MEASURING AND PROMOTING PRIMARY SCHOOL CHILDREN'S STATISTICAL LITERACY

Dissertation
zur Erlangung des Doktorgrades
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von
M.Sc. Lucas Stark
aus Esslingen am Neckar

Tübingen
2025

1. Betreuer	Prof. Dr. Ulrich Trautwein
2. Betreuer	Prof. Dr. Benjamin Nagengast
3. Betreuerin:	Prof. Dr. Jessika Golle
Tag der mündlichen Prüfung:	24.04.2025
Dekanin:	Prof. Dr. Taiga Brahm
Dekan:	Prof. Dr. Dominik Papies
1. Gutachter:	Prof. Dr. Ulrich Trautwein
2. Gutachter:	Prof. Dr. Elisabeth Kraus

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my supervisors, Prof. Dr. Ulrich Trautwein, Prof. Dr. Benjamin Nagengast, and Prof. Dr. Jessika Golle, for their unwavering guidance and support throughout this process. Your intellectual insights, thoughtful feedback, and encouragement have been invaluable. Working under your supervision has been an inspiring experience, and I am immensely grateful for the opportunities I was given to contribute to such an innovative and interdisciplinary research environment.

I am also profoundly grateful to my colleagues at the Hector Research Institute of Education Sciences and Psychology and the LEAD Graduate School. In particular, I want to extend my appreciation to Dr. Ann-Kathrin Jaggy, Dr. Maja Flaig, Fabienne Kremer, Amelie Rebmann, Markus Kleinhansl, Maria Vilella, Lisa Martin, Xenia Stein, and Julia-Kim Walther. Your encouragement, critical discussions, and camaraderie made this journey not only productive but also deeply enjoyable. I would also like to express my gratitude to my student assistants, who played a crucial role in the realization of this dissertation. Their hard work, dedication, and attention to detail were invaluable in ensuring the success of the research studies.

Special thanks go to Dr. Jens Krummenauer and Prof. Dr. Sebastian Kuntze from the Ludwigsburg University of Education. Your expertise and dedication were instrumental in the development and refinement of the statistical literacy intervention, as well as in realizing and evaluating the associated studies. Your mentorship and collaboration have left a lasting impression on me.

Furthermore, I am sincerely grateful to all managing directors, course instructors, children, and parents who participated in or supported the studies as part of the Hector Children's Academy Program. Your willingness to contribute and engage made this research possible. I also want to acknowledge the Hector Foundation II for their generous funding, which provided the essential resources to bring this project to life.

Lastly, I want to thank my family, friends, and partner. Your unwavering support and belief in me, even during the most challenging moments, were the bedrock of my perseverance. Thank you for your patience, understanding, and encouragement, which carried me through this journey.

ABSTRACT

In this time of increasing digitization, fake news, and artificial intelligence, knowing how to deal with data has become increasingly essential. The ability to understand and evaluate statistical data is known as *statistical literacy* (Ben-Zvi & Garfield, 2004; Gal, 2002; Rumsey, 2002; Schield, 1999; Wallman, 1993). Statistical literacy empowers citizens to interpret and evaluate statistical data themselves, without relying on other people's interpretations and evaluations (Bodemer, 2014). Moreover, statistical literacy helps combat misinformation by empowering people to question data sources and identify biases, thus enabling citizens to make meaningful choices and helping democracies function (e.g., United Nations 2012; Wallman, 1993). Despite the importance of statistical literacy in today's societies, most adults lack the necessary skills to critically evaluate and interpret statistical data (Galesic & Garcia-Retamero, 2010; Gigerenzer et al., 2007). As an understanding of statistical concepts is already beginning to develop during the primary school years (English & Watson, 2013; Piaget & Inhelder, 1975; Watson & Moritz, 2000), promoting statistical literacy from early on is crucial for building a statistically literate society. However, research on the promotion of statistical literacy is scarce. Only a small number of intervention studies exist, and most of them have faced methodological challenges such as small sample sizes or the absence of a control group (e.g., Bakker, 2004; Ben-Zvi, 2006; Paparistodemou & Meletiou-Mavrotheris, 2008).

The present dissertation comprises three empirical studies, which addressed the questions of how to measure and promote statistical literacy in primary school children. Specifically, the dissertation focused on (a) the evaluation of the efficacy and effectiveness of an extracurricular statistical literacy intervention for third and fourth graders and (b) the development of a new measurement instrument for assessing children's individual decision thresholds.

Study 1 focused on evaluating the efficacy of an 8-week statistical literacy intervention under standardized conditions. In this study, university staff conducted the intervention at five different locations of an extracurricular enrichment program for gifted primary school children (the *Hector Children's Academy Program, HCAP*). The study included 53 children who took part in a randomized controlled field trial. The intervention focused on promoting aspects of children's statistical literacy. The intervention demonstrated positive effects on children's data-based argumentation, understanding of the concept of variability, and self-concept in data-related tasks.

Study 2 focused on the development and evaluation of a new measurement instrument for assessing children's individual decision thresholds. A decision threshold describes the level of certainty a person needs to be confident enough to make a decision in a probabilistic reasoning task. With the use of a *signal detection theory* (SDT; Green & Swets, 1966; Stanislaw & Todorov, 1999) approach to data from 299 first to fourth graders, the accuracy of the instrument was found to be high. As expected, the decision threshold was found to covary significantly with several related constructs.

Based on the positive results of the efficacy study, *Study 3* investigated the effectiveness of the statistical literacy intervention under less standardized conditions. In this study, course instructors from the field were trained to conduct the intervention at nine local HCAP sites. The study included 87 third and fourth graders who participated in the randomized controlled field trial. The intervention had positive effects on the children's data-based argumentation, some aspects of their understanding of the concept of variability, as well as their attainment value and self-concept in data-related tasks.

In summary, this dissertation provides evidence of the effectiveness of an extracurricular statistical literacy intervention for third and fourth graders and the accuracy and validity of a newly developed measurement instrument for assessing children's decision thresholds. The findings of all three studies are summarized and discussed in the broader context of statistical literacy research and educational policy and practice.

