
Intervention	0.301	0.363	0.407
Gender	0.007	0.156	0.962
Age	0.087	0.080	0.276
Pretest	0.642	0.069	<0.001
R^2		0.455	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 53

Gender Interaction Model (intention-to-treat) on domain-specific motivational outcomes.

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention:Gender	-0.080	0.288	0.781	-0.110	0.261	0.673	-0.212	0.259	0.415	-0.188	0.267	0.482	0.052	0.237	0.828
Intervention	-0.243	0.226	0.282	-0.077	0.206	0.709	0.099	0.204	0.626	-0.033	0.210	0.876	-0.041	0.186	0.824
Gender	-0.242	0.179	0.177	-0.022	0.163	0.893	0.153	0.163	0.350	0.304	0.167	0.069	-0.089	0.148	0.550
Age	-0.084	0.074	0.259	0.009	0.070	0.896	-0.038	0.068	0.577	0.029	0.069	0.676	-0.023	0.062	0.712
Pre-test	0.403	0.068	<0.001	0.663	0.072	<0.001	0.622	0.066	<0.001	0.609	0.069	<0.001	0.712	0.060	<0.001
R^2		.216			.415			.388			.366			.495	

Note. Pre-test and dependent variables were standardized before the analyses.

Table 54

Gender Interaction Model (per-protocol) on domain-specific motivational outcomes.

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention:Gender	0.004	0.299	0.990	0.023	0.265	0.930	-0.316	0.267	0.236	-0.184	0.283	0.517	0.021	0.252	0.933
Intervention	-0.348	0.234	0.138	-0.184	0.209	0.378	0.061	0.210	0.769	-0.100	0.222	0.653	-0.047	0.198	0.813
Gender	-0.244	0.179	0.173	-0.020	0.159	0.900	0.149	0.161	0.353	0.306	0.170	0.072	-0.088	0.151	0.563
Age	-0.101	0.076	0.183	-0.001	0.070	0.991	-0.026	0.069	0.705	0.031	0.072	0.668	-0.012	0.065	0.855
Pre-test	0.402	0.068	<0.001	0.641	0.071	<0.001	0.639	0.066	<0.001	0.593	0.071	<0.001	0.711	0.062	<0.001
R^2		.226			.418			.404			.350			.484	

Note. Pre-test and dependent variables were standardized before the analyses.

Table 55

Gender Interaction Model (finishers only) on domain-specific motivational outcomes.

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention:Gender	-0.288	0.495	0.561	-0.079	0.426	0.852	0.173	0.404	0.669	0.244	0.437	0.577	0.211	0.388	0.586
Intervention	-0.403	0.419	0.336	-0.287	0.362	0.427	-0.681	0.342	0.046	-0.715	0.371	0.054	0.037	0.330	0.910
Gender	-0.246	0.179	0.170	-0.014	0.154	0.928	0.147	0.148	0.320	0.311	0.160	0.051	-0.092	0.142	0.518
Age	-0.027	0.090	0.763	0.076	0.081	0.347	0.043	0.074	0.563	0.151	0.080	0.060	0.030	0.072	0.682
Pre-test	0.447	0.083	<0.001	0.648	0.081	<0.001	0.673	0.069	<0.001	0.640	0.074	<0.001	0.773	0.069	<0.001
R^2		.225			.390			.466			.426			.513	

Note. Pre-test and dependent variables were standardized before the analyses.

6.7.4 Interaction Analyses - Interaction of Treatment and Age

Table 56

Age Interaction Model for proof competency (intention-to-treat)

Effect	β	SE	p
Intervention:Age	-0.013	0.131	0.922
Intervention	-0.042	0.122	0.733
Gender	0.063	0.123	0.610
Age	0.117	0.086	0.174
Pretest	0.598	0.061	<0.001
R^2		0.403	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 57

Age Interaction Model for proof competency (per-protocol)

Effect	β	SE	p
Intervention:Age	-0.061	0.136	0.655
Intervention	0.008	0.128	0.952
Gender	0.055	0.126	0.664
Age	0.113	0.086	0.189
Pretest	0.611	0.063	<0.001
R^2		0.408	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 58

Age Interaction Model for proof competency (finishers only)

Effect	β	SE	p
Intervention:Age	-0.136	0.223	0.541
Intervention	0.100	0.191	0.600
Gender	-0.024	0.146	0.869
Age	0.103	0.086	0.232
Pretest	0.646	0.069	<0.001
R^2		0.455	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 59

Age Interaction Model (intention-to-treat) on domain-specific motivational outcomes.

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention:Age	-0.259	0.149	0.083	-0.261	0.135	0.054	-0.118	0.136	0.386	-0.208	0.140	0.138	-0.079	0.124	0.523
Intervention	-0.297	0.139	0.032	-0.150	0.125	0.232	-0.034	0.126	0.787	-0.154	0.129	0.232	-0.012	0.115	0.919
Gender	0.023	0.096	0.808	0.118	0.090	0.188	0.009	0.088	0.917	0.116	0.091	0.201	0.010	0.080	0.902
Age	-0.268	0.139	0.054	-0.059	0.127	0.643	0.073	0.127	0.565	0.236	0.130	0.068	-0.066	0.115	0.566
Pre-test	0.410	0.067	<0.001	0.668	0.071	<0.001	0.620	0.066	<0.001	0.629	0.070	<0.001	0.709	0.060	<0.001
R^2		.237			.437			.387			.385			.496	

Note. Pre-test and dependent variables were standardized before the analyses.

Table 60

Age Interaction Model (per-protocol) on domain-specific motivational outcomes.

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention:Age	-0.322	0.153	0.036	-0.288	0.135	0.034	-0.107	0.139	0.439	-0.213	0.148	0.148	-0.055	0.131	0.673
Intervention	-0.348	0.143	0.015	-0.171	0.127	0.178	-0.135	0.130	0.299	-0.214	0.136	0.116	-0.035	0.123	0.775
Gender	0.024	0.095	0.799	0.114	0.087	0.190	0.011	0.087	0.903	0.113	0.092	0.220	0.010	0.082	0.907
Age	-0.227	0.141	0.108	0.002	0.126	0.985	0.042	0.130	0.747	0.251	0.135	0.063	-0.077	0.121	0.525
Pre-test	0.412	0.067	<0.001	0.650	0.070	<0.001	0.635	0.066	<0.001	0.614	0.072	<0.001	0.708	0.062	<0.001
R^2		.252			.442			.401			.370			.484	

Note. Pre-test and dependent variables were standardized before the analyses.

Table 61

Age Interaction Model (finishers only) on domain-specific motivational outcomes.

	Self concept			Interest			Attainment value			Utility value			Persistence		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Intervention:Age	-0.395	0.255	0.122	-0.263	0.218	0.228	0.179	0.209	0.392	0.242	0.236	0.305	0.104	0.207	0.616
Intervention	-0.613	0.219	0.005	-0.346	0.187	0.065	-0.556	0.179	0.002	-0.540	0.194	0.006	0.193	0.177	0.275
Gender	0.026	0.096	0.786	0.112	0.086	0.194	0.020	0.079	0.797	0.117	0.087	0.181	0.019	0.077	0.809
Age	-0.270	0.166	0.105	-0.012	0.144	0.933	0.160	0.138	0.247	0.336	0.148	0.023	-0.068	0.132	0.605
Pre-test	0.454	0.082	<0.001	0.643	0.079	<0.001	0.680	0.069	<0.001	0.620	0.077	<0.001	0.779	0.070	<0.001
R^2		.237			.395			.469			.425			.514	

Note. Pre-test and dependent variables were standardized before the analyses.

6.7.5 Interaction Analyses - Interaction of Treatment and Pretest (PC)

Table 62

Pre-test Interaction Model for proof competency (intention-to-treat)

Effect	β	SE	p
Intervention:Pretest	-0.021	0.125	0.869
Intervention	-0.042	0.122	0.732
Gender	0.061	0.123	0.622
Age	0.112	0.067	0.094
Pretest	0.604	0.073	<0.001
R^2		0.402	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 63

Pre-test Interaction Model for proof competency (per-protocol)

Effect	β	SE	p
Intervention:Pretest	0.004	0.131	0.975
Intervention	0.008	0.128	0.948
Gender	0.052	0.127	0.683
Age	0.090	0.069	0.194
Pretest	0.608	0.074	<0.001
R^2		0.409	

Note. Pre-test and dependent variable were standardized before the analyses.

Table 64

Pre-test Interaction Model for proof competency (finishers only)

Effect	β	SE	p
Intervention:Pretest	0.273	0.200	0.170
Intervention	0.100	0.190	0.597
Gender	-0.024	0.144	0.870
Age	0.078	0.080	0.329
Pretest	0.610	0.073	< 0.001
R^2		0.452	

Note. Pre-test and dependent variable were standardized before the analyses.

6.7.6 Prediction Model for Dropout

Table 65

Logistic Regression Model for Dropout Risk - Group Variable Included

	β	SE	p
Age	-0.034	0.214	0.874
Gender	0.319	0.403	0.428
Self-Concept	0.373	0.231	0.107
Utility Value	0.212	0.241	0.379
Attainment Value	0.456	0.261	0.080
Domain Specific Interest	-0.121	0.247	0.624
Perstistence	-0.145	0.222	0.513
Proof Competency	0.211	0.206	0.306
Intervention (Group)	-0.013	0.381	0.973

Table 66

Logistic Regression Model for Dropout Risk - Group Variable Not Included

	β	SE	p
Age	-0.035	0.213	0.871
Gender	0.320	0.402	0.425
Self-Concept	0.373	0.231	0.107
Utility Value	0.213	0.240	0.375
Attainment Value	0.455	0.260	0.080
Domain Specific Interest	-0.121	0.246	0.624
Perstistence	-0.144	0.221	0.513
Proof Competency	0.211	0.207	0.305

6.7.7 Censored Variable Analyses

Table 67

Regressions According to Censored Variable Analyses

	Attainment Value (T2)			Utility Value (T2)			Domain Specific Interest (T2)			Self Concept (T2)			Persistence (T2)		
	β	SE	p	β	SE	p	β	SE	p	β	SE	p	β	SE	p
Attainment Value (T1)	**0.569**	**0.100**	**<0.001**	0.370	0.224	0.098	0.274	0.113	0.016	0.016	0.061	0.790	0.003	0.080	0.973
Utility Value (T1)	**0.248**	**0.070**	**<0.001**	**1.057**	**0.142**	**<0.001**	0.172	0.075	0.021	**0.131**	**0.044**	**0.003**	0.008	0.064	0.903
Domain Specific Interest (T1)	-0.038	0.094	0.689	-0.162	0.258	0.530	**0.495**	**0.098**	**<0.001**	0.038	0.055	0.489	0.106	0.096	0.269
Self Concept (T1)	-0.021	0.108	0.846	-0.339	0.250	0.175	-0.149	0.121	0.217	**0.159**	**0.064**	**0.013**	**0.328**	**0.090**	**<0.001**
Persistence (T1)	**0.433**	**0.111**	**<0.001**	**0.667**	**0.261**	**0.011**	**0.535**	**0.147**	**<0.001**	**0.336**	**0.063**	**<0.001**	**0.748**	**0.101**	**<0.001**
Intervention	0.076	0.097	0.435	-0.256	0.217	0.237	-0.140	0.104	0.179	**-0.104**	**0.053**	**0.049**	-0.033	0.079	0.676
Gender	0.087	0.094	0.355	0.322	0.224	0.151	-0.110	0.104	0.292	**-0.120**	**0.054**	**0.027**	-0.124	0.084	0.140

Note. T1 = Pretest, T2 = Posttest

7 STUDY 4: INVESTIGATING MOTIVATIONAL FLUCTUATION AND DROPOUT AMONG TALENTED PRIMARY SCHOOL CHILDREN IN A SELF-PACED ONLINE MATHEMATICS COURSE

The content of this chapter is currently under review for a special issue of ZDM – Mathematics Education. The proportional contributions of the (co-)authors to the manuscript are presented in the subsequent table. This article will not exactly replicate the final version published in the journal. It is not the copy of record.

Author	Author position	Scientific ideas (%)	Data generation (%)	Analsysis & interpretation (%)	Study writing (%)
Armin Fabian	first	10	0	60	85
Xenia Stein	second	55	95	10	5
Florian Berens	third	0	0	5	1
Katerina Tsarava	fourth	5	5	0	1
Wolfgang Wagner	fifth	0	0	15	0
Luise von Keyserlingk	sixth	5	0	0	1
Jessika Golle	seventh	5	0	0	1
Jeffrey A. Greene	eighth	5	0	5	1
Matthew L. Bernacki	ninth	5	0	5	1
Walther Paravicini	tenth	0	0	0	1
Ulrich Trautwein	eleventh	10	0	0	3

Abstract

Many students disengage from mathematics throughout their school careers, a pattern often attributed to declines in motivation. While long-term motivational trajectories are well documented, far less is known about short-term motivational fluctuations in self-paced online courses; a setting that becomes increasingly popular in mathematics education and which places high demands on learners' autonomy, possibly increasing the risk of disengagement when motivation decreases. Drawing on situated expectancy-value theory, we investigated the dynamic interplay between dispositional (trait-like) motivation, situational (state-like) motivation, and parental help in shaping talented children's motivational fluctuation in an extracurricular enrichment online course designed to foster children's proof competency. Longitudinal data from 159 talented primary school students indicated considerable within-person variability in motivation. Dispositional motivation at the beginning of the course predicted situational motivation only modestly. Notably, across different course sections, motivational declines seemed more pronounced when students encountered cognitively demanding content (e.g., construction of formal-deductive proofs). For utility value, parental help emerged as a critical moderator, significantly buffering these declines. Critically, our data indicated high drop-out rates among children. Findings from a survival analysis suggested that greater situational motivation (particularly self-concept and intrinsic value), but not greater dispositional motivation, reduced the risk of dropout, highlighting the importance of immediate motivational states in sustaining engagement.

Keywords: situated expectancy-value theory, self-paced online course, motivational fluctuation, dropout, parental help, talented primary school children

7.1 Introduction

Motivation has been widely recognized as a key predictor of students' engagement in mathematics (Flunger, Hollmann, Hornstra, & Murayama, 2022; Musu–Gillette, Wigfield, Harring, & Eccles, 2015). At the same time, many students experience a gradual decline in motivation, possibly contributing to disengagement with mathematics over the course of their school careers (Gaspard et al., 2015; Musu–Gillette et al., 2015). To investigate the reasons for disengagement, prior researchers have predominantly focused on investigating *long-term* motivational trajectories within traditional, on-site learning contexts, such as classrooms (Flunger et al., 2022; Parrisius, Gaspard, Zitzmann, Trautwein, & Nagengast, 2022). In contrast, considerably less is known about *short-term* motivational fluctuation, especially in the context of self-paced online courses; a learning environment where momentary motivational fluctuations may play a particular role in sustaining engagement due to the high degree of autonomy (K. Chen & Jang, 2010). This lack of insights into motivational dynamics in self-paced online settings is concerning given their possibly increasing role in mathematics education in providing extracurricular content that goes beyond the regular school curriculum.

Advancements in the field of motivation research have contributed to the development of a theoretical framework for better understanding the dynamics of motivation and dropout behavior in challenging learning environments. Specifically, according to the situated expectancy-value theory (SEVT, Eccles & Wigfield, 2020, 2024), motivation is inherently context dependent, shifting quickly when students encounter changing learning conditions. Indeed, empirical evidence suggests that situational (state-like) motivation may be more predictive of academic outcomes than dispositional (trait-like) motivation (see Wolff, Hilpert, Vongkulluksn, Bernacki, & Greene, 2024, for an example in the context of self-efficacy). However, rigorous research combining both situational and dispositional motivation in online settings is scarce.

Moreover, despite a large body of research focusing on contextual determinants of students' situational motivation - such as teachers' autonomy (Flunger et al., 2022; Witte, Spinath, & Ziegler, 2024) or peer group comparisons (Hasenbein, Trautwein, Hahn, Soller, & Göllner, 2024; Schunk & Mullen, 2012) - the role of parental help in online learning contexts remains unexplored. Understanding this is crucial, particularly for young learners, for whom parental scaffolding may play an important role in shaping situational motivation. Possibly, effective parental help may reduce motivational declines and dropout risk (Williams–Johnson & Gonzalez–Dehass, 2022).

To investigate motivational dynamics in self-paced online learning, we analyzed data from an extracurricular enrichment course on mathematical proving. The course was designed for primary

school children with indications of high cognitive potential. Given that the activity of proving is cognitively demanding (e.g., Heinze, Cheng, Ufer, Lin, & Reiss, 2008) and aligns well with characteristics of giftedness (Stein, Tsarava, Fabian, et al., 2025), we considered it a particularly suitable content area for our study, offering both a challenge for talented learners and a likely context for observing motivational fluctuation.

7.2 Theoretical Background

7.2.1 The Central Role of Motivation in Mathematics Engagement and Learning

Motivation is a key factor in driving engagement and success in mathematics (Schukajlow, Rakoczy, & Pekrun, 2023, see also Eccles & Wigfield, 2020 for evidence across disciplines). According to expectancy-value theory (Eccles & Wigfield, 2002), motivation is mainly determined by students' expectancies to perform well on a task, and the subjective value they ascribe to it. Task value is typically differentiated into several components: intrinsic value ("I enjoy doing this"), utility value ("This is useful for my life"), attainment value ("It is important for me to do well on this"), and cost ("Doing this takes too much effort").

Recently, the word "situated"has been explicitly added to expectancy-value theory (Eccles & Wigfield, 2020) to emphasize the context-dependent and dynamic nature of motivation. Accordingly, motivation is not considered a static, trait-like construct that is temporally stable (Eccles & Wigfield, 2024; Schukajlow et al., 2023), but rather a dynamic, context-sensitive one that is situational and characterized by its "embeddedness in tasks, social environment, and sociocultural contexts"(Schukajlow et al., 2023, p.252). Prior empirical research in mathematics education has confirmed the context-dependent nature of motivation by commonly applying between-subject designs, documenting variation across mathematical tasks (Krawitz & Schukajlow, 2018) and content areas (Street, Malmberg, & Stylianides, 2022). While such designs capture differences between learners, within-person approaches - as used in the present study - can complement this work by tracing how motivation fluctuates over time within the same individual (Murayama et al., 2017).

7.2.2 The Dynamic Nature of Motivation: Empirical Studies Investigating Fluctuation Within Mathematics Learners

The shift to situated expectancy-value framework (Eccles & Wigfield, 2020) necessitates a methodological shift toward within-person designs when studying motivational dynamics; a methodological approach strongly advocated by mathematics education scholars (e.g., Schukajlow et al., 2023) and education scientists alike (Flunger et al., 2022; Murayama et al., 2017; Tsai, Kunter, Lüdtke,

Trautwein, & Ryan, 2008). Although mathematics education has often relied on cross-sectional designs (Schukajlow et al., 2023), longitudinal studies in mathematics education have emerged, demonstrating overall substantial motivational fluctuations found within learners. For example, Tsai et al. (2008) conducted a three-week repeated-measures study and showed that in mathematics classrooms, around 36% of variance in students' situational interest was located within students (see Flunger et al., 2022; Witte et al., 2024, for more examples.).

Together, such longitudinal studies underscore a vibrant and rapidly expanding literature of within-person designs in motivational research. What remains scarce, however, is work investigating situational (state-like) fluctuation and its relationship to dispositional (trait-like) motivation, particularly in self-paced online learning where learners' initial motivation may decline quickly without teacher regulation.

7.2.3 The Relationship between Dispositional and Situational Motivation

Trait-like dispositional motivation reflect more enduring patterns of motivation that students bring with them into new situations (see Flunger et al., 2022, for a discussion). In contrast, state-like situational motivation capture how students feel "in the moment" when working on a particular topic or activity (Eccles & Wigfield, 2020). In theory, motivational dispositions can predispose learners to experience higher levels of momentary motivation in a given context (Keller, Yanagida, Lüdtke, and Goetz 2025; Parrisius et al. 2022, see also Moeller, Viljaranta, Tolvanen, Kracke, and Dietrich 2022 for a comprehensive framework on the reciprocal relationship of motivational traits and states).

Initial empirical evidence for this relationship exists, at least for some motivational constructs like interest or emotions. For example, Tsai et al. (2008) showed that adolescent students who possessed higher dispositional interest in mathematics also exhibited higher situational interest in specific subsequent lessons. For self-concept, Witte et al. (2024) found that dispositional and situational self-concept were positively related (mean correlation of aggregated states and disposition: $r = .68$). Relatedly, Keller et al. (2025) demonstrated that aggregated situational emotions, assessed across three years, and self-reported dispositional emotions among school students were highly correlated ($r \approx .50$).

7.2.4 The Role of Motivation in Self-Paced Online Learning Environments

Short-term motivational fluctuations may be especially important in online courses, where learners must persist with minimal external regulation (e.g., teacher support). For example, a learner with strong dispositional self-concept in mathematics might still experience sharp declines in situational

self-concept during a challenging online module if they perceive minimal support or believe the content is irrelevant. Conversely, a learner with only moderate dispositional motivation may become momentarily highly motivated if the specific task is of personal value or interest to them. Very recent empirical evidence indicated deeper engagement with online materials in a technology-supported learning environment when situational interest was high (Guo & An, 2025). On the other hand, prior research has shown that situational motivation tends to decline over the course of online learning (Kyewski & Krämer, 2018), which may increase the risk of dropout as online courses evolve.

Dropouts in Self-Paced Asynchronous Online Learning Environments High dropout rates have been consistently documented for online learning environments across disciplines and diverse contexts (Muljana & Luo, 2019; Tinto, 1975), typically exceeding those in face-to-face courses by a large margin (Xavier & Meneses, 2020). In this study, we consider dropout to be a behavioral indication of disengagement. High dropout rates pose a huge threat to the practical feasibility and effectiveness of online courses (W. Wang, Guo, He, & Wu, 2019). Existing research has identified several reasons for the increased risk of dropout in online learning environments, and these reasons are often broadly distinguished between factors associated with the external context and factors that may be located in the learners themselves (Schunk & Mullen, 2012). For example, Y. Lee, Choi, and Kim (2013) distinguished between internal (i.e., motivation, knowledge and self-regulation) and external factors (e.g., parental help) and investigated their respective power for predicting the completion rate in an online course of university students in Korea. Based on a single pre- and post-survey, their findings indicated that students who dropped out showed lower metacognitive self-regulation than completers, whereas no meaningful differences emerged regarding perceived support from family.

Support from family may be particularly crucial for children engaging in self-paced online learning due to the lack of external support from a teacher. Possibly, in the absence of a teacher, children are likely to ask for help from their parents. However, despite a large body of empirical research indicating the crucial role of parents in the long-term development of students' achievement and motivation (X. Wang & Wei, 2024) and its explicit theoretical grounding in Eccles and Wigfield's cultural milieu (2020), there is limited knowledge about the role that parental help in children's situational motivation. Potentially, parental help may help sustain motivation during challenging online tasks and prevent dropout.

7.2.5 Fostering Proof Competencies Among Talented Children in A Self-Paced Online Course

To investigate motivational dynamics in self-paced online courses, this study draws on data from an extracurricular online course. Such courses remain rare in primary education but may gain importance as enrichment opportunities for talented children, offering flexibility beyond the time and space constraints of traditional on-site programs (Stein, Tsarava, Fabian, et al., 2025).

Mathematically talented children—characterized by high performance potential and motivation (Subotnik et al., 2011)—require cognitively demanding content to stay engaged and develop their abilities fully (Rotigel & Fello, 2016). Insufficiently challenging learning environments risk boredom and underachievement, whereas complex, non-routine problems foster interest and higher-order thinking (Deal & Wismer, 2010). The topic of proving provides exactly such challenges and thus offers promising opportunities for mathematical enrichment beyond the regular curriculum.

Building on this rationale, we developed an asynchronous online course to foster proof competencies among talented children. We define proof competency as the ability to comprehend and construct formal-deductive proofs, following Stylianides' (2007) definition of proof as a "mathematical argument, a connected sequence of assertions for or against a mathematical claim, with the following"(p.291). Although proofs are central to mathematics, students often struggle with them (e.g., Sommerhoff & Ufer, 2019) experiencing high cognitive demands and anxiety (e.g., Heinze, Cheng, & Yang, 2004; Häsä, Westlin, & Rämö, 2023). A. Stylianides (2007) therefore advocated introducing proofs in primary school already to provide meaningful engagement and reduce later anxiety. Talented learners may particularly benefit from such an early exposure, as shown by Stein, Tsarava, Fabian, et al. (2025), since the proving process aligns well with their cognitive strengths.

The course Logical Detectives is an asynchronous extracurricular online course (see Stein, Tsarava, Fabian, et al., 2025, for details on its development). attended by talented children. Whereas the first two chapters of the course introduced the notion of proof and need of proof, the last three chapters required them to construct formal proofs in the context of Boolean Logic, Set Theory, and Elementary Number Theory [1]. A detailed overview of the content and learning objectives of each chapter, together with exemplary screenshots, is provided in OSF (`https://osf.io/qf6by/files/osfstorage`).

7.2.6 The present study

As outlined above, situational motivation appears to be a key factor in self-paced online learning. In this study, we focused on three situated expectancy-value components - self-concept, intrinsic

[1] In the framework of this dissertation, the course chapters are numbered as Chapters 0 to 4. In this section, however, the same chapters will be divergently numbered as Chapters 1 to 5 due to different standards applied during the evaluation process.

value, and utility value - which were assessed repeatedly during the course (see Section 7.3.2 for details). Despite their relevance, little is known about how these motivational facets relate to dispositional factors or contribute to dropout. To address these gaps, we draw on data from a self-paced online mathematics course designed for primary school children as part of an enrichment program for talented students. Although this selective sample is likely to display above-average motivation with little variance (see Gottfried & Gottfried, 1996, for an empirical example in the context of mathematical interest), it offers a unique opportunity to examine how even talented learners respond motivationally to cognitively demanding online mathematics content - and whether such challenges may still lead to dropout during the course. Based on this rationale, we formulated the following research questions.

1. How much fluctuation of motivation is there within talented children across the online course?
2. To what extent is dispositional (trait-like) motivation at the beginning of the course related to situational (state-like) motivation during the course?
3. How does situational motivation change from chapter to chapter during the course?
4. Does parental help moderate the chapter to chapter changes in motivation?

   Given the considerable high drop-out rates observed in our study (> 85 %), we further examined possible reasons for disengagement, and formulated the following research question:

5. What factors may reduce the risk of dropout: dispositional motivation, situational motivation, or parental help?

Based on situated expectancy-value theory, we expected that most motivational variance would be located within children across chapters, while between-person variance is small due to the selective nature of our talented sample (RQ1). For RQ2, we anticipated that higher dispositional motivation would be associated with higher situational motivation. By contrast, RQ3-RQ5 were treated as exploratory, given the limited empirical evidence in the context of proof learning and online enrichment courses.

7.3 Method

7.3.1 Study Context and Design

In total, the study draws on data from 159 primary school children ($M_{\text{age}} = 9.58$; $SD_{\text{age}} = 0.59$; 38% girls) enrolled in Logical Detectives, a course which was implemented in the context of Hector Children's Academy Program in Baden-Württemberg, South Germany, a large enrichment program

to foster talented primary school children (see Trautwein et al., 2023, for details). Our sample consisted of children who (a) had been nominated by their teachers to attend the enrichment program based on high overall interest and engagement, as well as assumed intellectual potential (i.e., no formal testing occurs; see Trautwein et al., 2023) and who (b) voluntarily enrolled in the six-week online course Logical Detectives. As such, we expected children in our study to show high levels of motivation with little overall variance after all. Strikingly, 136 children in total dropped out during the course, either within chapters or between chapters (see 7.3.2 for details).

The study followed a longitudinal research design. More specifically, children's dispositions (i.e., socio-demographics, motivation and proof competency) were collected before they entered the course with additional repeated measures of situational motivation and parental help during the course.

7.3.2 Instruments

Dispositional Motivation and Knowledge We assessed motivational disposition before the online course started. Following expectancy-value theory, we assessed domain specific dispositional self-concept as a conceptualization of self-efficacy (see H. Marsh et al., 2019, for empirical justification), intrinsic value and utility value using 4-point Likert scales. To assess domain-specific intrinsic value, we adapted three items from Stalder, (2013, e.g. "I find everything interesting that has to do with mathematical reasoning.", α= .78). To assess domain-specific self-concept, we used three items from Arens et al. (2011, e.g. "I'm good at everything that has to do with mathematical reasoning."; α= .77). To assess utility value, we used the following item by Leifheit et al. (2019): "Being good at mathematical reasoning gives me advantages in school". To assess prior proof competency, we used a previously developed test instrument (α= .84) that measures proof competency in the dimensions of Boolean Logic, Set Theory and Elementary Number Theory with 15 closed items (Stein, Tsarava, & Goecke, 2025).

Situational Motivation and Parental Help We assessed situational motivation immediately *before* and *after* each of the five chapters with items specifically addressing the content of the respective chapters. The two short questionnaires were implemented directly into the learning management system to minimize disruption to students' learning (see Bernacki, Nokes–Malach, & Aleven, 2015, for a related approach). Each item was rated on a four-point response scale.

For each sub-scale of motivation (i.e., self-concept, intrinsic value and utility value [2]) used during the course, we used single-item indicators as recommended when assessing motivation frequently

[2]Attainment value and costs were omitted due to challenges in reliably assessing these constructs with single items in younger students and to keep assessments brief in the self-paced online format.

(Eccles & Wigfield, 2024; Gogol et al., 2014). These single items were taken from previously developed and tested scales (Arens et al., 2011; Leifheit et al., 2019), and slightly adapted in a child-friendly language. The items translated from German answered by the children before each chapter were as follows:

- Self-concept: "I am confident that I will be able to solve the tasks in this chapter."
- Intrinsic value: "I am looking forward to this chapter."
- Utility value: "I want to learn useful things for school in this chapter."

The corresponding post-chapter items were:

- Self-concept: "I did well in this chapter."
- Intrinsic value: "I enjoyed this chapter."
- Utility value: "I learned useful things for school in this chapter."

Note, that pre-chapter items captured children's prospective motivational expectations for the upcoming chapter, whereas post-chapter items assessed retrospective evaluations of their experience during the chapter (see limitations). For consistency in reporting, we used the same labels for both pre- and post-chapter measures.

To assess parental help of each chapter, the post-chapter questionnaire contained the following single item: "I received no/little/some/a lot of help from an adult[3]."

Dropout To capture patterns of dropout throughout the course, we distinguished between two types of dropouts for each of the five chapters: dropout *during* a chapter and dropout *after* a chapter. A student was classified as having dropped out during a chapter if they had started the chapter (i.e., completed the pre-chapter situational questions) but did not finish it, as indicated by missing data on the post-chapter questionnaire. In contrast, a student was classified as having dropped out *after* a chapter if they completed the post-chapter questionnaire but did not provide data for the pre-chapter questionnaire of the subsequent chapter, indicating disengagement prior to the start of the next chapter. Importantly, the course was designed so that students could only progress to the next chapter (and hence stay in the course) after completing the post-chapter questionnaire. This allowed us to model whether a student dropped out during or after a chapter. Although a malfunction of the learning record management system (see 7.3.4) resulted in missing data for situational motivation on some days, it is important to emphasize that dropout during a chapter could be identified precisely.

[3]The survey item referred to "adult"help, which likely included parents in most cases but may also have involved other caregivers (e.g., relatives, tutors)

7.3.3 Procedure

Participants were recruited through the Hector Children's Academy Program's annual summer course adverts. Information about the study was provided to parents and children eight weeks before the summer holidays. Parents submitted informed consent up to one day before course start. At course launch, children completed a questionnaire including socio demographics, dispositional motivation, and prior proof competency. After the pre-test, the children were randomly assigned to either the intervention or a waitlist control group, as this study was part of a larger project including an effectiveness study of the course on children's proof competency. Children in the waitlist control group started the same asynchronous online enrichment course but several weeks later. Because both groups ultimately received theidentical intervention under structurally equivalent conditions and no baseline differences were present, their data were combined for the present analyses.

During the course, the children completed the brief surveys on situational motivation and parental help immediately before and after each of the five course chapters. Participants who completed both the pre- and post-tests were entered into a raffle for a bookstore voucher. Before data collection, the study received approval from the local ethics committee.

7.3.4 Analyses

Occurrence of Missing Data There were two sources for missing data. First, children started the course but dropped out while participating (see 7.3.2) resulting in missing data after their disengagement. Second, due to a temporary malfunction of Moodle's Learning Record Store used for capturing situational motivation and parental help, data on these variables were not recorded for ten days (systematic but not dependent on any outcome variables).

To account for missing data, we estimated the mixed-effects models by Maximum Likelihood, which uses all available observations per learner yielding unbiased estimates under the missing at random assumption. This assumption appears plausible given the exogenous nature of the technical failure and the fact that dropout was partly explained by observed situational motivation.

Statistical Analyses To address the first set of research questions (i.e., investigation of fluctuation and predictors of situational motivation), we applied a multilevel modelling framework suitable for analysing repeated measures time points (here: pre- and post- chapter values) nested within children (Hox, Moerbeek, & Van De Schoot, 2017). Compared to classical approaches like repeated-measures ANOVA, such models are more robust to unbalanced data and allow us to distinguish within-person predictors (e.g., parental help fluctuations) from between-person predictors (e.g., dispositional motivation).

We analysed each motivational construct (i.e., utility value, self-concept and intrinsic value) separately given their conceptual differences (see 7.3.2), fitting distinct multilevel models for each construct. Further, we examined pre-chapter and post-chapter measurements in distinct models, as our primary interest was in examining children's overall trajectories across chapters rather than investigating specific motivational changes within a chapter.

To answer how much situational motivational fluctuation is found within children (RQ1), we fit a random-intercept model without predictors and calculated the Intraclass Correlation Coefficient (ICC). Accordingly, the proportion of variance reflecting within-student variance is captured by the value of $1 - ICC$. Because $1 - ICC$ also includes measurement error, it should be interpreted as an upper bound of the true within-student variance.

To investigate the relationship between dispositional and situational motivation (RQ2), we added the respective dispositional motivational scores (grand mean centred) as a level 2 predictor to the empty models. This allowed us to examine whether dispositional motivation significantly predicted the average level of situational motivation across chapters.

To investigate how situational motivation changed from chapter to chapter (RQ 3), we fit a multilevel model that included successive difference contrasts as fixed effects. These contrasts were coded to represent mean differences between each chapter and its immediate predecessor. This coding allowed us to statistically examine whether, and how, situational motivation significantly increased or decreased across consecutive chapters, while simultaneously control for the nested structure. To elaborate, the following equation represents the model applied to each motivational construct (pre- and post-chapters):

$$\text{sit.motivation}_{ij} = \gamma_{00} + \gamma_{10} D_{(2,1),ij} \; + \gamma_{20} D_{(3,2),ij} + \gamma_{30} D_{(4,3),ij} + \gamma_{40} D_{(5,4),ij} + u_{0j} + e_{ij}$$

where γ_{00} is the grand mean motivation across all chapters, $D_{(k,k-1),ij}$ the contrast-coded predictors[4] indicating the mean difference from Chapter $k-1$ to Chapter k, γ_{10} to γ_{40} the fixed effects representing the average motivational change between successive chapters, u_{0j} the random intercept for child j and e_{ij} the residual of child j at Chapter i ($i = 1, \ldots, 5$).