ZUSAMMENFASSUNG

In Zeiten zunehmender Digitalisierung, Fake News und künstlicher Intelligenz ist der Umgang mit Daten immer wichtiger geworden. Die Fähigkeit, statistische Daten zu verstehen und auszuwerten, wird als *Statistical Literacy* bezeichnet, was auch als Datenkompetenz übersetzt werden kann (Ben-Zvi & Garfield, 2004; Gal, 2002; Rumsey, 2002; Schield, 1999; Wallman, 1993). Statistical Literacy befähigt die Bürger, statistische Daten selbst zu interpretieren und zu bewerten, ohne sich auf die Interpretationen und Bewertungen anderer zu verlassen (Bodemer, 2014). Darüber hinaus trägt Statistical Literacy dazu bei, Fehlinformationen zu bekämpfen, indem sie die Menschen befähigt, Datenquellen zu hinterfragen und menschliche Verhaltenstendenzen zu erkennen, wodurch die Bürgerinnen und Bürger in die Lage versetzt werden, sinnvolle Entscheidungen zu treffen und das Funktionieren von Demokratien unterstützt wird (z. B. United Nations 2012; Wallman, 1993). Trotz der Bedeutung von Statistical Literacy in der heutigen Gesellschaft verfügen die meisten Erwachsenen nicht über die notwendigen Fähigkeiten, um statistische Daten kritisch zu bewerten und zu interpretieren (Galesic & Garcia-Retamero, 2010; Gigerenzer et al., 2007). Da sich das Verständnis für statistische Konzepte bereits während der Grundschulzeit zu entwickeln beginnt (English & Watson, 2013; Piaget & Inhelder, 1975; Watson & Moritz, 2000), ist die frühzeitige Förderung von Statistical Literacy entscheidend für den Aufbau einer statistisch gebildeten Gesellschaft. Allerdings gibt es bisher wenig Forschung zur Förderung von Statistical Literacy. Es gibt nur eine kleine Anzahl von Interventionsstudien, und die meisten von ihnen weisen methodischen Probleme auf, wie z. B. eine geringe Stichprobengröße oder das Fehlen einer Kontrollgruppe (e.g., Bakker, 2004; Ben-Zvi, 2006; Paparistodemou & Meletiou-Mavrotheris, 2008).

Die vorliegende Dissertation umfasst drei empirische Studien, die sich mit der Frage befassen, wie Statistical Literacy von Grundschulkindern gemessen und gefördert werden kann. Konkret ging es in der Dissertation um (a) die Evaluierung der Wirksamkeit und Effektivität einer außerschulischen Intervention zur Förderung von Statistical Literacy bei Dritt- und Viertklässlern und (b) die Entwicklung eines neuen Messinstruments zur Beurteilung von individuellen Entscheidungsschwellen von Kindern.

Studie 1 konzentrierte sich auf die Bewertung der Wirksamkeit einer 8-wöchigen Intervention zur Förderung von Statistical Literacy unter standardisierten Bedingungen. In dieser Studie führten Universitätsmitarbeiter die Intervention an fünf verschiedenen Standorten eines außerschulischen Enrichment-Programms für besonders begabte und hochbegabte

Grundschulkinder durch (*Hector Kinderakademien, HKA*). Die Studie umfasste 53 Kinder, die an einem randomisierten kontrollierten Feldversuch teilnahmen. Die Intervention konzentrierte sich auf die Förderung von Aspekten von Statistical Literacy der Kinder. Die Intervention zeigte positive Auswirkungen auf das datenbasierte Argumentieren der Kinder, ihr Verständnis des Konzepts der Variabilität und ihr Selbstkonzept bei datenbezogenen Aufgaben.

Studie 2 konzentrierte sich auf die Entwicklung und Evaluierung eines neuen Messinstruments zur Beurteilung von individuellen Entscheidungsschwellen von Kindern. Eine Entscheidungsschwelle beschreibt das Maß an Gewissheit, das eine Person benötigt, um eine Entscheidung in einer wahrscheinlichkeitstheoretischen Aufgabe zu treffen. Mit Hilfe der *Signalentdeckungstheorie* (SET; Green & Swets, 1966; Stanislaw & Todorov, 1999) wurde eine hohe Genauigkeit des Instruments anhand der Daten von 299 Schülern der ersten bis vierten Klasse nachgewiesen. Wie angenommen, bestanden signifikante Zusammenhänge zwischen der Entscheidungsschwelle und mehreren verwandten Konstrukten.

Auf der Grundlage der positiven Ergebnisse der Wirksamkeitsstudie wurde in *Studie 3* die Wirksamkeit der Intervention zu Statistical Literacy unter weniger standardisierten Bedingungen untersucht. In dieser Studie wurden Kursleiter aus der Praxis geschult, um die Intervention an neun HKA-Standorten durchzuführen. Die Studie umfasste 87 Dritt- und Viertklässler, die an der randomisierten kontrollierten Feldstudie teilnahmen. Die Intervention hatte positive Auswirkungen auf das datenbasierte Argumentieren der Kinder und einige Aspekte ihres Verständnisses des Konzepts der Variabilität. Außerdem wurde die Wichtigkeit von datenbezogenen Aufgaben und ihr zugehöriges Selbstkonzept gefördert.

Zusammenfassend liefert diese Dissertation Belege für die Wirksamkeit einer außerschulischen Intervention zur Förderung von Statistical Literacy bei Dritt- und Viertklässlern sowie für die Genauigkeit und Validität eines neu entwickelten Messinstruments zur Beurteilung von Entscheidungsschwellen bei Kindern. Die Ergebnisse aller drei Studien werden zusammengefasst und im breiteren Kontext der Forschung zu Statistical Literacy sowie der Bildungspolitik und -praxis diskutiert.

CONTENTS

1

Introduction and Theoretical Framework

1 Introduction and Theoretical Framework

In people's everyday lives, it is becoming more and more important to understand statistical data. As data are increasingly available across sectors such as healthcare, education, government, and business, individuals equipped with statistical literacy are better positioned to make informed decisions. Statistical literacy is the ability to understand and evaluate statistical information (Ben-Zvi & Garfield, 2004; Gal, 2002; Rumsey, 2002; Schield, 1999; Wallman, 1993). It empowers citizens to interpret and evaluate statistical data themselves without relying on other people's interpretations and evaluations (Bodemer, 2014). Moreover, statistical literacy helps combat misinformation by empowering people to question data sources and identify biases, thus enabling citizens to make meaningful choices and helping democracies function (e.g., United Nations, 2012; Wallman, 1993).

Despite the importance of statistical literacy in today's societies, most adults lack the necessary skills to critically evaluate and interpret statistical data. For instance, a study by Galesic and Garcia-Retamero (2010) found that a significant proportion of adults from Germany and the United States struggle with basic statistical concepts such as percentages and probabilities, which are essential for understanding everyday information such as health statistics and election polls. The lack of statistical literacy is further highlighted by Gigerenzer et al. (2007), who demonstrated that even well-educated individuals, including doctors and politicians, often misinterpret health statistics, leading to poor decision-making. This widespread statistical illiteracy presents a challenge for societies when data-driven decisions are becoming more prevalent. Even students struggle to understand basic concepts such as variability (delMas & Liu, 2005), randomness (Batanero & Serrano, 1999), and sampling distributions (delMas et al., 1999). Efforts to improve statistical literacy are thus crucial for enhancing citizens' ability to make informed decisions.

To ensure that everyone can achieve statistical literacy, the question that arises is how to effectively promote it (e.g., Wallman, 1993). Previous research from tertiary education has indicated that cooperative learning methods can help enhance the understanding of statistics in introductory classes (Krause et al., 2009). Also, the predict-observe-explain approach has been shown to be useful for improving conceptual knowledge in the sciences (Gustina et al., 2023). However, there is still a lack of systematic research on how to promote aspects of statistical literacy—such as data-based argumentation or understanding central statistical concepts—especially in primary school children.

With this dissertation, I aimed to address this research gap by examining how aspects of statistical literacy can be fostered in primary school children. To answer this question, an extracurricular statistical literacy intervention was developed for third- and fourth-grade primary school children. An intervention must pass through six stages to gradually realize its real-world effectiveness (Herbein et al., 2018b; Humphrey et al., 2016; Lendrum & Wigelsworth, 2013): (1) detect the needs of the target group, (2) develop the intervention on the basis of scientific theories and evidence, (3) run a pilot study, (4) run an efficacy study, (5) run an effectiveness study, and (6) conduct a scaling-up study.

The statistical literacy intervention developed here followed the first five steps. Three main research questions were addressed: (1) whether the statistical literacy intervention's efficacy is adequately high, (2) whether a new measurement instrument that assesses the decision threshold as an indicator of the jumping to conclusions bias is valid, and (3) whether the effectiveness of the statistical literacy intervention is demonstrable by testing whether the effects of the treatment on children's statistical literacy from the efficacy study are still stable after implementation. These questions established the intervention's effectiveness under real-world conditions and served as the grounds for future efforts to design and evaluate statistical literacy interventions for primary school children.

This dissertation is structured to first identify the target group's needs and then evaluate the effectiveness of the statistical literacy intervention. The overall goal was to gather evidence on whether and how statistical literacy can be promoted in primary school children. The introductory chapter serves to establish the broader context of the intervention. It gives insights into why and how the statistical literacy intervention was designed. In Study 1, a randomized controlled field trial was conducted to test whether the intervention had a significant effect on aspects of children's statistical literacy under standardized conditions. Thus, the intervention was conducted by university staff. In Study 2, a new measurement instrument was developed to complement the existing measures and add more meaning to the observable effects of the intervention. The so-called decision threshold (e.g., Moritz et al., 2006; Moritz et al., 2020) was measured with objective probabilities to determine how the intervention affected the minimum probability one needed to feel certain enough to decide. These objective probabilities helped to measure whether the children in the intervention based their decisions on too little evidence or not. In Study 3, we examined the real-world effectiveness of the intervention. To do so, the intervention was conducted under less standardized conditions, meaning that it was conducted by course instructors from the field, who were trained in the content of the intervention. The dissertation closes with a discussion of all results and their implications.

1.1 Theoretical Conceptualizations of Statistical Literacy

1.1.1 Definitions of statistical literacy

As there is no consensus on what statistical literacy entails, I present several definitions of statistical literacy in this chapter and use their commonalities to derive a working definition of statistical literacy for this dissertation.

Wallman (1993) described statistical literacy as “the ability to understand and critically evaluate statistical results that permeate our daily lives—coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions” (Wallman, 1993, p. 1). This definition is one of the earliest and most prominent in the research field. It captures statistical literacy as an ability with cognitive and affective components and focuses on statistical information from daily life.

Schield (1999) described statistical literacy as "a basic skill: the ability to think critically about arguments using statistics as evidence" (Schield, 1999, p. 15). He focused on the cognitive aspects of statistical literacy and described statistical literacy as comprising mainly inductive and partially deductive reasoning skills. He dismissed affective components and stressed that the main use of statistical literacy is to use statistical evidence in arguments, without further specifying what kinds of problems are tackled.

Similarly, Rumsey (2002) defined statistical literacy as “understanding statistics well enough to be able to consume the information that they are inundated with on a daily basis, think critically about it, and make good decisions based on that information” (Rumsey, 2002, p. 1). This author also focused on the cognitive aspects of statistical literacy and on statistical information from daily life.

Gal (2002) presented statistical literacy as “the ability to interpret, critically evaluate, and communicate about statistical information and messages” (Gal, 2002, p. 1). He also summarized statistical literacy as mainly cognitive but added its utility: to communicate what is being processed.

Ben-Zvi and Garfield (2004) stated that “[s]tatistical literacy includes basic and important skills that may be used in understanding statistical information or research results. These skills include being able to organize data, construct and display tables, and work with different representations of data. Statistical literacy also includes an understanding of concepts, vocabulary, and symbols, and includes an understanding of probability as a measure of uncertainty” (Ben-Zvi & Garfield, 2004, p. 7). These authors focused on the cognitive aspects of statistical literacy. They also mentioned several skills that are included in it and are supposed

to be foundational. These foundational skills are presumed to be important for the average citizen.

In general, the question of how statistical literacy should be conceptualized is still being debated. However, when we look at the definitions above, the researchers appear to agree about several aspects (Ben-Zvi & Garfield, 2004; Gal, 2002; Rumsey, 2002; Schield, 1999; Wallman, 1993). They all define statistical literacy as *an ability*, which mostly includes being able to *read* or *understand* and *interpret* or *evaluate* statistical information. Statistical information is represented as numbers, diagrams, or text. Furthermore, most definitions describe statistical literacy as a *minimum skill level*, which all citizens are expected to have *to meet daily statistical challenges* (e.g., interpreting a graph in a newspaper). In addition, the direct benefit of statistical literacy is that statistical information can be communicated in *arguments*. Therefore, in this dissertation, statistical literacy is defined as the ability to understand and evaluate statistical information to meet daily statistical challenges and develop data-based arguments.