To investigate whether parental help affected motivational changes across successive chapters (RQ 4), we included parental help as a moderator (i.e., interaction term with each contrast code) to these models in a next step. In these models, parental help was within-person centred, meaning

[4]For chapters $k = 1, \ldots, 5$, the contrasts are defined as $D_{(k,k-1),ij} = \left\{ \begin{array}{l} \frac{k-6}{5} \text{ for } i \leq k \\ \frac{k-1}{5} \text{ for } i > k \end{array} \right\}$, where $i \in \{1, \ldots, 5\}$. This contrast coding corresponds to effect coding used to capture mean differences between successive chapters (see Breen, 2018, as an example).

each value represented the deviation from a child's average level of help received in the respective chapter. Consequently, the interaction coefficients tested whether deviations from a child's own average moderated the expected contrast-based shifts in situational motivation from one chapter to the next. Here, we concentrated on post-chapter situational motivation given that parental help was only asked post-chapter. To mitigate confounding effects from dispositional motivation, we included the respective (grand-mean centred) motivation as control variables in our models. Further, to control the false-discovery rate, we adjusted the p-values with the Benjamini-Hochberg procedure.

To examine predictors of dropout (RQ 5), we used Cox proportional hazards models (Zhang, Reinikainen, Adeleke, Pieterse, & Groothuis–Oudshoorn, 2018), a regression method suited for analysing both the occurrence and timing of events by simultaneously allowing to investigate time-invariant predictor and time-varying predictors. These models estimate predictor effects independently of the events distribution and can flexibly handle right-censored data (i.e., students who did not drop out by the end of the course). We included dispositional motivation as a time-invariant and situational motivation as a time-varying predictor, enabling direct comparison of their predictive strength. Separate models were estimated for dropout during and after chapters.

Given our focus on motivational effects beyond individual differences in proof competency, we included pretest proof competency as a control variable in all models. This is empirically supported, as motivational beliefs and achievement are closely related (H. Marsh et al., 2015), with related evidence in proving (Häsä et al., 2023). Further, sensitivity analyses without this control yielded largely the same conclusions, though a few effects lost significance, indicating that its inclusion reduces confounding and improves precision for motivational effects. As a robustness check, we also reran all models with the inclusion of gender as a covariate. Yet, no meaningful changes were observed which is why we will report the findings based on the parsimonious models without gender.

All analyses were conducted using R version 4.4.2. Anonymized data and scripts used for analysis are available in OSF (`https://osf.io/nx2p9/`)

7.4 Results

7.4.1 Preliminary Analyses and Descriptives

Before analysing the data regarding the research questions, we conducted several descriptive statistics which are presented in Table 68 (see also Supplementary Information 7.6.2 for correlation tables between situational motivation constructs). As can be seen, children in our sample reported consistently high levels of dispositional and situational motivation suggesting ceiling effects. Further, Figure 37 illustrates the (between-person) trajectories of situational motivation and parental help

Table 68

Descriptives of Dispositional Motivation, Knowledge and Situational Motivation for Each Chapter

Variables	Dispositions		Chapter 1			Chapter 2			Chapter 3			Chapter 4			Chapter 5		
	M	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Utility value	3.58	0.71															
Self-concept	3.49	0.47															
Intrinsic value	3.30	0.70															
Knowledge	9.79	3.12															
UV_{pre}			134	3.12	0.38	131	3.33	0.81	101	3.29	0.89	50	3.20	0.95	31	3.16	0.97
UV_{post}			134	3.72	0.47	104	3.78	0.42	54	3.28	0.66	33	3.15	0.79	23	3.17	0.83
SC_{pre}			134	3.82	0.38	131	3.87	0.34	101	3.91	0.29	50	3.74	0.59	31	3.58	0.56
SC_{post}			134	3.95	0.25	104	3.94	0.23	54	3.55	0.67	33	3.55	0.67	23	3.48	0.79
IV_{pre}			135	3.98	0.15	131	3.98	0.12	102	3.97	0.17	50	3.94	0.24	31	3.97	0.18
IV_{post}			134	3.99	0.17	104	3.99	0.09	54	3.96	0.19	33	3.91	0.29	23	3.87	0.34
Par. Support			129	1.62	0.73	102	1.75	0.85	52	2.02	0.80	32	1.84	0.81	23	1.74	0.81

Note. UV = Utility value, SC = self-concept, IV = Intrinsic value. The subscript *pre* indicates pre-chapter values, while the subscript *post* indicates post-chapter values

across chapters.

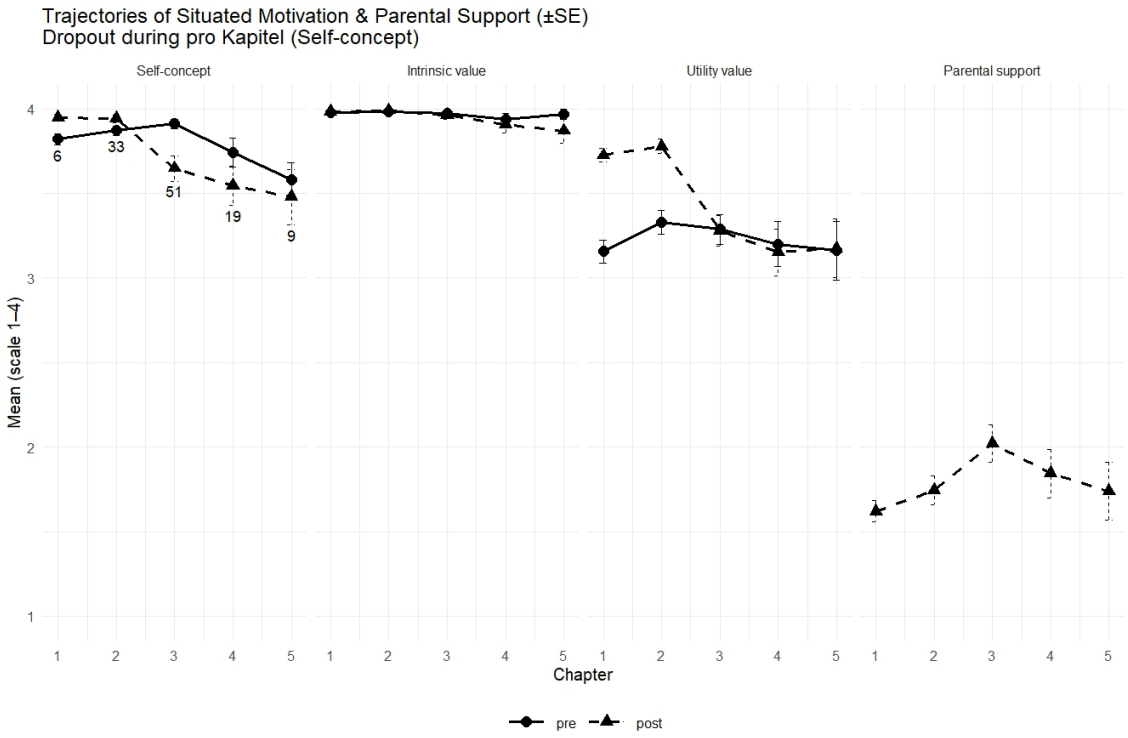
7.4.2 Research Questions 1 and 2: Variability of Situational Motivation Within Children and the Relationship to Dispositional Motivation

Pre-Chapter Motivation As indicated in Table 69, results of the empty pre-chapter models indicated substantial within-student fluctuations, though the magnitude varied by motivational construct. Specifically, self-concept (72%) and intrinsic value (85%) showed notable variability within children from one chapter to another, highlighting their sensitivity to changing contexts. In contrast, utility value was more stable, with a substantial portion of variance attributable to inter-individual differences (ICC = 59%) rather than situational fluctuations.

When including the respective dispositional motivation values as fixed effects into these pre-chapter models (controlling for prior knowledge), we found that children with higher dispositional self-concept showed significantly higher pre-chapter situational self-concept ($b = 0.146$, $p = .005$) with a small to medium effect size. Similarly, higher dispositional intrinsic value significantly predicted situational intrinsic value, $b = 0.026$, $p = .040$. Conversely, dispositional utility value was not a significant predictor of pre-chapter situational utility value ($p = .534$), suggesting dispositional utility may not influence children's immediate pre-chapter motivation.

Figure 37

The Trajectories of Situational Motivation and Parental Help Across the Five Course Chapters



Note. Dashed lines indicate post-chapter mean values and solid lines pre-chapter mean value. The numbers in the figure self-concept indicate the numbers of children dropping out during the respective chapter.

Table 69

Models Predicting Pre-Chapter Situational Motivation

Variables	Utility value		Self-concept		Intrinsic value	
	Empty model	Including disposition	Empty model	Including disposition	Empty model	Including disposition
Fixed Effects						
Intercept	3.25***	3.70***	3.83***	4.02***	3.97***	3.99
Disposition		-0.05		0.146**		0.026*
Random effects						
Intercept variance	0.39	0.37	0.05	0.04	0.003	0.003
Residual variance	0.28	0.28	0.12	0.12	0.02	0.02
ICC	0.59		0.28		0.15	
Model fit indices						
AIC	951.3	948.5	434.9	427.4	-367.0	-368.2
BIC	963.6	969.0	447.2	447.9	-354.7	-347.7
Deviance	945.3	938.5	428.9	417.4	-373.0	-378.2

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Deviance = - 2 × log-likelihood. Lower values indicate better model fit. ***$p < .001$, **$p < .01$, *$p < .05$

Table 70

Models Predicting Post-Chapter Situational Motivation

Variables	Utility value		Self-concept		Intrinsic value	
	Empty model	Including disposition	Empty model	Including Disposition	Empty model	Including Disposition
Fixed Effects						
Intercept	3.59***	3.81***	3.85***	4.00***	3.97***	3.90***
Disposition		0.10*		0.12*		0.07**
Random effects						
Intercept variance	0.05	0.04	0.03	0.03	0.01	0.01
Residual variance	0.31	0.32	0.16	0.16	0.02	0.02
ICC	0.14		0.17		0.39	
Model fit indices						
AIC	635.8	632.2	416.9	414.9	-187.0	-196.2
BIC	647.4	651.5	428.4	434.1	-175.5	-176.9
Deviance	629.8	622.2	410.9	404.9	-193.0	-206.2

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Deviance = - 2 × log-likelihood. Lower values indicate better model fit. ***$p < .001$, **$p < .01$, *$p < .05$

Post-Chapter Motivation Results of the post-chapter models similarly revealed considerable within-student fluctuations, again varying by motivational construct (see Table 70). Both self-concept (17%) and utility value (14%) showed substantial variability within children across chapters, indicating high situational sensitivity. Intrinsic value, in contrast, showed greater within-stability (61%), with a larger portion of variance arising from stable differences between children (ICC = 39%).

When including dispositional motivation as fixed effects in these models (controlling for prior knowledge), we found that all dispositional motivation values statistically significantly predicted post-chapter situational motivation with varying effect sizes from 0.07 for intrinsic value to 0.12 for self-concept.

7.4.3 Research Question 3 and 4: The Mean Trajectories of Situational Motivation Across all Chapters and the Effect of Parental Help

Detailed findings of the contrast analyses to investigate mean changes between successive chapters are found in Table 71.

Pre-Chapter Trajectories of Situational Motivation For utility value, the models indicated no statistically significant contrast effects indicating that there were no (between-subject) mean changes of pre-chapter utility values for any chapter transitions. For self-concept, no statistically significant changes emerged from Chapter 1 to Chapter 2 (b = 0.036, SE = 0.040, p = .372) or from Chapter 2 to Chapter 3 (b = 0.034, SE = 0.044, p = .446). However, statistically significant decreases were observed from Chapter 3 to Chapter 4 (b = -0.193, SE = 0.059, p = .001) and from Chapter 4 to Chapter 5 (b = -0.190, SE = 0.076, p = .013) indicating that children's situational self-concept before chapters declined notably towards the end of the course. For intrinsic value, none of the successive chapter contrasts reached statistical significance (all $p_s > 0.17$).

Post-Chapter Trajectories of Situational Motivation For utility value, the model revealed no statistically significant mean change from Chapter 1 to Chapter 2, but a statistically significant decrease occurred from Chapter 2 to Chapter 3 (b = -0.497, SE = 0.085, p <.001) indicating that children on average considered the value of Chapter 3 for their school life considerably less than Chapter 2. The contrasts from Chapter 3 to Chapter 4 and from Chapter 4 to Chapter 5 were not statistically significant, indicating that after an initial decline, between-subject utility values stabilized through the remaining chapters.

In the self-concept model, no statistically significant change occurred from Chapter 1 to Chapter

Table 71

Fixed Effects From Linear Mixed Models Testing Successive Chapter Contrasts on Situational Motivation

Contrasts	Pre chapter			Post chapter		
	Utility value	Self-concept	Intrinsic value	Utility value	Self-concept	Intrinsic value
Intercept	3.21 (0.060)	3.77 (0.027)	3.96 (0.010)	3.41 (0.041)	3.71 (0.030)	3.94 (0.015)
$D_{(2,1)}$	0.13 (0.066)	0.04 (0.040)	0.01 (0.018)	0.05 (0.065)	-0.01 (0.048)	0.00 (0.020)
$D_{(3,2)}$	-0.04 (0.072)	0.03 (0.044)	-0.02 (0.019)	-0.49*** (0.085)	-0.29*** (0.063)	-0.03 (0.027)
$D_{(4,3)}$	-0.12 (0.096)	-0.19** (0.059)	-0.04 (0.026)	-0.15 (0.111)	-0.10 (0.082)	-0.05 (0.034)
$D_{(5,4)}$	-0.14 (0.124)	-0.19* (0.076)	0.03 (0.034)	-0.01 (0.135)	-0.09 (0.100)	-0.02 (0.041)

Note. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, standard errors in parentheses.

2 (b = -0.008, SE = 0.048, p = .862). However, a statistically significant decline was present from Chapter 2 to Chapter 3 (b = -0.295, SE = 0.063, $p < .001$), indicating a reduction in children's self-concept mid-course. Subsequent contrasts remained statistically non-significant, suggesting stabilization thereafter. For intrinsic value, the model indicated no statistically significant differences across chapters, implying consistent (high) intrinsic motivation throughout the course duration.

The Effect of Parental Help on Post-Chapter Changes As indicated in Table 72, we found two significant effects after adjusting the p-values. First, the self-concept mode revealed a statistically significant positive main effect of parental help (b = 0.151, SE = 0.046, p_{adj} = .001), indicating higher-than-average parental help was linked to higher situational self-concept across chapters. Second - and critically - , parental help significantly moderated the decline in utility value from Chapter 2 to Chapter 3 (b = 0.42, SE = 0.152, p_{adj} = .005), suggesting that higher-than-average parental help may have attenuated the pronounced decline observed between these chapters.

As a safe guard against potential bias in our results due to selective dropout, we re-estimated the models among the subset of children who completed all five chapters (N = 23). These sensitivity analyses replicated the main findings of the full-sample models, supporting the robustness of our results (see Supplementary Information 7.6.3 for details).

7.4.4 Research Question 5: What Predicts Dropout?

In total, 136 out of initial 159 (i.e., 85.5%) children dropped out in the course of the study. Notably, only a small proportion of non-completers dropped out after a chapter (13.2%), whereas the majority did so during a chapter (118 of 136; 86.8%). Of those who dropped out during a chapter, most did so in Chapter 2 (28.0%) and Chapter 3 (43.2%).

Table 72

Moderating Effects of Within-Person Centred Parental Help on Motivational Changes

	Utility value	Self-concept	Intrinsic value
Main Effects			
(Intercept)	3.35 (0.195)	3.81 (0.085)	3.88 (0.045)
$D_{(2,1)}$	0.08 (0.064)	-0.01 (0.049)	0.00 (0.021)
$D_{(3,2)}$	-0.55*** (0.085)	-0.34*** (0.065)	-0.04 (0.028)
$D_{(4,3)}$	-0.13 (0.110)	-0.07 (0.085)	-0.04 (0.036)
$D_{(5,4)}$	0.01 (0.129)	-0.07 (0.100)	-0.02 (0.042)
Help	0.12 (0.060)	0.15** (0.046)	-0.02 (0.019)
Interactions			
$D_{(2,1)}$ x help	-0.01 (0.157)	-0.16 (0.119)	-0.03 (0.056)
$D_{(3,2)}$ x help	0.42* (0.152)	0.23 (0.116)	0.08 (0.052)
$D_{(4,3)}$ x help	0.19 (0.203)	0.20 (0.156)	-0.12 (0.069)
$D_{(5,4)}$ x help	-0.25 (0.255)	-0.03 (0.196)	0.10 (0.084)

Note. Intercept = grand mean across chapters (at average help). $d_{(k+1,k)}$ = mean difference between successive chapters (negative values indicate declines). Help = within-child deviations from average. Interaction terms indicate whether parental help alters the motivational change between two chapters: positive values mean that more-than-usual help buffered declines or amplified gains, while negative values mean it intensified declines or dampened gains. SES in parentheses.
$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, adjusted p values

The survival analyses revealed the following: Regarding dropout during chapters, the analysis of utility value revealed no statistically significant relationship between dropout and either dispositional utility value (b = 0.048, SE = 0.135, p = .720) or situational utility value (b = -0.183, SE = 0.108, p = .090).

The self-concept model showed a statistically significant positive association between dispositional self-concept and dropout (b = 0.457, SE = 0.224, p = .041). Specifically, an increase of 0.1 points in dispositional self-concept corresponded to approximately a 4.6% increase in dropout risk. In contrast, situational self-concept was negatively and significantly associated with dropout (b = -0.620, SE = 0.209, p = .003), indicating that each additional 0.1 points in situational self-concept reduced the dropout risk by approximately 5.6%. This suggests that children's perceptions of their abilities immediately before the chapter (situational self-concept) appeared protective against dropout, whereas higher more general perceptions of ability (dispositional self-concept) were associated with

increased dropout risk.

The analysis of intrinsic value indicated no significant association between dispositional intrinsic value and dropout (b = 0.042, SE = 0.151, p = .778). However, higher situational intrinsic value significantly predicted lower dropout rates (b = -1.165, SE = 0.408, p = .004: Each additional point in situational intrinsic value reduced the risk of dropping out by approximately 11.1%. This underscores that children's situational intrinsic value immediately before a chapter substantially lowered their likelihood of dropout, while general dispositional intrinsic value did not. For none of the motivational constructs, our results indicated that parental help was significantly related to dropout behavior.

For dropping out after a chapter, no statistically significant link between motivation, parental help and dropout was revealed for any of the motivational constructs (all $p_s > 0.260$) indicating that dropping out between chapters may be mainly due to other reasons (see also limitation section).

7.5 Discussion

7.5.1 Main findings

First, we found considerable within-children variability of situational motivation across all three motivational facets under scrutiny despite a very high overall level of situational motivation (see 7.5.3), extending prior findings of studies conducted with older students in on-site settings (e.g., Flunger et al., 2022; Tsai et al., 2008). These findings reinforce the situational nature of motivation as described in SEVT (Eccles & Wigfield, 2024). Interestingly, much more within-variance (86 %) for utility value was found for post-chapter models compared to pre-chapter models (36 %) suggesting that the particular content of each chapter had a strong, immediate impact on how useful children perceived the material to be.

Second, dispositional motivation did positively predict children's situational motivation, yet only to a very small degree (possibly partly due to ceiling effects, see limitations). Again, this finding aligns well with the situational nature of motivation posing that dispositional traits cannot fully account for moment-to-moment fluctuation of motivation.

Third, the pre- and post-chapter motivational trajectories revealed declining trends, particularly regarding children's situational self-concept and perceived utility value, from the mid-point to the later stages of the course. These observations align with prior empirical research indicating that situational motivation typically declines during online courses (Kyewski & Krämer, 2018). Importantly, this general motivational decline was likely intensified by the increasing cognitive complexity of the content, specifically in Chapters 2 and 3, which introduced the challenging topic of

formal proof construction. From a mathematics-education perspective, these findings were expected, as the transition from chapter 2 to 3 possibly marked a substantial increase in cognitive demand, as chapter 3 required children to construct formal proofs by applying logical reasoning and using symbolic nation (Boolean operators) - a practice commonly only introduced at later stages of education (Sommerhoff & Ufer, 2019).

Fourth, our findings indicated that these motivational declines may have been buffered by the help children received from their parents, specifically for children's perceived utility value. This suggests that parental help may be particularly relevant when children evaluate the usefulness of challenging tasks for school.

Finally, our survival analyses revealed extremely high dropout rates throughout the course despite initial high levels of motivation. This is unfortunate but consistent with prior studies in asynchronous, self-paced online courses across disciplines (Muljana & Luo, 2019; Xavier & Meneses, 2020). Importantly, situational motivation (particularly intrinsic value and self-concept), rather than dispositional motivation, emerged as significant predictors of reducing dropout risk. Unexpectedly, dispositional self-concept was positively associated with increased dropout risk. Possibly, children with higher dispositional self-concept refrained from engaging in challenging course content to safeguard their positive self-image when facing potential failure (see Urdan & Midgley, 2001, for a general discussion on "academic self-handicapping"). These findings cautiously support the idea that momentary motivational states play a more decisive role in self-regulation capabilities (Greene, Bernacki, & Hadwin, 2023), highlighting the need to support children's situational motivation.

7.5.2 Theoretical and Practical Implications

Our results entail three major theoretical contributions. First, our findings extend existing empirical support for motivation's situated nature, showing that even talented children can experience motivational shifts - and sometimes dropout - when content becomes challenging. Our data also provide preliminary evidence that parents, as part of the cultural milieu (Eccles & Wigfield, 2020), may help sustain children's situational motivation in moments of difficulty. In line with SEVT, such findings highlight that the immediate social context can shape motivational dynamics, though we advise that the observed associations should not be overgeneralized to all situational constructs as only the effect for utility value reached significance. Second, the work advances mathematics education research by documenting how studying the challenging topic of proof competency interacts with motivation in a digital environment; the motivational decline at Chapter 3 may be a result of the specific cognitive demands posed by formal proof construction (e.g., Tall, 1999). Third, by adopting within person modelling and survival analysis, we followed recent calls for within-person

designs (e.g., Schukajlow et al., 2023) and illustrated the importance of examining both dispositional and situational motivational constructs simultaneously to better understand learning behavior in online settings.

Practically, this study offers insights into designing self-paced online courses for young talented children in the context of enrichment programs. Sustaining motivation in such environments requires timely support, which may come from parents or, where unavailable, from automated scaffolds such as dynamic feedback or pedagogical agents. Additionally, since dropout primarily occurred within chapters, continuous motivational support is important - not only during active learning moments but also during inactive phases. In this context, first promising evidence suggests that technologies, such as mobile applications, may help children to monitor and plan their learning (Biedermann, Breitwieser, Nobbe, Drachsler, & Brod, 2025).

Taken together, our findings underscore that self-paced online courses can be a promising pathway to provide talented children with cognitively demanding content—independent of time, place, or school constraints. At the same time, the high dropout rates observed in the present study poses a huge threat to the applicability of such courses. Further, the findings highlight that formal proof construction remains a particularly challenging topic, and that sustained motivational and (meta-)cognitive scaffolding may be essential if such enrichment programs are to reach their full potential.

7.5.3 Limitations and Future Research

This study has several limitations. First and foremost, generalizability is limited to highly interested, self-selected children in an extracurricular online course. Consistent with this selective sample, motivation scores were consistently high with low variability (in particular for intrinsic value), indicating ceiling effects. Still, the emergence of significant within-effects suggests that even highly motivated children show some motivational fluctuation in self-paced online courses. Future work should systematically examine motivational dynamics in more diverse samples to test the generalizability of our findings. Second, dropout rates were high, which may cautiously be interpreted as an indication that the course was too demanding—even for talented children who show high levels of motivation. Future iterations of similar courses will benefit from further simplifying content, and incorporating cognitive and motivational scaffolds (e.g., short videos on strategies for monitoring and regulating one's motivation when facing challenging content, in line with metamotivation theory; Miele & Scholer, 2018) to help sustain motivation and potentially reduce dropout. Third, although our data indicate significant links between situational motivation and dropout, we cannot exclude other reasons for dropouts, such as technical difficulties or contextual factors beyond the

scope of our study. Relatedly, the item reflecting parental help may have been interpreted slightly differently by distinct children—for instance as motivational or technical support—rather than solely cognitive help which should be kept in mind when interpreting our findings. Fourth, technical issues in the collection of data caused missing data on some days. While within-chapter dropout detection was unaffected, caution is advised when interpreting dropout between chapters. Fifth, our pre- and post-chapter items assessed anticipated versus experienced aspects of task motivation. This distinction means that pre- and post-scores reflect related but not identical constructs. In particular, the post-chapter self-concept and utility value items both involve ognitive evaluations of the chapter (e.g., 'did well' vs. 'learned useful things'), which was also reflected in their relatively stronger correlation (r = .67, see Supplementary Information 7.6.2). An important avenue for future research is to examine how prospective and retrospective evaluations of motivation interrelate, and how this relationship depends on learners' actual experiences during task engagement. Finally, our findings relied on self-report assessments of motivation, and dropout was inferred solely from questionnaire response patterns. Future research should employ more fine-grained measures, such as digital log data, to better capture risk factors for dropout. Though increasingly used in educational research (e.g., Bernacki et al., 2015), learning analytics based on log data remain mostly unexplored in mathematics education (see Hershkovitz, Noster, Siller, & Tabach, 2024, for a rare example). In addition, it would be valuable to complement the findings with qualitative data (e.g., student or parent interviews) to gain richer insights into learners' reasons for dropping out.

Together, our results underscore the importance of designing responsive online learning environments that support highly interested children's moment-to-moment engagement and scaffold their (meta-) motivational needs. As self-paced online courses continue to grow, attending to the dynamic interplay between personal, contextual, and situational factors becomes an essential endeavour for future mathematics education.

7.6 Appendix D

7.6.1 Supplementary Information 1

This information includes technical details on the analyses for examining research questions 3 and 4. The following equation represents the model applied to each motivational construct (pre- and post-chapters) to estimate successive mean differences of adjacent chapters:

$$sit.\ mot._{ij} = \gamma_{00} + d_{(2.1)}D_{(2,1),ij} + d_{(3.2)}D_{(3,2),ij} + d_{(4.3)}D_{(4,3),ij} + d_{(5.4)}D_{(5,4),ij} + u_{0j} + e_{ij}$$

γ_{00} is the grand mean motivation across all chapters, $D_{(k,k-1),ij}$ the contrast-coded predictors[5], $d_{(k.k+1)}$ the fixed effects representing the average motivational change between successive chapters, u_{0j} the random intercept for child j and e_{ij} the residual of child j at Chapter i ($i = 1, \ldots, 5$).

7.6.2 Supplementary Information 2

Table 73

Pearson Correlations Between Situational Motivation Across all Chapters

Variable	1	2	3	4	5
1. Self-Concept (pre)	—				
2. Utility Value (pre)	.42***	—			
3. Intrinsic Value (pre)	.41***	.15**	—		
4. Self-Concept (post)	.22***	.07	.12*	—	
5. Utility Value (post)	.26***	.10*	.12*	.67***	—
6. Intrinsic Value (post)	.16**	.03	.11*	.38***	.28***

Note. Correlations are based on all available person-by-chapter observations across all chapters.

**$p < .001$, **$p < .01$, *$p < .05$

7.6.3 Supplementary Information 3

The following tables and figure represent the sensitivity analysis with the *finisher-only* sample (i.e., only children who finished the fifth chapter are included in the following).

The key full-sample including non-finishers pattern replicated: post-chapter utility value and post-chapter self-concept both declined from Chapter 2 to 3 (utility: b = -0.567, $p = .004$; self-

[5]For chapters $k = 1, \ldots, 5$, the contrasts are defined as $D_{(k,k-1),ij} = \begin{cases} \frac{k-6}{5} & for\ i \leq k \\ \frac{k-1}{5} & for\ i > k \end{cases}$, where $i = 1, \ldots, 5$. This contrast coding corresponds to effect coding used to capture mean differences between successive chapters (see Breen, 2018, as an example).

Table 74

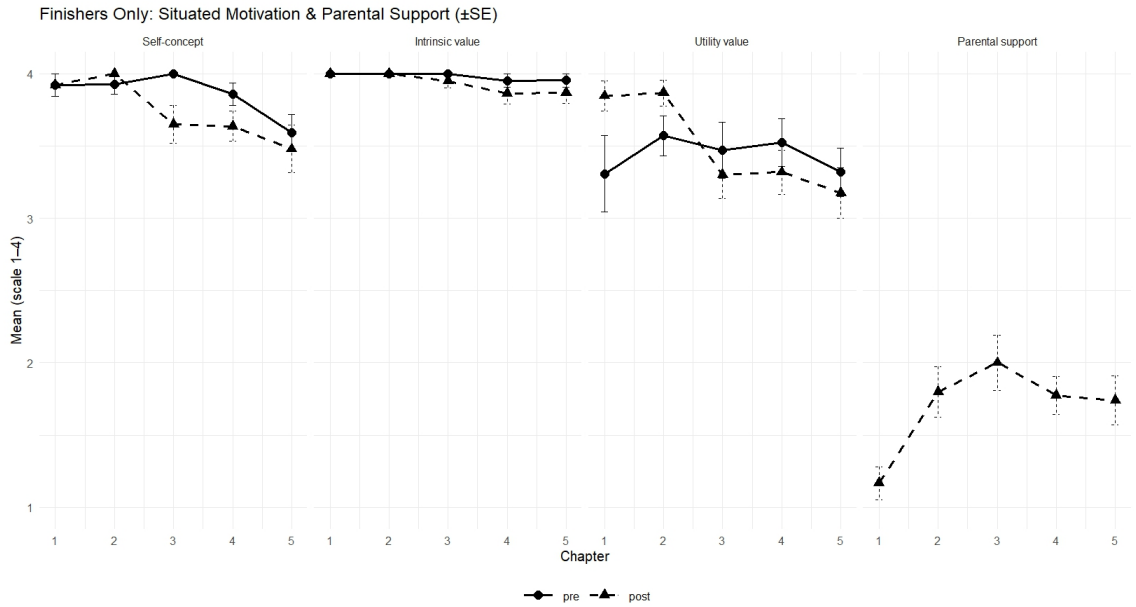
Successive Chapter Contrasts on Situational Motivation

	Pre chapter			Post chapter		
Contrast	**Self-concept**	**Intrinsic value**	**Utility value**	**Self-concept**	**Intrinsic value**	**Utility value**
Intercept	3.88*** (0.04)	3.99*** (0.02)	3.50*** (0.13)	3.75*** (0.08)	3.92*** (0.04)	3.53*** (0.10)
$d_{(2,1)}$	0.00 (0.14)	0.00 (0.05)	0.22 (0.16)	0.06 (0.16)	-0.01 (0.07)	-0.02 (0.21)
$d_{(3,2)}$	0.08 (0.13)	0.00 (0.05)	-0.16 (0.15)	-0.34 (0.15)*	-0.04 (0.07)	-0.57** (0.19)
$d_{(4,3)}$	-0.16 (0.11)	-0.05 (0.04)	0.05 (0.14)	-0.04 (0.13)	-0.09 (0.06)	-0.02 (0.17)
$d_{(5,4)}$	-0.27* (0.11)	0.00 (0.04)	-0.23 (0.13)	-0.15 (0.13)	0.03 (0.06)	-0.14 (0.16)

Note. The intercept represents the grand mean across all chapters. $d_{(k+1,k)}$ represents the mean difference in motivation between successive chapters k and k–1. Standard errors in parentheses. ***$p < .001$, **$p < .01$, *$p < .05$, adjusted p values.

Figure 38

The Trajectories of Situational Motivation and Parental Help Across the Five Course Chapters



Note. Based on finishers only, $N = 23$. Dashed lines indicate post-chapter mean values and solid lines pre-chapter mean value. The numbers in the figure self-concept indicate the numbers of children dropping out during the respective chapter.

Table 75

Moderating effects of within-person parental help on motivational changes

Predictor	Self-concept	Intrinsic value	Utility value
Intercept	3.78 (0.26)***	3.70 (0.13)***	3.47 (0.52)***
$d_{(2,1)}$	-0.07 (0.22)	0.02 (0.10)	0.23 (0.28)
$d_{(3,2)}$	-0.40 (0.14)**	-0.05 (0.07)	-0.70 (0.18)***
$d_{(4,3)}$	0.02 (0.13)	-0.08 (0.06)	0.11 (0.16)
$d_{(5,4)}$	-0.14 (0.12)	0.03 (0.05)	-0.14 (0.15)
Help	0.07 (0.09)	-0.05 (0.04)	0.02 (0.12)
$d_{(2,1)}$ × help	-0.49 (0.39)	-0.11 (0.18)	0.26 (0.49)
$d_{(3,2)}$ × help	0.36 (0.28)	0.16 (0.13)	0.66 (0.35)
$d_{(4,3)}$ × help	-0.48 (0.26)	-0.21 (0.12)	-0.49 (0.33)
$d_{(5,4)}$ × help	0.52 (0.26)	0.19 (0.12)	0.10 (0.33)

Note. Intercept = grand mean across chapters (at average help). $d_{(k+1,k)}$ = mean difference between successive chapters (negative values indicate declines). Help = within-child deviations from average. Interaction terms indicate whether parental help alters the motivational change between two chapters: positive values mean that more-than-usual help buffered declines or amplified gains, while negative values mean it intensified declines or dampened gains. SEs in parentheses.

***$p < .001$, **$p < .01$, *$p < .05$

concept: b = -0.344, $p = .023$), whereas intrinsic value showed no systematic change. For pre-chapter self-concept, the decline from Chapter 4 to 5 remained significant (b = -0.269, $p = .015$); the decline from chapter 3 to 4 was similar in magnitude to the full sample but not statistically significant with the smaller sample. Parental-help effects attenuated (Table S3), consistent with reduced statistical power, although the directions of the interaction effects were the same as in the full-sample analyses.

8 Results and Discussion

The main goal of this dissertation was to determine how enjoyable and effective proof learning can be for talented primary school children in an asynchronous online setting. Given the suitability of proof learning for gifted children (e.g., Bardy & Bardy, 2020) and their ability to work at their own pace (Risemberg & Zimmerman, 1992), an enrichment course on this content and of this format seemed promising. To investigate the potential of such a concept, I developed an asynchronous online course on mathematical proving for gifted primary school children.

In a pilot study (Study 1, Section 4), this course was tested for its general feasibility and for content and design choices that needed revision. After the course had been refined, the goal was to measure its effect on the children's PC. For this purpose, a novel test instrument was needed, appropriate for the target group. Therefore, the Preformal Proving Test (PfPT) was developed, which was piloted and validated in Study 2 (Section 5). With Study 3 (Section 6), an efficacy study was then conducted, in which the PfPT was applied alongside motivational measures to determine the effect the course had on performance and proof-related motivation. To delve deeper into the connections between motivational traits, situational motivation, and the proof course, Study 4 analyzed the non-cognitive data from the pre- and post-tests of the efficacy study as well as additional data on the children's motivational state which we collected using single-item indicators throughout the course (Section 7).

8.1 Findings Across the Studies

In Study 1, I present the novel concept of an asynchronous online course on mathematical proving for talented primary school children. Delivered via the online platform of the extracurricular enrichment program HCAP, the course was trialed with n=304 talented third and fourth graders. The study gathered qualitative and quantitative feedback on the course and its components from both the children and their legal guardians. Furthermore, log data analysis allowed for the investigation of dropout numbers and patterns. The feedback from both groups was generally favorable of the chosen topic and format. Children and adults stated that the course was enjoyable and promoted learning opportunities. The interactive elements, like *explorations, games* and *cases*, were much appreciated. According to the ICAP (Interactive Constructive Active Passive) framework, such interactive learning tools are also the most effective (Chi & Wylie, 2014). The log data analyzes indicated that about 4 in 10 children finished the course. Compared to the majority of asynchronous online courses in which about 1 in 10 children completes the course (Eriksson et al., 2016). A possible explanation for this rate is the higher proficiency in self-regulated learning that gifted children

tend to possess in comparison to their pears (Risemberg & Zimmerman, 1992) or the structured program within the course was offered (see Trautwein et al., 2023, for details about the HCAP). However, a closer look at the dropout patterns revealed that the natural language proof writing activities formed a considerable bottleneck, being related to higher numbers of discontinuation than other course elements. Also, most of the critical feedback addressed these activities: The participants wished for an implementation without keyboard use, featuring more worked examples and instructions on demand. These valuable comments formed the basis for the revision of the course before the consecutive studies.