1.1.2 Models of statistical literacy

After defining statistical literacy, the question that arises is: What are the aspects that constitute statistical literacy? There is a wide range of conceptualizations of statistical literacy, and several models have been proposed to describe statistical literacy or the more general term *statistical thinking* (Gal, 2002; Jones et al., 2000; Moore, 1990; Wild & Pfannkuch, 1999). In this chapter, I present several models to answer the question: Which key aspects are essential for statistical literacy and should be promoted?

Moore's (1990) core elements of statistical thinking

Moore (1990) was one of the first to describe important elements of statistical thinking. He defined five core elements. First, one must acknowledge the omnipresence of variability in processes. This phenomenon describes the idea that measurements are generally not stable. Repeated measures most likely lead to different results. Second, one must acknowledge that there is a need for data about processes. In statistical thinking, data must be used to support assumptions. Third, one must keep in mind how the data were produced. Variability in the data can arise due to controlled and uncontrolled conditions, which means, for example, that differences in the data can be attributed to data collection methods or randomization. Fourth, one must be able to quantify variability. For example, it can be quantified in terms of probability or measures of spread. Fifth, one must be able to explain variability, which means that methods of statistical analysis must be used to find systematic effects in the data at hand.

Altogether, Moore (1990) wrote a compact model of statistical thinking with a focus on the concept of variability and an understanding of it in different parts of scientific research. He defined different types of thinking and focused on the importance of data as evidence, but he did not describe personal characteristics or thoughts as part of his model.

Wild and Pfannkuch's (1999) framework for statistical thinking in empirical inquiry

Wild and Pfannkuch (1999) explored the complexity of statistical thinking by interviewing students and professional statisticians to understand their approaches to real-world problem-solving using statistics. Their goal was to develop a framework for thinking patterns and strategies in statistical problem-solving, focusing in particular on how statistical knowledge can be applied to critically interpret information in reports and the media. They arrived at a four-dimensional framework that describes elements of statistical thinking during data-based inquiry.

The first dimension is the Investigative Cycle, which follows the Problem, Plan, Data, Analysis, Conclusions (PPDAC) model (MacKay & Oldford, 2000) to guide the process of statistical inquiry (Wild & Pfannkuch, 1999). It emphasizes understanding and solving statistical problems in real-world contexts, with each cycle aimed at improving system dynamics and informing future actions. This process often initiates additional investigative cycles as new knowledge is acquired.

The second dimension comprises the types of thinking involved in statistical inquiry, categorized into both general thinking and inherently statistical thinking (Wild & Pfannkuch, 1999). Fundamental to statistical thinking is the recognition of the need for data, the process of transnumeration (i.e., transforming data representations to enhance understanding), the use of distinctive statistical models, and especially the understanding of variability, which is postulated as the main concept and is omnipresent in statistics. These elements emphasize the synthesis of contextual knowledge with statistical knowledge, constantly moving between the two to interpret and explain data. Finally, strategic thinking is essential for planning and problem-solving within real-world constraints, and modeling plays a key role in predicting and understanding the behavior of systems.

The third dimension is the Interrogative Cycle, a recursive thinking process used throughout statistical problem-solving (Wild & Pfannkuch, 1999). This cycle involves generating possibilities, seeking information, interpreting the results, and criticizing and judging them on consistency and accuracy. The process is continuous, with steps such as generating hypotheses, recalling or collecting data, and comparing new insights with existing

knowledge, followed by critical reflection and judgment. At each stage, thinkers assess their assumptions, the reliability of data, and the practicality of their conclusions, leading to a distillation of ideas and information into actionable insights.

The fourth dimension is composed of dispositions, which encompass personal qualities that are crucial for initiating and sustaining effective thinking, particularly in statistical problem-solving (Wild & Pfannkuch, 1999). These dispositions include curiosity and awareness, which drive question generation and innovation; engagement, which heightens sensitivity and interest in relevant information; imagination, which is essential for creating and evaluating mental models; scepticism, which involves critically assessing the validity of information; and logical reasoning, which is necessary for constructing valid arguments and understanding implications. These dispositions are not merely innate but can be influenced by the level of engagement with a problem, suggesting that fostering these qualities through experience and education can enhance statistical thinking.

Altogether, Wild and Pfannkuch's (1999) framework is a rather complex model because it is based not only on the average citizen's needs but on professional statisticians' procedures. It has some theoretical implications for the needs of the average citizen. Going through the investigative cycle is thought to be of high importance, and a conclusion must be reached at the end. To reach a conclusion, the individual must be able to make an argument that is based on data. In their model, the authors also showed that various fundamental ways of thinking are relevant for statistical thinking, in particular variability, which is claimed to be the starting point of all statistical investigations. In addition, the model emphasizes the idea that dispositions such as engagement and perseverance are important for performing statistical thinking.

Jones et al.'s (2000) framework for characterizing children's statistical thinking

Jones et al. (2000) developed a framework that describes the cognitive development of primary school children's statistical thinking. They conceptualized four different behavioral constructs that develop across four levels of thinking. First, *describing data* is the explicit reading and extraction of information directly from data displays. It involves recognizing graphical conventions such as titles and axis labels. It also includes evaluating and comparing different visual representations of the same data but does not extend to interpreting or analyzing the data further. Second, *organizing and reducing data* involves mental actions such as grouping, ordering, and summarizing data, often through measures such as central tendency and spread. This process includes recognizing how data are represented in different ways and

understanding concepts such as the arithmetic mean and variability, while noting that some information might be lost during reorganization. Third, *representing data* involves constructing visual displays that organize a data set and adhere to basic graphical conventions. Representing data includes tasks such as completing partially constructed graphs and creating new visual representations of data, tasks that younger elementary school students often find challenging. Fourth, *analyzing and interpreting data* involves recognizing patterns, trends, and exceptions, as well as using the data to make inferences and predictions. This construct includes comparing and combining data and extrapolating or predicting future outcomes, additional tasks that primary school students often struggle with.

In comparison with Wild and Pfannkuch (1999), Jones et al. (2000) did not specify any dispositions as being important to statistical thinking. Instead, they concentrated on relevant processes. Describing, analyzing, and interpreting data were described as essential tasks that can be linked to producing data-based arguments. Furthermore, the importance of understanding the concept of variability was made clear, as it is necessary for organizing and reducing data.

Gal's (2002) model of statistical literacy

Gal (2002) proposed a rather simple model of statistical literacy. He saw two major groups of elements as essential to statistical literacy: knowledge and dispositional elements. He mentioned five knowledge elements (literacy skills, statistical knowledge, mathematical knowledge, context knowledge, and critical skills) and two dispositional elements (beliefs/attitudes and critical stance). He proposed that these components are essential for citizens' daily challenges as data consumers, not data producers. Contexts included watching the news, attending a political event, or reading a report at work.