In Study 2, I investigated the nomological net of PC, including reasoning (gf), crystallized intelligence (gc), and computational thinking (CT). As existing PC assessments so far only targeted learners from secondary school and above (e.g., Kempen & Biehler, 2014; Senk, 1989), I developed a broad, child-appropriate item set to investigate PC in children without any prior knowledge of formal mathematics. The items resembled different proof steps from Boolean Logic, Set Theory, and Elementary Number Theory using an iconic representation. This item set was piloted in a large-scale online study with n=409 children. The study provided valuable insights and allowed us to sample items for a final, shorter test with promising item statistics. In a second study, which was conducted in a proctored on-site setting with n=180 children. This shortened test was rolled out accompanied by instruments for measuring the abovementioned related constructs in order to assess the validity of the test. Additionally, proof-related motivational self-reports were included to investigate how the test result correlates with motivational belief levels. The studies showed a broad difficulty range for both the initial and the shortened item set. The results regarding the familiar cognitive constructs suggest that PC is distinct from CT, gf, and gc, but still related to each of these. Unlike it was found for other cognitive constructs (e.g., Hansford & Hattie, 1982), no significant relation between PC and the corresponding domain-specific motivational beliefs was observed. A threefold correlated factor model of PC with the underlying dimensions of Boolean Logic, Set Theory, and Elementary Number Theory showed satisfying statistical properties. Altogether, the results from the online pilot study and the on-site validation study implied that the resulting Preformal Proving Test (PfPT) is valid and feasible to be used in proof studies with primary school children.

In Study 3, we used the PfPT as well as motivational self-reports to measure the effects of the revised intervention course regarding PC and domain-specific non-cognitive constructs. For this online efficacy study, I chose the design of a randomized controlled trial. A total of n=269 participants registered. From this sample, 132 children were randomly assigned to the treatment group, while the other 137 children were allocated to the wait list control group. The study revealed no intervention effects on PC, proof interest, intrinsic value, persistence, or attainment value for

proving. Instead, it yielded negative intervention effects on the students' proof self-concept. This could be explained by the fact that the children, who previously had not encountered proving and voluntarily registered for this mathematical enrichment course, started the course with a generally high self-concept related to all mathematical tasks (Hoge & Renzulli, 1993). Throughout the course, they most likely reassessed their self-concept after exploring this very new field (see Kruger & Dunning, 1999). Considering the incomplete course participation by many children in the treatment group, it seems possible that potential effects on PC remained undetected, as the children did not receive the full treatment. Therefore, dropout and situational motivation in this course afford further investigation so the treatment can be refined. Even though gifted children tend to be more proficient in self-regulated learning (Risemberg & Zimmerman, 1992), it seems like more scaffolding is needed in this context.

Finally, to further explore what prevented the children from successfully participating in an enrichment course basically tailored to their needs, Study 4 examined the motivational state during the runtime of the course more closely. It investigated motivation fluctuation, the relationship between dispositional motivation and situational motivation, and changes in situational motivation between chapters. Furthermore, the impact of parental support on motivation was examined, as were factors that may reduce the risk of dropout. For this purpose, the motivational data from the pre- and post-test in Study 3 were used, alongside log data and motivational single-item indicators woven into the course, collected from the sample of Study 3. This study revealed notable within-children variability in line with the situational character of motivation proposed by Eccles and Wigfield (2024). The children's dispositional motivation positively predicted this situational motivation, but only to a very small degree. In the course of the intervention, decreasing trends regarding self-concept and utility value emerged, which is in line with prior research indicating that situational motivation declines during online courses (Kyewski & Krämer, 2018). This observation was particularly evident in Chapters 2 and 3, which introduced formal proof construction. Parental help, however, seemed to have buffered these motivational declines. Still, our analyzes revealed high dropout rates, consistent with prior studies in asynchronous online courses (Muljana & Luo, 2019; Xavier & Meneses, 2020). Situational rather than dispositional motivation emerged as a significant predictor of reduced dropout risk, whereas dispositional self-concept was associated with an increased dropout risk. Possibly, children with higher dispositional self-concept refrained from continuing with the challenging parts of the course to safeguard their positive self-concept (Urdan & Midgley, 2001). These findings can guide further course refinements, addressing the still challenging chapters on formal mathematics as well as additional scaffolding to provide more help independent of parental presence. Additionally, incorporating intervention components that foster situational

motivation could provide an opportunity to reduce dropout rates.

8.2 Implications and Future Directions

This dissertation followed the research goal of examining if an enrichment course on mathematical proving can be successfully implemented in a self-paced online environment. In the pursuit of this aim, I took a theoretical perspective on proving as a cognitive skill and developed a new test instrument to measure PC, corresponding to a three-dimensional measurement model. The findings from the investigations of the self-paced course and the test instrument carry implications for future research as well as for educational practice. In the following I will discuss which potential research questions arise from these findings. After that, I will point out possible directions for educational practice that emerge from the results of my research.

8.2.1 Implications for Future Research

Mathematical Proving as Enrichment The overarching question of this work was how enjoyable and effective proof learning can be for talented primary school children in an asynchronous online course. Therefore, in Studies 1 and 3, a novel enrichment course on mathematical proving in an asynchronous online setting was examined for its feasibility and efficacy. Generally, the topic of mathematical proving and most of the corresponding exercises were appreciated by the children and their parents. This is in line with previous research on mathematical giftedness that suggests proving and related activities to be suitable enrichment content (Bardy & Bardy, 2020; Krutetskii, 1976; Käpnick, 1998). However, a majority of children were unable to complete the course, and, likely due to this circumstance, no significant effect on the children's PC was observed. To identify potential room for improvement, I will now revisit the four core components that guided the implementation of the course through a critical lense.

Iconic and symbolic reasoning make use of the two more abstract modes of representation for mathematical content from the EIS-model (Lambert, 2011). In the intervention course, the transition from iconic to symbolic reasoning was mostly taught in the so-called *explorations*, which were quite popular among the children. After that, in the proof activities, the children were asked to write their own proofs, solely using symbolic reasoning. These activities were rather unpopular and corresponded to dropout and loss of motivation (see: Study 1, 3, and 4). According to Selden (2013), proving involves two aspects: chains of reasoning and the laws of formal mathematics. The results of my research suggest that – at least at primary school level – the reasoning-related part of proving could be more suitable for children, and the transition from iconic to symbolic (i.e., formal) representations needs to happen with more intermediate steps. Studies with a less formal design in

the same age group could examine these ideas further.

Natural language proof writing in the drill-and-practice environment has worked well in previous studies with proof novices in a university context (Carl et al., 2022). However, in the context of the primary school enrichment course, it was not perceived as useful and did not significantly enhance the children's proof skills, despite the surrounding, supportive course environment and the exclusion of secondary school mathematical content. One possible explanation is that the written natural language may have posed a severe challenge to the children, as their literacy is still developing. Controlling for prior reading and writing skills could determine if these are factors that reduce the risk of dropout and moderate possible learning effects.

Automated real-time feedback was included in all interactive course elements. Receiving this feedback was perceived quite positively by the course participants. Additionally, parents and children wished for more detailed feedback prompts. This is in line with van der Kleij, Feskens, and Eggen (2015), who found that immediate feedback was more helpful than delayed feedback and elaborate feedback was better than short feedback.

The last core component was self-paced learning. Previous research suggests that marked self-regulation could be a characteristic of giftedness (Risemberg & Zimmerman, 1992) and that learning autonomy is essential for enrichment (Brink, 2025). However, most students' motivation declined during the course, and many of them did not complete it. Parental presence, on the other hand, decreased the dropout risk. This could imply that, even if gifted primary school children have a huge potential for self-regulated learning, their self-regulation skills must develop further for them to be able to profit from such learning scenarios. Until then, external guidance might be necessary or additional self-regulation training, as Risemberg and Zimmerman (1992) suggested to foster gifted children's academic development.

Measuring Proof Competency New technologies not only have the potential to create asynchronous online courses but also offer the chance to create adaptive, self-regulated courses through online assessment tools which can evaluate the respective competences on the spot (Leikin, 2021). In Study 2, a tool to measure PC without the use of formal expressions, the Preformal Proving Test (PfPT), was developed. The test is time-efficient, as it only takes 15 minutes, and has a colorful design, appropriate for primary school children. It was shown to be feasible in both an online and a supervised on-site setting. As the study design was not exactly parallel, it remains for future research to determine if the test outcomes differ between the two settings. Such differences could be caused by the presence of a test supervisor, generating audience effects on executive functions (Camos et al., 2025).

In the course of these two studies, a threefold correlated factor model of PC evolved, featuring three sub-dimensions, namely Boolean Logic, Set Theory, and Elementary Number Theory. It meets the standards for close fit (RMSEA) and fair fit (CFI) according to McNeish (2018).

Furthermore, the nomological net of PC was examined. The results suggest that PC is related to, yet distinct from, CT, gf, and gc. The familiarity of gf, which constitutes the reasoning dimension of current intelligence models (W. Schneider & McGrew, 2012), with PC emphasizes the close connection between proving and reasoning that is postulated by a substantial part of researchers within the didactics of mathematics (G. Stylianides, 2008; G. Stylianides & Silver, 2007; Thompson et al., 2012). The correlation of PC and CT suggests a similar overlap between proving and problem solving, as CT is a skill applied to solving problems (Román–González et al., 2018) and is highly correlated to problem solving abilities (Tsarava et al., 2019). Additional studies are needed to explore if the correlation persists when PC is compared to tests assessing problem solving ability on a more general level. Lastly, studies could be undertaken to investigate the relation of PC to other cognitive constructs, to determine its position relative to recent intelligence models like the Cattell-Horn-Carroll established by (W. Schneider & McGrew, 2012).

Open Qustions Apart from the aspects discussed in the previous paragraphs, there are some additional directions in which further research could go. Firstly, it might be worthwhile to investigate the intervention described here in a digital pull-out program, allowing talented children to work through the course during a school lesson. Such fostering programs can be beneficial for mathematically gifted children (Huth et al., 2024; Kohnen & Fischer–Ontrup, 2023). In a digital pull-out program, children could benefit from the structured environment and the teacher's presence in the room, which could help compensate for their not yet fully developed self-regulation strategies. Also, it remains open whether older children, for example from secondary school, could work through the self-paced course with a higher completion rate. This would allow for additional studies on the effectiveness of the training. Lastly, one could investigate the approaches of self-paced learning and proving separately for the target group of gifted primary school children by offering a synchronous course on mathematical proving or a self-paced course on another topic. Comparing the results to those of the original intervention could provide answers to the question whether the challenges that come with each approach may be overcome more easily if they are not combined.

8.2.2 Implications for Educational Practice

In this section, I will shine a light on the practical implications of this dissertation and how they contribute to the field of gifted education, especially at the primary school age.

Firstly, I would like to stress the recommendation to acknowledge mathematical proving as appropriate enrichment content on par with the more popular fields of reasoning and problem solving. Proving is not only listed alongside these fields in several works on mathematical giftedness (e.g., Bardy & Bardy, 2020) but also shows close familiarity with each of these disciplines and a substantial fit with the characteristics of mathematically gifted children (Krutetskii, 1976). This dissertation endorses this fit by demonstrating a high demand (enrollment) for a proof enrichment course and generally positive feedback regarding the topic and content.

Secondly, my dissertation gives advocates that it is technically possible to teach formal proving to gifted primary school children. However, only some of the examined children managed to complete the course in full. Thus, educators should consider the duality of proving, as it consists of a formal and a reasoning-related part (Selden, 2013). My research adds to the existing evidence that the formal part poses several challenges to young children while the reasoning part can be quite motivating to them. Starting with very broad proof standards and the classroom definition of proving (A. Stylianides, 2007), the reasoning part can be practiced, with the formal part slowly being introduced in parallel. For different levels of proficiency, increasingly formal proof approaches can then follow, accompanied by steady reflection on the difference between the respective approach and formal proving (Biehler & Kempen, 2016; Wittmann & Müller, 1988).

Furthermore, when designing self-paced courses for primary school children, educators should consider to install an adult in the role of a mentor within the child's reach who can assist with technical difficulties and reading, and who can motivate the child to keep going. As Study 4 has shown, this is recommendable because parental help significantly reduced the risk of dropping out of the self-paced intervention.

Lastly, this dissertation presents the online course *Logical Detectives* as the first asynchronous online course within the HCAP. Although the pilot and efficacy study indicate that the course requires further refinement due to its difficulty, the children greatly appreciated the design and the automated feedback. I recommend using the course with the assistance of an adult until a revised version with more scaffolding and an advanced paedagogical agent is released. Asynchronous online courses like this one can add to the rich program of the Hector Children's Academies, allowing many children to participate who otherwise would not have gotten the chance, due to a limited number of course seats or because of geographical obstacles.

8.3 Strengths and Limitations

To critically interpret the results of my dissertation, several strengths and limitations of the included studies need to be considered.

A major strength is the large sample size across all four studies which the online course and test design made possible. Children from various locations in the state of Baden-Württemberg could participate in the intervention, allowing for a more heterogeneous sample than an on-site study could have yielded. This speaks for the results of the study being unbiased from the influence of a specific area.

The second strength lies in the design process for the investigated course on mathematical proving. Here, I followed recent literature on stepwise intervention development (Herbein et al., 2018; Humphrey et al., 2016; Lendrum & Wigelsworth, 2013): In a first step, the topic of mathematical proof was identified as an extracurricular and challenging topic that showed a promising overlap with the needs and characteristics of gifted children. Thus, it was likely to match the needs of the target group. Next, the intervention was developed, grounded on core components derived from previous research, and evaluated in a pilot study. After that, the revised intervention was examined in an efficacy study. This procedure already covers four of the six steps recommended by Herbein et al. (2018) for achieving an intervention that is both feasible and effective.

One more strength is the design of the efficacy study: It was implemented as a randomized controlled trial (RCT), which is often referred to as the gold standard for intervention studies (Lendrum & Wigelsworth, 2013). Following the RCT design, all registered children participated in a pre-test after which about half of the children were randomly assigned to the treatment group, the others to the control group. The treatment group was given access to the course for the following three weeks. After that, a post-test was carried out involving all children. In this way, the development of both groups could be compared and the effect of the intervention could be differentiated from other influences, such as retest effects (National Research Council, 2004). For ethical reasons, the control group was given access to the course for the three weeks following the post-test.

Furthermore, the PfPT that was used in the intervention study had been piloted and validated in an online and on-site setting. This reduces the risk of the test results being influenced by technical difficulties. Also, no ceiling effects were found in the online study and the item difficulty showed a broad spectrum in both studies, indicating that it was indeed the children and not their parents who completed the test at home.

Finally, according to Nelson et al. (2012), the influence of any course component on the intervention outcomes is affected by the fidelity of the course instructor who implements it. In an asynchronous online course, all children are presented with the same predefined activities, and a single course instructor can answer all support queries. Thus, in the intervention presented here, it can be assumed that the intervention fidelity was not only independent of the children's location but also

very high. This allows for conclusions directly addressing the suitability of the course components without distortions and fast adjustments to the course after each iteration.

Nevertheless, the limitations of my research need to be considered as well to ensure a thorough reflection.

Firstly, the samples of all studies included in this dissertation were recruited from the HCAP which selects talented children for enrichment courses based on teacher nominations (see Trautwein et al., 2023). These children demonstrate above-average interest, dedication, and skills. This limits the applicability of the results, as this sample is not representative of all primary school children. Thus, one cannot derive a general suitability of the topic of proving for primary school children from the results of this research.

Also, the ceiling effects in the motivational pretest, which are most likely caused by the previously mentioned sample composition, limit the significance of the motivational findings of the efficacy study. Generally, the fact that motivation was only measured via self-reports constitutes a weakness of the study. Future research could include other indicators, such as parental observation records or behavioral traces, to assess the children's motivational state and traits.

A major limitation lies in the high dropout rates registered in Studies 1,3, and 4. The dropout rates do not only indicate that specific aspects of the course design need to be improved, they also decrease the sample size for all analyzes of course log data, chapter questionnaires, and post-course measures. This limits the significance of the observations from the efficacy study.

Lastly, it remains to mention that the course was held in an unsupervised online setting. Therefore, it cannot be controlled if the children did indeed only receive the amount of parental help they indicated in the post-test. Also, other domestic factors that may have influenced the children's engagement in the course could have remained undetected (e.g., presence of siblings, profession of parents). Future studies could include such questionnaires or trial the course in a classroom setting.

8.4 Conclusion

At the core of this dissertation, the mathematical proof intervention *Logical Detetctives*, an asynchronous online course for gifted and talented children, was developed and evaluated. Firstly, the intervention's feasibility was assessed in a large-scale online study. Then an efficacy study was carried out as a randomized controlled trial to examine the intervention's effects on its participants. For this purpose, a new measurement tool to assess children's mathematical PC without formal language was developed and validated. For the current course version, no effects on the children's PC could be detected. In general, the use of formal language and the self-regulated learning environment challenged the children while the extracurricular content of mathematical proving was perceived as

interesting. Therefore, it seems worthwhile to further explore the possibilities of teaching this topic to gifted primary school children.

References

Aloni, M., & Harrington, C. (2018). Research based practices for improving the effectiveness of asynchronous online discussion boards. *Scholarship of Teaching and Learning in Psychology*, *4*(4), 271—289.