Gal (2002) defined the seven elements of his model not as fixed and separate but as context-dependent. He proposed the following: First, general literacy skills are deeply intertwined with statistical literacy, as individuals must navigate both written and graphical information to understand statistical messages. Second, statistical knowledge is needed to understand statistical terms, concepts, and methods. For example, it is important to know why data are needed and how concepts such as probability and variability are related to the data. Third, mathematical knowledge is used to perform certain calculations such as probabilities and averages. Fourth, context knowledge is needed to make sense of the data at hand and to know how variations and errors can be interpreted. Fifth, critical skills are needed to explore why certain statistics, or an individual's interpretation of them, could be biased. Sixth and

seventh, beliefs and attitudes as well as critical stance are affective dispositions that are needed to activate the knowledge elements, for example, feeling that one is competent enough to handle statistical problems.

Gal's (2002) model of statistical literacy highlights the importance of both knowledge and dispositional elements by enabling individuals to critically interpret and engage with statistical data in everyday contexts. Basic concepts such as variability must be understood to be able to understand and develop data-based arguments as described in previous models (Jones et al., 2000; Wild & Pfannkuch, 1999). Also, the model emphasizes that statistical literacy is not just about acquiring knowledge but also involves a critical mindset, beliefs, and attitudes (e.g., motivation) to question and evaluate data as in Wild and Pfannkuch's (1999) model.

Intersections of the models of statistical literacy

From the analysis of several frameworks, three key aspects emerged as central to statistical literacy. First, data-based argumentation is fundamental across all models. Whether describing data, interpreting trends, or drawing conclusions, the ability to use data to support claims is an essential outcome of statistical thinking (Gal, 2002; Jones et al., 2000; Moore, 1990; Wild & Pfannkuch, 1999). This skill involves not only reading and understanding data but also using it as the basis for critical inquiry, communicating interpretations, and decision-making.

Second, understanding variability is emphasized in every model, particularly as a core concept in statistics (Moore, 1991; Wild & Pfannkuch, 1999). Variability provides the foundation for recognizing patterns, making comparisons, and addressing uncertainty. Whether in organizing data (Jones et al., 2000) or interpreting real-world contexts (Gal, 2002; Moore, 1990), acknowledging and understanding variability is key to becoming statistically literate.

Finally, the importance of motivational dispositions (e.g., curiosity, engagement, and perseverance) cannot be overlooked. Wild and Pfannkuch (1999) and Gal (2002) particularly highlighted how dispositions such as scepticism and perseverance drive individuals to question, explore, and critically analyze statistical information. These affective elements are essential for sustaining statistical inquiry and ensuring that statistical literacy extends beyond mere knowledge acquisition.

In summary, the ability to form data-based arguments, understand the concept of variability, and engage with statistical information are three key aspects of statistical literacy that need to be promoted in educational interventions.

1.1.3 Key aspects of statistical literacy

In summary, current models for describing statistical literacy focus on three main aspects: data-based argumentation, understanding the concept of variability, and motivation in data-related tasks. These aspects are described in greater detail below.

Data-based argumentation

The first key aspect of statistical literacy is data-based argumentation. It describes the process of using data to support or disconfirm statements about reality (e.g., Krummenauer & Kuntze, 2019b). Data-based argumentation can be seen as a more specific version of Toulmin's (2003) model of argumentation. He defined two components of an argument. The first component is the *conclusion*, which is a statement whose validity is in question. The second component is the *datum*, which is the basis of the argument that is supposed to justify the statement. In data-based argumentation specifically, data are used as the datum (Krummenauer et al., 2020).

Data-based argumentation can also be viewed from the perspective of scientific reasoning theories (e.g., Kuhn, 2011; Sodian et al., 1991; Zimmermann, 2007). In this perspective, the conclusion of the argument is called a hypothesis, and the datum is the empirical evidence. Kuhn (2011) described the process of matching a hypothesis with evidence as the *coordination of theory and evidence*. For instance, when confronted with an interpretation about the effectiveness of jogging on sleep quality, data about individuals' sleep and jogging behavior must be regarded.

For this process to succeed, people need to use various strategies (Zimmermann, 2007). Kuhn (2011) stated that they must first be able to distinguish theory from evidence. Research has demonstrated that by the age of 4, children begin to grasp the idea of false beliefs (Wellman et al., 2001). This phenomenon suggests that they are capable of recognizing that theories can be either true or false, and thus, they can learn to differentiate between theory and evidence. Kuhn (2011) suggested that the next developmental stage is recognizing evidence as a foundation for knowledge. In the context of forming arguments based on statistical data, younger children often believe that knowledge is absolute (Conley et al., 2004) and may feel certain about their conclusions even without evidence. However, as they grow, they come to understand that knowledge must be backed by empirical evidence. By the age of 6, children are typically more skilled at incorporating evidence into their explanations compared with 4-year-olds (Kuhn & Pearsall, 2000). Another strategy that can be used is to look for contradictory evidence instead of just searching for confirmatory evidence (Sodian et al., 1991).

There is evidence that young children are already able to develop data-based arguments. In two studies by Krummenauer and Kuntze (2018, 2019), around 33% to 40% of fourth graders were able to use data as evidence in arguments. Depending on the difficulty of the task context, even up to 90% were able to successfully formulate data-based arguments (Krummenauer & Kuntze, 2019; Krummenauer et al., 2022). Also, over half of middle school students were found to be capable of using data to support their arguments (Ruiz-Primo et al., 2010). Therefore, primary school children already possess the ability to use evidence in arguments, but there is still a lot of room for improving it.

Understanding the concept of variability

The second key aspect of statistical literacy is understanding the concept of variability, which many scholars regard as the most central concept in statistics (e.g., McKenzie, 2004; Watson & Callingham, 2003; Wild & Pfannkuch, 1999). Despite the importance of variability, school curricula often emphasize measures of central tendency far more than measures of variability (Torok & Watson, 2000). This imbalance is noteworthy because variability is not only omnipresent but also fundamental to the very existence of statistics (Moore, 1990; Watson, 2006). Variability arises from the inherent differences between individuals and errors in measurement, which together create uncertainty (Moore, 1990; Wild & Pfannkuch, 1999). Consequently, statistics is about quantifying variability, predicting it, and explaining where it comes from (Moore, 1990). These explanations can then be used in data-based arguments.

Young children usually tend to ignore the concept of variability. Krummenauer et al. (2022) found that primary school children have difficulties considering statistical variability when developing data-based arguments. Also, Conley et al. (2004) found that primary school children tend to see knowledge as rather certain and as a true reflection of reality. They may hold this view because, in school, children often learn that some answers are right and others are wrong—they learn to see the world as deterministic (Moore, 1990). However, such beliefs can change into more sophisticated beliefs about knowledge (Schiefer et al., 2021). Therefore, it is possible for young children to develop an understanding of variability.

Even adults seem to fail to acknowledge the concept of variability. This manifestation can be recognized in many different so-called *cognitive biases*, which are systematic errors in thinking that occur when people process and interpret information, leading to deviations from rational judgment (e.g., Tversky & Kahnemann, 1971). Cognitive biases can be understood as misunderstandings of statistical concepts (Ben-Zvi & Garfield, 2008; Tversky & Kahnemann, 1971). For instance, the *gambler's fallacy* (e.g., Barron & Leider, 2010) is a classic example

of misunderstanding statistical independence. Many people believe that after a series of similar outcomes (e.g., multiple coin tosses landing on heads), the next outcome is more likely to be different, when in reality, each toss is independent, and the probability remains constant. Thus, people with this bias expect variability to follow a certain pattern. Similarly, the *confirmation bias* (e.g., Nickerson, 1998) shows how people often seek out or interpret data that support their preconceived notions, neglecting the importance of considering all available evidence, akin to ignoring the concept of representative sampling. Therefore, people with this bias reduce variability in favor of their own pre-existing views. Furthermore, the *jumping to conclusions bias* (Garety et al., 2005) can be viewed as underestimating the influence of variability on small samples. People with this bias tend to believe that small samples are just as representative as larger ones and therefore make hasty decisions on the basis of too little evidence. This bias is also known as the belief in the *law of small numbers* (Tversky & Kahnemann, 1971).

Altogether, understanding statistical variability is a critical aspect of statistical literacy, yet primary school children often struggle with this concept, tending to see knowledge as certain and deterministic. However, these beliefs and related cognitive biases could be targeted by educational interventions. Research suggests that there is potential for young children to develop a more sophisticated understanding of variability.

Motivation in data-related tasks

The third key aspect of statistical literacy is motivation in data-related tasks. Many scholars see this aspect as crucial for statistical literacy (e.g., Gal, 2002; Wallman, 1993; Watson, 2006; Wild & Pfannkuch, 1999). To be able to make use of one's cognitive abilities or knowledge in data-related tasks (i.e., develop data-based arguments), one must be motivated to do so. Motivation has been found to be connected to achievement in various subjects (e.g., Nasser, 2004). More specifically, statistics motivation has been positively linked to statistics engagement (e.g., Gopal et al., 2018; Schutz et al., 1998) and achievement (e.g., Hood et al., 2012) in tertiary education.

Many forms of motivation have been described. Wild and Pfannkuch (1999) emphasized the crucial role of motivation in fostering statistical literacy. Attributes such as curiosity, awareness, and engagement trigger an individual's initial interest in statistical problems, driving them to ask important questions and explore data. Engagement, in particular, intensifies these attributes, leading to deeper involvement and better performance, as people are more motivated to explore topics they find personally interesting. The authors argued that without such intrinsic motivation, students may struggle to engage deeply with statistical

thinking. Ultimately, motivation helps cultivate the critical mindset necessary for effective statistical analysis.

Gal (2002) emphasized that motivation—in the form of attributes such as critical stance, beliefs, and attitudes—is essential for statistical literacy. He argued that statistical literacy requires more than passive interpretation—it involves actively questioning and evaluating statistical information. Characterized by a willingness to challenge and question quantitative messages, a critical stance depends on certain beliefs and attitudes. People must believe in their ability to engage with statistical reasoning and feel empowered to critique information. A positive attitude toward statistics combined with self-confidence in one's reasoning abilities are necessary to sustain motivation for critical engagement with statistical data.

However, statistics as a subject is not very popular amongst students in tertiary education (Levpuscek & Cukon, 2022). Students from the social sciences have particularly negative attitudes toward it and even tend to feel statistics anxiety (Onwuegbuzie & Wilson, 2003; Tremblay et al., 2000; Zeidner, 1991). Negative attitudes toward statistics are more common among women and students with lower performance levels (Levpuscek & Cukon, 2022). However, to my knowledge, there are no studies that have assessed statistics motivation in primary school children. Thus, there is a need for further research that focuses on young children.

1.1.4 Assessment of statistical literacy

Instruments for assessing statistical literacy are required to measure children's levels and progress, for instance, at pretest and posttest in a randomized controlled trial. Important criteria for such instruments include validity, which ensures that the tool accurately measures the aspect of statistical literacy, and reliability, which guarantees consistent results across different applications and raters (Scholtes et al., 2011). Furthermore, practical aspects such as usability, particularly in diverse educational contexts, must be considered to avoid biases that are linked to linguistic or cultural backgrounds, ensuring that the tool is accessible and interpretable across varied populations (Wallace et al., 2009). These criteria collectively ensure that the assessment tool is both methodologically sound and contextually appropriate for assessing statistical literacy in children across time points. In the following, I introduce the approaches that exist for measuring the three aspects of statistical literacy highlighted above.

Data-based argumentation

Being able to critically evaluate interpretations of data can be summarized as data-based argumentation (Krummenauer & Kuntze, 2018) and is a requirement for being statistically

literate, for instance, to be able to participate in social discourse (e.g., Gal, 2002; Wild & Pfannkuch, 1999). Consequently, the appropriate measurement of data-based argumentation is required. Several studies have tried to assess data-based argumentation, but these studies employed different approaches to measure the construct.

In a study by Ben-Zvi (2008), data-based argumentation was assessed by reviewing video material, making research observations, conducting interviews, and reviewing the notebooks and project reports of 75 fifth-grade students in an exploratory data analysis project conducted in their mathematics and science lessons. The researchers qualitatively analyzed the materials to check for whether the students provided as much evidence as possible for proposed hypotheses and whether the students tried to produce and support opposing hypotheses. These qualitative measures helped to show that the students were able to produce some data-based arguments. However, the analyses were very complex and costly, and this study did not produce any statistical coefficients that could be used to judge its reliability.