Amaefule, C. O., Britzwein, J., Yip, J. C., & Brod, G. (2025, April). Children's perspectives on self–regulated learning: A co–design study on children's expectations towards educational technology. *Education and Information Technologies*, *30*(5), 6117—6140. Retrieved 2025–05–22, from `https://doi.org/10.1007/s10639-024-13031-0`

Arens, A., Trautwein, U., & Hasselhorn, M. (2011, January). Erfassung des Selbstkonzepts im mittleren Kindesalter: Validierung einer deutschen Version des SdQ I. *Zeitschrift für Pädagogische Psychologie*, *25*(2), 131—144.

Arizmendi, C., Bernacki, M., Raković, M., Plumley, R., Urban, C., Panter, A., ... Gates, K. (2022, 8 26). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, *55*(6), 3026–3054.

Ball, D., & Bass, H. (2003). Making mathematics reasonable in school. In D. Schifter, J. Kilpatrick, & W. G. Martin (Eds.), (p. 27—44). National Council of Teachers of Mathematics.

Bandura, A. (1977). *Social learning theory.* Prentice–Hall.

Bardy, T., & Bardy, P. (2020). *Mathematisch begabte Kinder und Jugendliche.* Springer.

Barr, D., Harrison, J., & Conery, L. (2011). Computational thinking: A digital age skill for everyone. *Learning and leading with technology*, *38*, 20–23.

Bass, H. (2011, jan). Proof in mathematics education: An endangered species? A review of teaching and learning proof across the grades: A k–16 perspective. *Journal for Research in Mathematics Education*, *42*(1), 98—103.

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). *ASKU – Allgemeine Selbstwirksamkeit Kurzskala.* ZPID (Leibniz Institute for Psychology) – Open Test Archive.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238—246.

Bernacki, M., Nokes–Malach, T., & Aleven, V. (2015). Examining self–efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning*, *10*(1), 99–117.

Bickerton, R. T., & Sangwin, C. J. (2021, March). Practical online assessment of mathematical proof. *International Journal of Mathematical Education in Science and Technology*, *53*(10), 2637—2660.

Bicknell, B. (2008). Gifted students and the role of mathematics competitions. australian primary mathematics classroom. *Australian Primary Mathematics Classroom*, *13*(4), 16 - 20.

Biedermann, D., Breitwieser, J., Nobbe, L., Drachsler, H., & Brod, G. (2025, 2). Memorizing plans with an app: Large individual differences in the effectiveness of retrieval–based and generative learning activities in a naturalistic context. *Learning and Individual Differences*, *118*, 102641.

Biehler, R., & Kempen, L. (2016, March). Didaktisch orientierte Beweiskonzepte – Eine Analyse zur mathematikdidaktischen Ideenentwicklung. *Journal für Mathematik–Didaktik*, *37*(1), 141—179. Retrieved 2025–07–21, from https://doi.org/10.1007/s13138-016-0097-1

Bleiler, S. (2009). Integration of proof and reasoning in school mathematics: A guide for elementary school teachers. *Dimensions in Mathematics*, *29*(2), 6–12.

Bock, H., & Borneleit, P. (2000). Logisches denken – Einige Gedanken über ein altes Ziel und seine Verfolgung im Mathematikunterricht. In L. Flade & W. Herget (Eds.), *Mathematik: Lehren und lernen nach timss* (p. 59—68). Volk und Wissen.

Boero, P. (1999). *Argumentation and mathematical proof: A complex, productive, unavoidable relationship in mathematics and mathematics education* (Vol. 7) (No. 8). International newsletter on the teaching and learning of mathematical proof. Retrieved 2024–11–15, from http://www.lettredelapreuve.org/OldPreuve/Newsletter/990708Theme/990708ThemeUK.html

Breitner, J. (2016). Visual theorem proving with the incredible proof machine. In J. C. Blanchette & S. Merz (Eds.), *Interactive Theorem Proving* (p. 123—139). Springer International Publishing.

Brink, H. (2025, April). The complexity, autonomy, authenticity, and support (caas) framework for gifted students' needs in technology education: A systematic literature review. *Roeper Review*, *47*(2), 125—135. Retrieved 2025–08–01, from https://doi.org/10.1080/02783193.2025.2466514

Bronkhorst, H., Roorda, G., Suhre, C., & Goedhart, M. (2021, June). Student development in logical reasoning: Results of an intervention guiding students through different modes of visual and formal representation. *Canadian Journal of Science, Mathematics and Technology Education*, *21*(2), 378—399. Retrieved 2024–10–15, from https://doi.org/10.1007/s42330-021-00148-4

Browne, M. W., & Cudeck, R. (1992, November). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230—258.

Bruner, J. S., Olver, R. R., & Greenfield, P. M. (1971). *Studien zur kognitiven Entwicklung. Eine*

kooperative Untersuchung am "center for cognitive Studies" der Harvard–Universität. 1. Aufl. (H. Aebli & J. R. Hornsby, Eds.) [gedruckt]. Klett.

Brunner, E. (2014). *Mathematisches Argumentieren, Begründen und Beweisen: Grundlagen, Befunde und Konzepte*. Springer.

Brunner, M., Nagy, G., & Wilhelm, O. (2012, June). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796—846.

Cadwallader Olsker, T. (2011, January). What do we mean by mathematical proof? *Journal of Humanistic Mathematics*, *1*(1), 33—60.

Camos, V., Mariz Elsig, S., Öncü, Y., Wohlhauser, M., & Belletier, C. (2025, April). Does the experimenter presence impact children's working memory? *Cognitive Development*, *74*, 101569.

Carl, M. (2022). *Diproche – Entwurf, Umsetzung und Erprobung eines automatischen Systems zur Unterstützung des Beweisenlernens bei StudienanfängerInnen* (Unpublished doctoral dissertation). Europa Universität Flensburg.

Carl, M., Lorenzen, H., & Schmitz, M. (2022). Natural language proof checking in introduction to proof classes — first experiences with diproche. *Electronic Proceedings in Theoretical Computer Science*, *354*, 59—70. Retrieved 2022–12–22, from `http://arxiv.org/abs/2202.08131`

Cattell, R. (1957). Personality and motivation theory based on structural measurement. In *Psychology of personality: Six modern approaches.* (p. 63–119). Logos Press.

Cervantes-Barraza, J. A., Hernandez Moreno, A., & Rumsey, C. (2020). Promoting mathematical proof from collective argumentation in primary school. *School Science and Mathematics*, *120*(1), 4—14. Retrieved 2025–07–25, from `https://onlinelibrary.wiley.com/doi/abs/10.1111/ssm.12379`

Chan, N., & Kennedy, P. E. (2002, April). Are multiple-choice exams easier for economics students? a comparison of multiple-choice and "equivalent" constructed-response exam questions. *Southern Economic Journal*, *68*(4), 957—971.

Chang, H., Chong, Y., & Song, S. (2006, 11). Mathematically gifted 6th grade students' proof ability for a geometric problem. *Journal of Educational Research in Mathematics*, *16*(4), 327–344. Retrieved from `https://koreascience.kr/article/JAKO200606140727067.pdf`

Cheema, J. R., & Galluzzo, G. (2013, November). Analyzing the gender gap in math achievement: Evidence from a large–scale us sample. *Research in Education*, *90*(1), 98—112. Retrieved 2025–07–16, from `https://doi.org/10.7227/RIE.90.1.7`

Chen, J., Yun Dai, D., & Zhou, Y. (2013, July). Enable, enhance, and transform: How technology use can improve gifted education. *Roeper Review*, *35*(3), 166—176. Retrieved 2025–08–15,

from https://doi.org/10.1080/02783193.2013.794892

Chen, K., & Jang, S. (2010, July). Motivation in online learning: Testing a model of self–determination theory. *Computers in Human Behavior*, *26*(4), 741—752.

Chi, M. T. H., & Wylie, R. (2014, October). The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219—243. Retrieved 2025–05–23, from https://doi.org/10.1080/00461520.2014.965823

Dai, D. Y., Moon, S., & Feldhusen, J. (1998, March). Achievement motivation and gifted students: A social cognitive perspective. *Educational Psychologist*, *33*(2), 45—63.

Dawkins, P. C., & Weber, K. (2017, June). Values and norms of proof for mathematicians and students. *Educational Studies in Mathematics*, *95*(2), 123—142. Retrieved 2025–08–12, from https://doi.org/10.1007/s10649-016-9740-5

Deal, L., & Wismer, M. (2010, 7). NCTM principles and standards for mathematically talented students. *Gifted Child Today*, *33*(3), 55–65.

Diezmann, C., & Watters, J. (2000, 7). Catering for mathematically gifted elementary students: Learning from challenging tasks. *Gifted Child Today*, *23*(4), 14–52.

Dimitriadis, C. (2016, July). Gifted programs cannot be successful without gifted research and theory: Evidence from practice with gifted students of mathematics. *Journal for the Education of the Gifted*, *39*(3), 221—236.

Dreyfus, T. (1999, March). Why Johnny can't prove. *Educational Studies in Mathematics*, *38*(1), 85—109. Retrieved 2024–12–09, from https://doi.org/10.1023/A:1003660018579

Eccles, J., & Wigfield, A. (2002, 2). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132.

Eccles, J., & Wigfield, A. (2020, 4). From expectancy–value theory to situated expectancy–value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, 101859.

Eccles, J., & Wigfield, A. (2024, 5 17). The development, testing, and refinement of Eccles, Wigfield, and colleagues' situated expectancy–value model of achievement performance and choice. *Educational Psychology Review*, *36*(2), 51.

Elgendi, E., & Shaffer, C. (2020, 2 19). Dynamic concept maps for etextbook glossaries: Design and evaluation. *Frontiers in Computer Science*, *2*, 517867.

Engledowl, C. (2020). Constructing and validating an early algebra assessment. In *Proceedings of the 47th annual meeting of the research council on mathematics learning 2020* (p. 51-58).

Eriksson, T., Adawi, T., & Stoehr, C. (2016, nov 24). Time is the bottleneck: A qualitative study exploring why learners drop out of moocs. *Journal of Computing in Higher Education*, *29*(1),

133—146.

Ertl, B., Hartmann, F., & Heine, J. (2020, 12 9). Analyzing large–scale studies: Benefits and challenges. *Frontiers in Psychology*, *11*, 577410.

Feldman, D. H., & Goldsmith, L. T. (1986). *Nature's gambit: Child prodigies and the development of human potential*. Basic Books.

Flunger, B., Hollmann, L., Hornstra, L., & Murayama, K. (2022, 2). It's more about a lesson than a domain: Lesson–specific autonomy support, motivation, and engagement in math and a second language. *Learning and Instruction*, *77*, 101500.

Foth, M., & van der Meer, E. (2013). Mathematische Leistungsfähigkeit – Prädiktoren überdurchschnittlicher Leistungen in der gymnasialen Oberstufe. In T. Fritzlar & F. Käpnick (Eds.), *Mathematische begabungen: Denkansätze zu einem komplexen themenfeld aus verschiedenen perspektiven* (p. 191—220). WTM.

Fuchs, M. (2006). *Vorgehensweisen mathematisch potentiell begabter dritt–und Viertklässler beim Problemlösen: empirische Untersuchungen zur Typisierung spezifischer Problembearbeitungsstile*. LIT Verlag Münster.

Fölling–Albers, M. (2000). Entscholarisierung von Schule und Scholarisierung von Freizeit? Überlegungen zu Formen der Entgrenzung von Schule und Kindheit. *ZSE: Zeitschrift für Soziologie der Erziehung und Sozialisation*, *20*(2), 118—131. Retrieved 2024–12–09, from `http://nbn-resolving.de/urn:nbn:de:0111-pedocs-109324`

Gadanidis, G., Hughes, J. M., Minniti, L., & White, B. J. G. (2017, August). Computational thinking, grade 1 students and the binomial theorem. *Digital Experiences in Mathematics Education*, *3*(2), 77—96. Retrieved 2024–08–30, from `https://doi.org/10.1007/s40751-016-0019-3`

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Gaspard, H. (2015). *Promoting value beliefs in mathematics: A multidimensional perspective and the role of gender* (Doctoral dissertation).

Gaspard, H., Dicke, A., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, *107*(3), 663—677.

Gernes, D. (1999, May). Sharing teaching ideas: The rules of the game. *The Mathematics Teacher*, *92*(5), 424—429.

Glosauer, T. (2019). *(Hoch)Schulmathematik: Ein Sprungbrett vom Gymnasium an die Uni*. Springer.

Goecke, B., Staab, M., Schittenhelm, C., & Wilhelm, O. (2022, November). Stop worrying about

multiple–choice: Fact knowledge does not change with response format. *Journal of Intelligence*, *10*(4), 102.

Goecke, B., Zimny, L., Hartung, J., Lösche, P., Golle, J., & Wilhelm, O. (2024, November). Measuring cognitive ability in children and adolescents. *Psychological Test Adaptation and Development*. Retrieved 2025–07–17, from `https://econtent.hogrefe.com/doi/10.1027/2698-1866/a000089`

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., ... Preckel, F. (2014, 7). "my questionnaire is too long!" the assessments of motivational–affective constructs with three–item and single–item measures. *Contemporary Educational Psychology*, *39*(3), 188–205.

Golle, J., Schils, T., Borghans, L., & Rose, N. (2022, July). Who is considered gifted from a teacher's perspective? A representative large–scale study. *Gifted Child Quarterly*, *67*(1), 64—79.

Gottfried, A., & Gottfried, A. (1996, 10). A longitudinal study of academic intrinsic motivation in intellectually gifted children: Childhood through early adolescence. *Gifted Child Quarterly*, *40*(4), 179–183.

Grabiner, J. (2012). Why proof? a historian's perspective. In G. Hanna & M. de Villiers (Eds.), *Proving in mathematics education. new ICMI study series* (Vol. 15, p. 147-167). Springer Netherlands.

Greene, J., Bernacki, M., & Hadwin, A. (2023, 9 12). Self–regulation. In P. A. S. . K. R. Muis (Ed.), *Handbook of educational psychology* (p. 314–334). Routledge.

Grieser, D. (2017). *Mathematisches Problemlösen und Beweisen* (2nd ed.). Springer.

Groskurth, K., Bluemke, M., & Lechner, C. M. (2023, August). Why we need to abandon fixed cutoffs for goodness–of–fit indices: An extensive simulation and possible solutions. *Behavior Research Methods*, *56*(4), 3891—3914.

Guo, J., & An, F. (2025, 3 26). Exploring the categories of students' interest and their relationships with deep learning in technology supported environments. *Scientific Reports*, *15*(1), 10370.

Gutierrez, A., & Jaime, A. (1998). On the assessment of the van hiele levels of reasoning. *Focus on Learning Problems in Mathematics*, *20*, 27—46.

H5P Group. (2013). *Course presentation.* Retrieved 2025–08–19, from `https://h5p.org/presentation`

H5P Group. (2014). *Memory game.* Retrieved 2025–08–19, from `https://h5p.org/memory-game`

H5P Group. (2016a). *Interactive video.* Retrieved 2025–08–19, from `https://h5p.org/interactive-video`

H5P Group. (2016b). *Personality quiz.* Retrieved 2025–08–19, from `https://h5p.org/personality-quiz`

H5P Group. (2018). *Image pairing.* Retrieved 2025–08–19, from https://h5p.org/image-pairing

Hanna, G., & Knipping, C. (2020, 8 31). Proof in mathematics education, 1980-2020: An overview. *Journal of Educational Research in Mathematics*, *30*(S), 1-13.

Hansford, B. C., & Hattie, J. A. (1982, March). The relationship between self and achievement/performance measures. *Review of Educational Research*, *52*(1), 123—142. Retrieved 2025–03–25, from https://doi.org/10.3102/00346543052001123

Hasenbein, L., Trautwein, U., Hahn, J., Soller, S., & Göllner, R. (2024). An experimental test of the big–fish–little–pond effect using an immersive virtual reality classroom. *Instructional Science*, *52*(4), 583–612.

Healy, L., & Hoyles, C. (2000, July). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, *31*(4), 396—428. Retrieved 2025–02–27, from https://www.jstor.org/stable/749651

Heinze, A., Cheng, Y., Ufer, S., Lin, F., & Reiss, K. (2008, 4 30). Strategies to foster students' competencies in constructing multi–steps geometric proofs: teaching experiments in taiwan and germany. *ZDM – Mathematics Education*, *40*(3), 443–453.

Heinze, A., Cheng, Y., & Yang, K. (2004). *Students' performance in reasoning and proof in taiwan and germany: Results, paradoxes and open questions.*

Heller, K., & Schofield, N. (1993). International trends and topics of research on giftedness and talent. In K. A. Heller, F. J. Mönks, R. Subotnik, & R. J. Sternberg (Eds.), *International handbook of giftedness and talent* (p. 123-137). Elsevier.

Heller, K. A., Mönks, F. J., Subotnik, R., & Sternberg, R. J. (Eds.). (2000). *International hand book of giftedness and talent* (2nd ed.). Pergamon.

Herbein, E., Golle, J., Tibus, M., Zettler, I., & Trautwein, U. (2018, September). Putting a speech training program into practice: Its implementation and effects on elementary school children's public speaking skills and levels of speech anxiety. *Contemporary Educational Psychology*, *55*, 176–188.

Hersh, R. (1997). *What is mathematics, really?* Oxford University Press.

Hershkovitz, A., Noster, N., Siller, H., & Tabach, M. (2024, 3 3). Learning analytics in mathematics education: the case of feedback use in a digital classification task on reflective symmetry. *ZDM – Mathematics Education*, *56*(4), 727–739.

Hidi, S. (2006, 1). Interest: A unique motivational variable. *Educational Research Review*, *1*(2), 69–82.

Hofstadter, D. (1999). *Gödel, Escher, Bach: An eternal golden braid.* Penguin Books. Retrieved 2025–05–19, from http://archive.org/details/douglas-hofstadter-godel

-escher-bach-an-eternal-golden-braid

Hoge, R. D., & Renzulli, J. S. (1993, December). Exploring the link between giftedness and self–concept. *Review of Educational Research*, *63*(4), 449—465. Retrieved 2025–07–14, from https://doi.org/10.3102/00346543063004449

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010, November). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior*, *26*(6), 1278—1288. Retrieved 2025–05–19, from https://www.sciencedirect.com/science/article/pii/S0747563210001718

Hox, J., Moerbeek, M., & Van De Schoot, R. (2017, 9 14). Multilevel analysis. In *Multilevel analysis: Techniques and applications (3. aufl.)*. Routledge.

Hu, L., & Bentler, P. M. (1999, January). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1—55. Retrieved 2024–09–30, from https://doi.org/10.1080/10705519909540118

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (ipe) for interventions in educational settings: An introductory handbook*. Education Endowment Foundation.

Huth, M., Pollok, J., & Schreiber, C. (2024, January). Digitale drehtür Hessen Mathematik – Mathematik für Interessierte. In B. B. . L. Bröll (Ed.), *Digitaler mathematikunterricht in forschung und praxis ii. tagungsband zur vernetzungstagung 2023 in siegen* (p. 129-144). WTM-Verlag.

Häsä, J., Westlin, L., & Rämö, J. (2023, 7 15). Undergraduate students' attitudes towards mathematical proving in an introduction to proof course. *Educational Studies in Mathematics*, *114*(3), 393–415.

Jablonski, S., & Ludwig, M. (2022, December). Examples and generalizations in mathematical reasoning – a study with potentially mathematically gifted children. *Journal on Mathematics Education*, *13*(4), 605—630.

Jasmin, D. R., & Ongcoy, P. J. (2024, November). Study on school students' blended learning experiences and mathematical self–concept during covid–19. *Journal of Learning for Development*, *11*(3), 492—501. Retrieved 2025–07–14, from https://jl4d.org/index.php/ejl4d/article/view/1136

Jones, E. F., Pritchard, A., Jacobson, L. A., Mahone, E. M., & Zabel, T. A. (2021, June). How much testing can a kid take? feasibility of collecting pediatric patient experience ratings of neuropsychological and psychological assessment. *Applied Neuropsychology: Child*, *11*(4),

610—617.

Jonsson, B., Mossegård, J., Lithner, J., & Karlsson Wirebring, L. (2022, January). Creative mathematical reasoning: Does need for cognition matter? *Frontiers in Psychology*, *12*.

Keller, M., Yanagida, T., Lüdtke, O., & Goetz, T. (2025, 3). How similar are students' aggregated state emotions to their self–reported trait emotions? results from a measurement burst design across three school years. *Educational Psychology Review*, *37*(1), 26.

Kempen, L. (2019). *Begründen und Beweisen im Übergang von der Schule zur Hochschule*. Springer.

Kempen, L., & Biehler, R. (2014). The quality of argumentations of first–year pre–service teachers. *International Group for the Psychology of Mathematics Education*. Retrieved from `https://api.semanticscholar.org/CorpusID:228162626`

Kennedy, P., & Walstad, W. (1997, 10). Combining multiple–choice and constructed–response test scores: An economist's view. *Applied Measurement in Education*, *10*(4), 359–375.

Kiesswetter, K. (1985). Die Förderung von mathematisch besonders begabten und interessierten Schülern – ein bislang vernachlässigtes sonderpädagogisches Problem. Mit Informationen über das Hamburger Modell. [gedruckt]. *Der mathematische und naturwissenschaftliche Unterricht*, *38*(5), 300—306.

Kline, M. (1973). *Why Johnny can't add: The failure of the new math*. St. Martin's Press.

Ko, J., & Song, S. (2011). Effect of proof education through informal activities on the proof abilities of students in the elementary gifted class. *Journal of Korea Society of Educational Studies in Mathematics - School Mathematics*, *13*, 501–524.

Kohnen, M., & Fischer–Ontrup, C. (2023, June). Die digitale drehtür. Enrichmentangebot für Schüler/innen und Praxiserfahrung für Studierende. In P. P. Christian Fischer (Ed.), *Aufholen nach corona? was schule zu mehr bildungsgerechtigkeit beitragen kann.* (Münstersche Gespräche zur Pädagogik ed., Vol. 39, p. 131—138). Waxmann.

Koshy, V., Ernest, P., & Casey, R. (2009, 3 15). Mathematically gifted and talented learners: Theory and practice. *International Journal of Mathematical Education in Science and Technology*, *40*(2), 213–228.

Krach, S. K., Paskiewicz, T. L., & Monk, M. M. (2020, December). Testing our children when the world shuts down: Analyzing recommendations for adapted tele–assessment during covid–19. *Journal of Psychoeducational Assessment*, *38*(8), 923—941.

Krapp, A. (1993). Bedingungen und Auswirkungen berufsspezifischer Lernmotivation in der kaufmännischen Erstausbildung. In *Bildung zwischen staat und markt* (Vol. 39, p. 55–57). VS Verlag für Sozialwissenschaften.

Krawitz, J., & Schukajlow, S. (2018). Do students value modelling problems, and are they confident they can solve such problems? value and self–efficacy for modelling, word, and intra–mathematical problems. *ZDM – Mathematics Education*, *50*(1–2), 143–157.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self–assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121—1134.

Krutetskii, V. (1976). *The psychology of mathematical abilities in schoolchildren.* The University of Chicago Press.

Kyewski, E., & Krämer, N. (2018, 3). To gamify or not to gamify? An experimental field study of the influence of badges on motivation, activity, and performance in an online learning course. *Computers & Education*, *118*, 25–37.

Kyllonen, P. C. (2020, July). Reasoning abilities. In *Oxford Research Encyclopedia of Education.* Retrieved 2025–07–08, from `https://oxfordre.com/education/display/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-878`

Käpnick, F. (1998). *Mathematisch begabte Kinder. Modelle, empirische Studien und Förderungsprojekte für das Grundschulalter.* [gedruckt]. Lang.

Lakatos, I. (2015). *Proofs and refutations: The logic of mathematical discovery* (J. Worrall & E. Zahar, Eds.). Cambridge University Press. Retrieved 2025–07–21, from `https://www.cambridge.org/core/books/proofs-and-refutations/A11138AE31DB52797A6E4C3F1856E1CB`

Lambert, A. (2011). Was soll das bedeuten? Enaktiv – ikonisch – symbolisch. Aneignungsformen beim Geometrielernen. In A. Filler & M. Ludwig (Eds.), *Vernetzungen und anwendungen im geometrieunterricht ziele und visionen 2020. ak geometrie* (p. 5—32).

Lau, E. Y. H., Jian–Bin, L., & Lee, K. (2021, August). Online learning and parent satisfaction during covid–19: Child competence in independent learning as a moderator. *Early Education and Development*, *32*(6), 830—842. Retrieved 2025–05–20, from `https://doi.org/10.1080/10409289.2021.1950451`

Lee, K. (2016, March). Students' proof schemes for mathematical proving and disproving of propositions. *The Journal of Mathematical Behavior*, *41*, 26—44.

Lee, K. H. (2005). Mathematically gifted students' geometrical reasoning and informal proof. In H. L. Chick & J. L. Vincent (Eds.), *Proceedings of the 29th conference of the international group for the psychology of mathematics education* (Vol. 3, pp. 241–248). PME.

Lee, Y., Choi, J., & Kim, T. (2013). Discriminating factors between completers of and dropouts from online learning courses. *British Journal of Educational Technology*, *44*(2), 328–337.

Leifheit, L., Tsarava, K., Moeller, K., Ostermann, K., Golle, J., Trautwein, U., & Ninaus, M.

(2019, 10 23). Development of a questionnaire on self–concept, motivational beliefs, and attitude towards programming. In *Proceedings of the 14th workshop in primary and secondary computing education* (p. 1–9). ACM.

Leifheit, L., Tsarava, K., Ninaus, M., Ostermann, K., Golle, J., Trautwein, U., & Moeller, K. (2020, June). Scapa: Development of a questionnaire assessing self–concept and attitudes toward programming. In *Proceedings of the 2020 acm conference on innovation and technology in computer science education* (p. 138—144). ACM.

Leikin, R. (2010, 12). Teaching the mathematically gifted. *Gifted Education International*, *27*(2), 161—175.

Leikin, R. (2021, December). When practice needs more research: The nature and nurture of mathematical giftedness. *ZDM – Mathematics Education*, *53*(7), 1579—1589. Retrieved 2025–08–08, from `https://doi.org/10.1007/s11858-021-01276-9`

Leikin, R., Koichu, B., & Berman, A. (2009, January). Mathematical giftedness as a quality of problem–solving acts. In R. Leikin, B. Koichu, & A. Berman (Eds.), *Creativity in mathematics and the education of gifted students* (p. 115—127). Sense Publishers.

Lendrum, A., & Wigelsworth, M. (2013). Special edition article: The evaluation of school–based social and emotional learning interventions: Current issues and future directions. *Psychology of Education Review*, *37*(2), 70–76.

Li, Q., & Baker, R. (2018, December). The different relationships between engagement and outcomes across participant subgroups in massive open online courses. *Computers & Education*, *127*, 41—65. Retrieved 2024–07–01, from `https://www.sciencedirect.com/science/article/pii/S0360131518302094`

Lin, X., & Gao, L. (2020). Students' sense of community and perspectives of taking synchronous and asynchronous online courses. *Asian Journal of Distance Education*, *15*(1), 169-179.

Lohaus, A., & Wild, E. (2021, January). Extracurriculare Förderangebote für benachteiligte Kinder und deren Eltern: Ein Angebot–Aneignungs–Modell zur Inanspruchnahme und Wirkung. *Zeitschrift für Pädagogische Psychologie*, *35*(1), 1—10.

Marinus, E., Powell, Z., Thornton, R., McArthur, G., & Crain, S. (2018, August). Unravelling the cognition of coding in 3–to–6–year olds: The development of an assessment tool and the relation between coding ability and cognitive compiling of syntax in natural language. In *Proceedings of the 2018 ACM Conference on International Computing Education Research* (p. 133—141). Association for Computing Machinery. Retrieved 2024–10–11, from `https://dl.acm.org/doi/10.1145/3230977.3230984`

Marsh, H., Lüdtke, O., Nagengast, B., Trautwein, U., Abduljabbar, A., Abdelfattah, F., & Jansen,

M. (2015, 2). Dimensional comparison theory: Paradoxical relations between self–beliefs and achievements in multiple domains. *Learning and Instruction*, *35*, 16–32.

Marsh, H., Pekrun, R., Parker, P., Murayama, K., Guo, J., Dicke, T., & Arens, A. (2019, 2). The murky distinction between self–concept and self–efficacy: Beware of lurking jingle–jangle fallacies. *Journal of Educational Psychology*, *111*(2), 331–353.

Marsh, H. W. (1987). The big–fish–little–pond effect on academic self–concept. *Journal of Educational Psychology*, *79*(3), 280—295.

Mason, J., Burton, L., & Stacey, K. (1982). *Thinking mathematically.* Addison–Wesley.

Mayer, R. (2009). Principles of multimedia learning based on social cues: Personalization, voice, and image principles. In R. Mayer (Ed.), *The cambridge handbook of multimedia learning* (p. 201–212). Cambridge University Press.

Mayring, P. (2022). *Qualitative Inhaltsanalyse Grundlagen und Techniken* (13., überarbeitete Auflage ed.). Beltz. Retrieved from `https://rds-tue.ibs-bw.de/link?kid=1800435738`

McBee, M. T., Peters, S. J., & Waterman, C. (2014, January). Combining scores in multiple–criteria assessment systems: The impact of combination rule. *Gifted Child Quarterly*, *58*(1), 69—89. Retrieved 2025–09–19, from `https://doi.org/10.1177/0016986213513794`

McBride, C., Ho, J. C., McQuade, M., Ngan, V. S. H., Ng, M. C. Y., Cheah, Z. R. E., & Maurer, U. (2025). Online assessment in young children: Challenges and considerations. *PsyCh Journal*, *14*(1), 5—14. Retrieved 2025–06–25, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/pchj.805`

McDonald, R. P. (1999). *Test theory: A unified treatment.* Psychology Press.

McHugh, C., & Way, J. (2018, January). What is reasoning? *Mind*, *127*(505), 167—196. Retrieved 2025–06–26, from `https://doi.org/10.1093/mind/fzw068`

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412—433.

Mcneish, D., & Wolf, M. G. (2023, September 12). Direct discrepancy dynamic fit index cutoffs for arbitrary covariance structure models. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 835-862.

Mejía–Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017, May). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, *19*(2), 130—146.

Meyer, H., & Junghans, C. (2021). *Unterrichtsmethoden* (2nd ed., Vol. 17) [gedruckt]. Cornelsen.

Miele, D., & Scholer, A. (2018). The role of metamotivational monitoring in motivation regulation. *Educational Psychologist*, *53*(1), 1–21.

Ministerium für Kultus, J. u. S. i. B. (2016a). Mathematik. *Bildungsplan der Grundschule*. Retrieved from http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GS_M.pdf

Ministerium für Kultus, J. u. S. i. B. (2016b). Mathematik. *Bildungsplan des Gymnasiums*. Retrieved from http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GYM_M.pdf

Moeller, J., Viljaranta, J., Tolvanen, A., Kracke, B., & Dietrich, J. (2022, October). Introducing the dynamics framework of moment–to–moment development in achievement motivation. *Learning and Instruction*, *81*, 101653. Retrieved 2025–08–22, from https://www.sciencedirect.com/science/article/pii/S0959475222000743

Monks, K., & Carter, N. (2014). *Lurch.* Retrieved 2025–09–08, from https://sourceforge.net/projects/lurch/

Moodle. (2023). *Glossary.*

Moon, B. (1986). *The "new maths" curriculum controversy: An international story.* Falmer Press.

Morales, S., & Torres, R. C. (2020, March). *Curse reverse from math snacks: Unlocking doors with expressions and equations.* MidSchoolMath.

Mory, E. H. (2004). Feedback research revisited. In (p. 745—783). Lawrence Erlbaum Associates Publishers.

Moya Pérez, J., Gutierrez, A., & Jaime, A. (2015, 02). Discriminating proof abilities of secondary school students with different mathematical talent. In (p. 171–177).

Muljana, P., & Luo, T. (2019). Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review. *Journal of Information Technology Education: Research*, *18*, 19—57.

Murayama, K., Goetz, T., Malmberg, L., Pekrun, R., Tanaka, A., & Martin, A. (2017). Within–person analysis in educational psychology: Importance and illustrations. In *Bjep monograph series ii: Part 12 the role of competence and beliefs in teaching and learning* (Vol. 2). British Psychological Society.

Musu–Gillette, L., Wigfield, A., Harring, J., & Eccles, J. (2015, 5 19). Trajectories of change in students' self–concepts of ability and values in math and college major choice. *Educational Research and Evaluation*, *21*(4), 343–370.

Na, G. (2011). Analysing the processes of discovery and proof of the mathematically gifted students. *Journal of Educational Research in Mathematics*, *21*, 105–120.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* National Council of Teachers of Mathematics. Retrieved 2025–08–06, from

http://archive.org/details/principlesstanda00nati

National Research Council. (2004). *Implementing randomized field trials in education: Report of a workshop* (L. Towne & M. Hilton, Eds.). National Academies Press.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012, aug). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research*, *39*(4), 374—396.

Niederer, K. (2017). The big–fish–little–pond effect: Self–concepts of gifted students in a part–time gifted programme. In N. Ballam & R. Moltzen (Eds.), (p. 155—178). Springer. Retrieved 2025–09–03, from https://doi.org/10.1007/978-981-10-6701-3_8

Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A., & Carraher, J. (2007). The contribution of logical reasoning to the learning of mathematics in primary school. *British Journal of Developmental Psychology*, *25*(1), 147—166. Retrieved 2024–10–15, from https://onlinelibrary.wiley.com/doi/abs/10.1348/026151006X153127

Nurlaelah, E., Pebrianti, A., Taqiyuddin, M., Dahlan, J., & Usdiyana, D. (2024, 10 6). Improving mathematical proof based on computational thinking components for prospective teachers in abstract algebra courses. *Infinity Journal*, *14*(1), 85–108.

OECD. (2019). *Pisa 2018 assessment and analytical framework*. OECD Publishing.

OECD. (2025). *Mathematics literacy*. Retrieved 2025–08–13, from https://www.oecd.org/en/topics/mathematics-literacy.html

Parrisius, C., Gaspard, H., Zitzmann, S., Trautwein, U., & Nagengast, B. (2022, 5). The "situative nature" of competence and value beliefs and the predictive power of autonomy support: A multilevel investigation of repeated observations. *Journal of Educational Psychology*, *114*(4), 791–814.

Peters, S., Matthews, M., Mcbee, M., & Mccoach, D. B. (2021). *Beyond gifted education: Designing and implementing advanced academic programs*.

Piaget, J., & Inhelder, B. (1977). *Von der Logik des Kindes zur Logik des Heranwachsenden: Essay über die Ausformung der formalen operativen Strukturen*. Klett–Cotta.

Polya, G. (2014). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Preckel, F., Golle, J., Grabner, R., Jarvin, L., Kozbelt, A., Müllensiefen, D., ... Worrell, F. C. (2020, March). Talent development in achievement domains: A psychological framework for within– and cross–domain research. *Perspectives on Psychological Science*, *15*(3), 691—722.

Pyryt, M. C. (2009). Recent developments in technology: Implications for gifted education. In L. V. Shavinina (Ed.), *International handbook on giftedness* (p. 1173—1180). Springer

Netherlands.

Pólya, G. (2010). *Schule des denkens: vom Lösen mathematischer Probleme* (Sonderausg. der 4. Aufl ed.). Francke.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.R-project.org/

Ramm, G., et al. (Eds.). (2006). *PISA 2003 – dokumentation der Erhebungsinstrumente.* [gedruckt]. Waxmann.

Ratz, S. (2016). Vocabulary learning with the moodle glossary tool: A case study. *Journal of Perspectives in Applied Academic Practice*, *4*(1), 44-51. Retrieved 2025–05–21, from https://jpaap.ac.uk/JPAAP/article/view/170

Rav, Y. (1999, February). Why do we prove theorems? *Philosophia Mathematica*, *7*(1), 5—41.