Similarly, in a study by Paparistodemou and Meletiou-Mavrotheris (2008), the data-based argumentation skills of 22 third graders were assessed with audio recordings, research observations, interviews, students' reports, and students' final presentations. These data were used to assess students' abilities to use data from descriptive to inferential statistics in three levels. The authors found that third graders are already able to make some statistical inferences. However, the results were described qualitatively, and thus, quantitative measures to judge students' abilities and the quality of the measurement instrument were missing. Also, the qualitative procedure that was used was complex and costly.

In a study by van Dijke-Droogers et al. (2024), 267 ninth graders' data-based argumentations skills in inferential statistics were assessed. Most of the 39 open-ended items they administered were based on work by Watson and Callingham (2003, 2005) and were scored on up to seven different levels of reasoning ability ranging from low ability (e.g., statements based on personal experience) to high ability (e.g., statements based on data, acknowledging possibility but low likelihood). Cronbach's alpha was .84, thereby showing good internal consistency. This measurement instrument is suitable for use in quantitative studies. However, it is still quite complex and costly with its seven levels and 39 items.

Krummenauer et al. (2022) conducted an exploratory interview study aimed at investigating young primary school students' data-based argumentation. As the children could not be expected to spontaneously produce data-based arguments, a specific elicitation method was developed. In one-on-one interviews, students were presented with a series of standardized tasks. Each of the 11 tasks included a data set visualized with pictograms and a related

statement interpreting the data. The students were asked to evaluate the statement's validity and justify their evaluation, thus requiring them to develop data-based arguments. Two studies were conducted: the first involved 11 first-grade students, and the second involved 29 first graders. The interviewer avoided giving any examples or hints. The students' responses were transcribed and analyzed using a dichotomous top-down coding to determine whether they provided consistent data-based arguments. A consistent argument needed to include a correct evaluation of the statement and reference to the data that supported their evaluation. Answers that did not meet these criteria underwent further bottom-up analysis to explore the types of difficulties the students experienced. The interrater reliability for this analysis was very high in both studies ($\kappa_1 = .96$; $\kappa_2 = .88$).

Krummenauer and Kuntze (2019) conducted a study using a newly designed test instrument to assess 167 fourth-grade students from southern Germany (91 female, 76 male, average age 10.4 years). In one task, students were required to evaluate whether a headline matched a diagram and to justify their response by developing at least one data-based argument. The meaning of "justify" was explained through standardized instructions before the test. The students' responses were analyzed through a combination of top-down coding (interrater reliability $\kappa = .91$) and bottom-up analysis to explore potential difficulties. The top-down analysis was based on Toulmin's (2003) model of argumentation, requiring at least a conclusion (e.g., "no, the headline does not match the diagram") and a data-based argument referencing specific aspects of the data. Further elements, such as a warrant or backing, were unnecessary if the connection between the data and the conclusion was clear. Responses that did not meet these criteria were analyzed with an inductive bottom-up approach based on Mayring's (2015) category formation method to classify the difficulties encountered by students. This instrument is relatively simple, compared with previous instruments. However, it should consist of more than one item to be more informative.

There are various methods for assessing data-based argumentation in students, ranging from qualitative methods to more quantitative measures. However, in all the previous studies described above, all methods had in common that the participants had to produce their own data-based arguments. No studies used simpler approaches such as multiple-choice items. Some of the instruments were costlier and more complex than others because they included many sources that had to be analyzed (e.g., videos and notebooks; Ben-Zvi, 2008) or had to be rated on seven different levels (van Dijke-Droogers et al., 2024). Also, some instruments were tested only with older students. Krummenauer et al. (2019, 2022) used a relatively cost-

effective approach to assess data-based argumentation in primary school children. This approach can be used in quantitative studies by including a dichotomous rating approach.

Understanding the concept of variability

Variability in statistics is omnipresent (Cobb & Moore, 1997) and needs to be anticipated when developing data-based arguments (Wild & Pfannkuch, 1999). Thus, it is crucial to measure the understanding of the concept of variability when assessing statistical literacy. Several approaches for measuring the understanding of variability exist.

Watson et al. (2003) developed a questionnaire to assess the understanding of variability with 48 items from different areas (sampling variability, displaying variability, chance variability, describing/measuring variability, and sources of variability). They validated the reliability and unidimensionality of the questionnaire with item response theory and derived four levels of competency from it. At Level 1 (*Prerequisites for Variation*), students justify responses through personal experiences and simple visual comparisons, demonstrating limited interpretation skills. At Level 2 (*Partial Recognition of Variation*), students begin to recognize patterns and use basic chance statements but struggle with expressing and quantifying variation. Level 3 (*Applications of Variation*) includes an improved understanding of sampling and variation, though students often miss key aspects, such as the importance of variation in averages. Finally, at Level 4 (*Critical Aspects of Variation*), students show a deep understanding of variability, accurately summarize data, and critically analyze bias and uncertainty with statistical reasoning. The test's wide coverage of areas of variability makes it a good measure of students' overall understanding of statistical variability. However, it does not capture a more nuanced view of abilities and cognitive biases that are related to this understanding. Also, the questionnaire is rather long with a duration of 45 min, making it less practicable to use in research along with other variables.

Garfield and Ben-Zvi (2005) noted that traditional approaches for assessing the understanding of variability have primarily emphasized definitions, calculations, and basic interpretations of measures of spread. Building on this foundation, they reviewed the tasks used in research articles and considered their implications, along with examining additional prior literature (e.g., Ben-Zvi & Garfield, 2004b). Garfield and Ben-Zvi (2005) proposed a broad set of assessment items to evaluate students' reasoning about variability. Emphasizing realistic contexts, they aimed to engage students and encourage the use of informal knowledge. They also outlined how to design assessments with real-world data to cover key areas in the understanding of variability. They identified seven key areas: (1) *developing intuitive ideas of*

variability, (2) *describing and representing variability*, (3) *using variability to make comparisons*, (4) *recognizing variability in special types of distributions*, (5) *identifying patterns of variability in fitting models*, (6) *using variability to predict random samples or outcomes*, and (7) *considering variability as part of statistical thinking*. They recommended that teachers and researchers create items that are based on the above list, embedding tasks in real or realistic contexts to engage students in reasoning about variability and revealing their understanding of the concept.