Rebholz, F. (2017). *Fostering mathematical competences by preparing for a mathematical competition* (Doctoral dissertation, University of Tuebingen).

Rebholz, F., Golle, J., Oschatz, K., & Trautwein, U. (2019). Fit für die Mathematik–Olympiade: Ein Trainingsprogramm zur Förderung mathematischer Fähigkeitenbesonders begabter Kinder im Grundschulalter. *Reihe Hector Core Courses*.

Rebholz, F., Golle, J., Tibus, M., Ruth–Herbein, E., Moeller, K., & Trautwein, U. (2017). Getting fit for the mathematical olympiad: Positive effects on achievement and motivation? *Zeitschrift für Erziehungswissenschaft*, *25*(5), 1175–1198.

Reis, S. M., & Park, S. (2001, October). Gender differences in high–achieving students in math and science. *Journal for the Education of the Gifted*, *25*(1), 52—73. Retrieved 2025–07–16, from https://doi.org/10.1177/016235320102500104

Renzulli, J. (2016, May). The three–ring conception of giftedness. In S. M. Reis (Ed.), *Reflections on gifted education* (p. 55-86). Prufrock Press.

Risemberg, R., & Zimmerman, B. J. (1992, November). Self-regulated learning in gifted students. *Roeper Review*, *15*(2), 98—101. Retrieved 2025–05–28, from https://doi.org/10.1080/02783199209553476

Roick, T., Gölitz, D., & Hasselhorn, M. (2018). DEMAT 3: deutscher Mathematiktest für dritte Klassen: Manual. Hogrefe.

Román–González, M., Pérez–González, J., Moreno–León, J., & Robles, G. (2018, March). Extending the nomological network of computational thinking with non–cognitive factors. *Computers in Human Behavior*, *80*, 441—459. Retrieved 2025–08–06, from https://www.sciencedirect.com/science/article/pii/S0747563217305563

Ross, K. A. (1998, March). Doing and proving: The place of algorithms and proofs in school

mathematics. *The American Mathematical Monthly*, *105*(3), 252—255. Retrieved 2025–08–12, from https://doi.org/10.1080/00029890.1998.12004875

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1—36.

Rota, G. (1997, May). The phenomenology of mathematical proof. *Synthese*, *111*(2), 183—196. Retrieved 2025–07–21, from https://www.jstor.org/stable/20117627

Rotelli, D., Noël, Y., Lallé, S., Luengo, V., & Pesce, D. (2023, August). A moodle plugin for rich xapi data logging. In O. Viberg, I. Jivet, P. Muñoz–Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and Sustainable Educational Futures* (p. 748—754). Springer Nature Switzerland.

Rothenbusch, S., Zettler, I., Voss, T., Lösch, T., & Trautwein, U. (2016, August). Exploring reference group effects on teachers' nominations of gifted students. *Journal of Educational Psychology*, *108*(6), 883—897.

Rotigel, J., & Fello, S. (2016). Mathematically gifted students: How can we meet their needs? *Gifted Child Today*, *27*(4), 46–51.

Schiefer, J., Stark, L., Gaspard, H., Wille, E., Trautwein, U., & Golle, J. (2021). Scaling up an extracurricular science intervention for elementary school students: It works, and girls benefit more from it than boys. *Journal of Educational Psychology*, *113*(4), 784—807.

Schneider, J., Börner, D., van Rosmalen, P., & Specht, M. (2016, October). Can you help me with my pitch? studying a tool for real–time automated feedback. *IEEE Transactions on Learning Technologies*, *9*(4), 318—327. Retrieved 2024–10–15, from https://ieeexplore.ieee.org/document/7745883

Schneider, W., & McGrew, K. (2012, January). The cattell–horn–carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (p. 99—144). Guilford Press.

Schoenfeld, A. H. (1985). Heuristics. In A. H. Schoenfeld (Ed.), *Mathematical problem solving* (p. 69—96). Elsevier.

Schoenfeld, A. H. (2009). The soul of mathematics. In D. A. Stylianou, M. L. Blanton, & E. J. Knuth (Eds.), *Teaching and learning proof across the grades: A K–16 perspective* (p. xii–xvi). Routledge.

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016, November). Do the smart get smarter? development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84—95. Retrieved 2024–10–23, from https://www.sciencedirect.com/science/article/pii/S0160289616301052

Schukajlow, S., Rakoczy, K., & Pekrun, R. (2023, 1 18). Emotions and motivation in mathematics education: Where we are today and where we need to go. *ZDM – Mathematics Education*, *55*(2), 249–267.

Schunk, D., & Mullen, C. (2012). Self–efficacy as an engaged learner. In . C. W. S. L. Christenson A. L. Reschly (Ed.), *Handbook of research on student engagement* (p. 219–235). Springer US.

Selden, A. (2013, January). Proof and problem solving at university level. *The Mathematics Enthusiast*, *10*, 303-334.

Senk, S. L. (1989, May). Van hiele levels and achievement in writing geometry proofs. *Journal for Research in Mathematics Education*, *20*(3), 309.

Shute, V. J., Sun, C., & Asbell–Clarke, J. (2017, November). Demystifying computational thinking. *Educational Research Review*, *22*, 142—158. Retrieved 2025–03–31, from `https://www.sciencedirect.com/science/article/pii/S1747938X17300350`

Sievertsen, H. H., Gino, F., & Piovesan, M. (2016, February). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences*, *113*(10), 2621—2624.

Singer, F. M., Sheffield, L. J., Freiman, V., & Brandl, M. (2016). *Research on and activities for mathematically gifted students.* Springer International Publishing. Retrieved 2025-09-25, from `http://link.springer.com/10.1007/978-3-319-39450-3`

Skulmowski, A., & Xu, K. M. (2022, March). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, *34*(1), 171—196. Retrieved 2025–05–19, from `https://doi.org/10.1007/s10648-021-09624-7`

Snapper, E. (1979). The three crises in mathematics: Logicism, intuitionism and formalism. *Mathematics Magazine*, *52*(4), 207—216. Retrieved 2025–08–13, from `https://www.jstor.org/stable/2689412`

Sommerhoff, D., & Ufer, S. (2019, 3 18). Acceptance criteria for validating mathematical proofs used by school students, university students, and mathematicians in the context of teaching. *ZDM – Mathematics Education*, *51*(5), 717–730. Retrieved 2025–08–18, from `https://doi.org/10.1007/s11858-019-01039-7`

Stalder, U. (2013). *Leselust in Risikogruppen: Gruppenspezifische Wirkungszusammenhänge.* Springer.

Stark, L. (2025). *Measuring and promoting primary school children's statistical literacy* (Doctoral dissertation, Universität Tübingen).

Staus, N., O'connell, K., & Storksdieck, M. (2021, 7 12). Addressing the ceiling effect when assessing stem out–of–school time experiences. *Frontiers in Education*, *6*, 690431.

Stein, X., Tsarava, K., Fabian, A., Carl, M., Kutkina, A., & Paravicini, W. (2025). Development and pilot study of an asynchronous online enrichment course on mathematical proving. *International Journal of Mathematical Education in Science and Technology*.

Stein, X., Tsarava, K., & Goecke, B. (2025). Development and validation of a preformal test for mathematicalproof competency. *Manuscript submitted for publication.*.

Stenning, K., & van Lambalgen, M. (2010, December). Reasoning, logic, and psychology. *WIREs Cognitive Science*, *2*(5), 555—567.

Stoll, R. R. (2012). *Set theory and logic*. Courier Corporation.

Street, K., Malmberg, L., & Stylianides, G. (2022, 7 14). Changes in students' self–efficacy when learning a new topic in mathematics: a micro–longitudinal study. *Educational Studies in Mathematics*, *111*(3), 515–541.

Stylianides, A. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, *38*(3), 289-321.

Stylianides, A. J. (2016). *Proving in the elementary mathematics classroom*. Oxford University Press. Retrieved from `https://doi.org/10.1093/acprof:oso/9780198723066.001.0001`

Stylianides, G. (2008). An analytic framework of reasoning–and–proving. *For the Learning of Mathematics*, *28*, 9—16.

Stylianides, G., & Silver, E. (2007). Reasoning and proving in school mathematics. In *Teaching and learning proof across the grades* (Vol. 38, p. 235–249). Routledge.

Sua Flórez, C., Gutierrez, A., & Jaime, A. (2020, 09). Design criteria of proof problems for mathematically gifted students..

Subotnik, R., Olszewski–Kubilius, P., & Worrell, F. (2011, 1). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest*, *12*(1), 3–54. Retrieved 2024–07–03, from `https://doi.org/10.1177/1529100611418056`

Taber, K. S. (2017, June). The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273—1296.

Tall, D. (1999). The cognitive development of proof: Is mathematical proof for all or for some? In Z. Usiskin (Ed.), *Developments in school mathematics around the world* (Vol. 4).

Tall, D. (2009, January). *The development of mathematical thinking: Problem–solving and proof.* Retrieved 2025–09–24, from `https://www.researchgate.net/publication/228731933_THE_DEVELOPMENT_OF_MATHEMATICAL_THINKING_PROBLEM-SOLVING_AND_PROOF`

Tall, D., Yevdokimov, O., Koichu, B., Whiteley, W., Kondratieva, M., & Cheng, Y. (2011, January). Proof and proving in mathematics education. In G. Hanna & M. de Villiers (Eds.), (Vol. 15, p. 13–49). Springer.

Tannenbaum, A. J. (1983). *Gifted children: Psychological and educational perspectives.* Macmillan.

Tao, T. (2018). *Qed [interactive textbook].* Retrieved 2025–09–08, from `https://github.com/teorth/QED`

Taub, M., Mudrick, N., & Azevedo, R. (2018, January). Strategies for designing advanced learning technologies to foster self–regulated learning. In R. Z. Zheng (Ed.), *Strategies for Deep Learning with Digital Technology: Theories and Practices in Education* (p. 137—169). Nova.

Terman, L. M. (1922, July). A new approach to the study of genius. *Psychological Review*, *29*(4), 310—318.

Thompson, D. R., Senk, S. L., & Johnson, G. J. (2012). Opportunities to learn reasoning and proof in high school mathematics textbooks. *Journal for Research in Mathematics Education*, *43*, 253–295.

Tinto, V. (1975, 3). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125.

Trautwein, U., Golle, J., Jaggy, A., Hasselhorn, M., & Nagengast, B. (2023). Mutual benefits for research and practice: Randomized controlled trials in the hector children's academy program. *Annals of the New York Academy of Sciences*, *1530*(1), 96—104. Retrieved 2024–09–05, from `https://onlinelibrary.wiley.com/doi/abs/10.1111/nyas.15074`

Tsai, Y., Kunter, M., Lüdtke, O., Trautwein, U., & Ryan, R. (2008, 5). What makes lessons interesting? the role of situational and individual factors in three school subjects. *Journal of Educational Psychology*, *100*(2), 460–472.

Tsarava, K. (2024, Feb). *Abbreviated ctt (german).* OSF. Retrieved from `osf.io/bv4kf`

Tsarava, K., Leifheit, L., Ninaus, M., Román–González, M., Butz, M. V., Golle, J., ... Moeller, K. (2019, October). Cognitive correlates of computational thinking: Evaluation of a blended unplugged/plugged–in course. In Q. Cutts & T. Brinda (Eds.), *Proceedings of the 14th workshop in primary and secondary computing education* (Vol. 10, p. 1—9). ACM.

Tsarava, K., Moeller, K., Román–González, M., Golle, J., Leifheit, L., Butz, M. V., & Ninaus, M. (2022, April). A cognitive definition of computational thinking in primary education. *Computers & Education*, *179*, 104425. Retrieved 2024–10–11, from `https://www.sciencedirect.com/science/article/pii/S036013152100302X`

Tullis, J. G., & Benjamin, A. S. (2011, February). On the effectiveness of self–paced learning. *Journal of Memory and Language*, *64*(2), 109—118. Retrieved 2024–10–15, from `https://`

www.sciencedirect.com/science/article/pii/S0749596X10000999

Urdan, T., & Midgley, C. (2001, 6). Academic self–handicapping: What we know, what more there is to learn. *Educational Psychology Review*, *13*(2), 115–138.

van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015, December). Effects of feedback in a computer–based learning environment on students' learning outcomes: A meta–analysis. *Review of Educational Research*, *85*(4), 475—511. Retrieved 2024–10–15, from https://doi.org/10.3102/0034654314564881

Villiers, M. (2012). *Rethinking proof with geometer's sketchpad* (Dan Bennett & Daniel Scher, Eds.). Key Curriculum Press.

Waluyo, M., Vidákovich, T., Ishartono, N., & Toyib, M. (2019, January). A review of assessing mathematical proving ability. In N. Ishartono, M. S. Adhantoro, Y. Sidiq, & Y. Sulistyono (Eds.), *Proceedings of the 4th progressive and fun education international conference.*

Wang, W., Guo, L., He, L., & Wu, Y. (2019). Effects of social–interactive engagement on the dropout ratio in online learning: insights from mooc. *Behaviour & Information Technology*, *38*(6), 621–636.

Wang, X., & Wei, Y. (2024, 12 17). The influence of parental involvement on students' math performance: a meta–analysis. *Frontiers in Psychology*, *15*, 1463359.

Watrin, L., Schroeders, U., & Wilhelm, O. (2023, February). Gc at its boundaries: A cross–national investigation of declarative knowledge. *Learning and Individual Differences*, *102*, 102267. Retrieved 2025–07–08, from https://www.sciencedirect.com/science/article/pii/S1041608023000122

Weaver, E. M., Shaul, K. A., & Lower, B. H. (2022). Implementation of an online poster symposium for a large–enrollment, natural science, general education, asynchronous course. *Frontiers in Education*, *7*(906995). Retrieved from https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2022.906995

Weber, K. (2005, January). Problem–solving, proving, and learning: The relationship between problem–solving processes and learning opportunities in the activity of proof construction. *The Journal of Mathematical Behavior*, *24*(3), 351—360. Retrieved 2024–11–14, from https://www.sciencedirect.com/science/article/pii/S0732312305000337

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wilder, R. L. (1944). The nature of mathematical proof. *The American Mathematical Monthly*, *51*(6), 309—323. Retrieved 2025–07–21, from https://www.jstor.org/stable/2304607

Wilhelm, O. (2005, January). Measuring reasoning ability. In *Handbook of understanding and*

measuring intelligence (p. 373—392).

Wilkins, J. L. M. (2004, July). Mathematics and science self–concept: An international investigation. *The Journal of Experimental Education*, *72*(4), 331—346.

William Revelle. (2023). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=psych` (R package version 2.3.6)

Williams–Johnson, M., & Gonzalez–Dehass, A. (2022, 10 2). Parental role construction leading to parental involvement in culturally distinct communities. *Educational Psychologist*, *57*(4), 231–237.

Winter, H. (1995). Mathematikunterricht und allgemeinbildung. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*(61), 37—46.

Witte, K., Spinath, B., & Ziegler, M. (2024, 5). Dissecting achievement motivation: Exploring the link between states, situation perception, and trait–state dynamics. *Learning and Individual Differences*, *112*, 102439.

Wittmann, E. C., & Müller, G. (1988). Wann ist ein Beweis ein Beweis? In P. Bender (Ed.), *Mathematikdidaktik: Theorie und praxis.festschrift für heinrich winter.* (p. 237—257). Cornelsen.

Wolf, M. G., & McNeish, D. (2023, January). dynamic: An r package for deriving dynamic fit index cutoffs for factor analysis. *Multivariate Behavioral Research*, *58*(1), 189—194.

Wolff, S., Hilpert, J., Vongkulluksn, V., Bernacki, M., & Greene, J. (2024, 12). Self–efficacy inertia: The role of competency beliefs and academic burden in achievement. *Contemporary Educational Psychology*, *79*, 102315.

Worrell, F. C., Subotnik, R. F., Olszewski–Kubilius, P., & Dixson, D. D. (2019, January). Gifted students. *Annual Review of Psychology*, *70*, 551—576. Retrieved 2024–07–03, from `https://www.annualreviews.org/content/journals/10.1146/annurev-psych-010418-102846`

Xavier, M., & Meneses, J. (2020). *Dropout in online higher education: A scoping review from 2014 to 2018.* eLearn Center, Universitat Oberta de Catalunya.

Yang, K., & Lin, F. (2008, January). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, *67*(1), 59—76. Retrieved 2024–11–12, from `https://doi.org/10.1007/s10649-007-9080-6`

Zaslavsky, O., Nickerson, S., Stylianides, A., Kidron, I., & Winicki–Landman, G. (2011). The need for proof and proving: Mathematical and pedagogical perspectives. *Proof and Proving in Mathematics Education*, *15*, 215–229.

Zeidner, M., & Schleyer, E. J. (1999, October). The big–fish–little–pond effect for academic

self–concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology*, *24*(4), 305—329.

Zhang, Z., Reinikainen, J., Adeleke, K., Pieterse, M., & Groothuis–Oudshoorn, C. (2018, 4). Time–varying covariates and coefficients in cox regression models. *Annals of Translational Medicine*, *6*(7), 121.

Ziegler, A. (2008). *Hochbegabung*. Reinhardt.

Zimmermann, B. (1991). Offene Probleme für den Mathematikunterricht und ein Ausblick auf Forschungsfragen. [gedruckt]. *Zentralblatt für Didaktik der Mathematik*, *23*(2), 38—46.

Zimny, L., Schroeders, U., & Wilhelm, O. (2024, January). Ant colony optimization for parallel test assembly. *Behavior Research Methods*, *56*(6), 5834—5848.

Øystein, H. P. (2011). What characterises high achieving students' mathematical reasoning? In B. Sriraman & K. H. Lee (Eds.), *The elements of creativity and giftedness in mathematics*. SensePublishers. Retrieved 2025–08–11, from `https://doi.org/10.1007/978-94-6091-439-3_13`