English and Watson (2015) focused on 115 fourth graders to evaluate their understanding of variability through two measurement lessons. In the first lesson, all students measured the arm span of the same child in the classroom, so that the children learned about variability in measurement. In the second lesson, each student measured their own arm span, allowing them to observe and compare variability between different people. Both lessons focused on how students could explain the differences between measurements. Throughout the lessons, students documented their work in a workbook and software formats, which enabled the authors to assess students' engagement and growth in understanding the need to observe, represent, and contrast variability in data. Using the components of Garfield and Ben-Zvi's (2005) framework, English and Watson examined the students' workbooks with respect to the first three categories and the last (see above). After the lessons, a written assessment evaluated students' retention of the concepts they learned about variation and their ability to transfer that understanding to a different scenario. On average, the children were able to answer 16.7 out of 26 questions correctly. The authors concluded that initial skills in inferential statistics are already present and can be acquired by primary school age children. However, the implementation was rather time-consuming, as it was carried out during normal school lessons.

Rather than assessing the understanding of variability generally, children's understanding of variability can also be measured in specific behaviors. Weaknesses in dealing with statistical variability have already previously been related to so-called cognitive biases (e.g., Ben-Zvi & Garfield, 2008; Tversky & Kahnemann, 1971). For instance, the jumping to conclusions bias (Garety et al., 2005) can be viewed as the underestimation of variability in small samples. People with this bias tend to feel certain about inferences that are based on too little data. This bias has usually been assessed with the *beads task* (Dudley et al., 1997a, 1997b; Garety et al., 2005; Huq et al., 1988). The beads task is a widely used probabilistic reasoning task in which participants are asked to decide which of two jars beads are drawn from. For example, there may be a jar with 15 orange beads and 85 black ones and a jar with 85 orange beads and 15 black ones. The beads are mixed, and then the participants are told that one jar

was chosen randomly and they can draw beads from it. After each bead is drawn, the participant is asked whether they know which jar the beads were drawn from or whether they want to see another bead. Up to 20 beads can be seen before the test ends. After each turn, the drawn bead is returned to its jar, but the participants can always see all the beads that have previously been drawn in a memory aid. The more beads are drawn, the smaller the jumping to conclusions bias. This measurement instrument offers one valid perspective on participants' understanding of statistical variability. Students who make a decision on the basis of a smaller number of beads show a greater underestimation of variability in small samples. However, even though this measure has been useful in previous research, it comes with some disadvantages. First, it has low reliability because usually only one item is administered (Moritz et al., 2012). Second, more draws to decision do not always reflect a more cautious thinking style, as the probability of being right jumps up and down as more beads are drawn. Third, the beads task has been accused of being misunderstood by some participants (Balzan et al., 2017). And lastly, more draws to decision might not always be better than drawing fewer beads because more draws could also indicate a decision process that is too cautious and inefficient. Therefore, several studies have concentrated on establishing a similar construct: the decision threshold (e.g., Moritz et al. 2006; Moritz et al., 2020; Veckenstedt et al., 2011). The decision threshold describes the minimum probability value that needs to be met in order for a person to be confident enough to make a decision. It can be compared to the concept of the alpha level in science. A decision threshold of 95% would mean that a person will make a decision only when there is a 95% probability of being correct. This threshold of 95% is equivalent to an alpha level of 5%. However, current instruments for measuring the decision threshold rely on participants' subjective probability estimates and do not take objective probabilities into account.

Another cognitive bias that can be related to the understanding of the concept of variability is the alternation bias (Kahneman & Tversky, 1972; Tversky & Kahneman, 1971). The alternation bias describes people's tendency to expect more alternation between random events than probabilistic laws would forecast on average. Therefore, people with this bias would expect certain patterns of variability more than other patterns. In small samples, they would tend to predict patterns that are similar to the patterns found in larger samples (i.e., a 50-50 distribution of heads and tails in a coin throw sequence). The alternation bias can be assessed in several ways. For example, participants could choose between different sequences of two alternating random events and decide which sequence looks the most random (e.g., Falk & Konold, 1997). Another possibility would be to let participants generate their own sequences

(e.g., Falk & Konold, 1997; Warren et al., 2018; Yu et al., 2018). In another approach, participants could manipulate the alternation rate of random sequences to alternate more or less and stop when they think that their sequence seems to reflect their concept of randomness (Yu et al., 2018).

Measuring people's understanding of variability in statistics can include broad assessments or specific behavioral evaluations to capture different facets of statistical literacy. Tools such as Watson et al.'s (2003) questionnaire and Garfield and Ben-Zvi's (2005) assessment provide broader insights, whereas more specific approaches can target behaviors that tend to accompany cognitive biases that affect variability perception, such as the jumping to conclusions bias (Garety et al., 2005). Together, these methods highlight the importance of different approaches for assessing aspects of understanding variability.

Motivation in data-related tasks

Many scholars have stated that motivation in data-related tasks is crucial for statistical literacy (e.g., Gal, 2002; Watson & Callingham, 2003). Statistics motivation has been positively linked to statistics engagement (e.g., Gopal et al., 2018; Schutz et al., 1998) and achievement (e.g., Hood et al., 2012) in tertiary education. To be able to assess change in motivation in intervention studies, appropriate measurement instruments are required. There are a few measurement instruments that were specifically designed for statistics motivation.

For instance, negative aspects of motivation can be assessed with the Statistics Anxiety Rating Scale (STARS; Cruise et al., 1985). This instrument is widely used for testing statistics anxiety (Levpušček & Cukon, 2022). It consists of 51 items that cover six different factors of statistics anxiety: (1) *interpretation anxiety*, (2) *test and class anxiety*, (3) *fear of asking for help*, (4) *worth of statistics*, (5) *computation self-concept*, (6) *fear of statistics teachers*. A sample item from the worth of statistics factor is "Statistics takes more time than it's worth." Cruise et al. (1985) found that all items loaded onto these six factors, and they found good internal reliabilities for all of them ($.68 \leq \alpha \leq .94$). The six-factor structure was also validated by Hanna et al. (2008). This scale is applicable for tertiary education. However, for primary education, this instrument might be too specific to college courses and too long to assess along with other constructs in an intervention study.

Krause et al. (2009) investigated whether cooperative learning and feedback support the development of statistical literacy. The study included 137 students from the University of Munich, all of whom participated voluntarily. A pretest and posttest on motivation were administered. For two single items, students rated their own perceived performance and

perceived competence from 1 (*very good*) to 6 (*very bad*/*none*). Topic interest was measured with six items (e.g., "I am interested in correlation analysis") that were rated on a 6-point response scale that ranged from 1 (*strongly disagree*) to 6 (*strongly agree*). Internal consistency was high with Cronbach's $\alpha = .81$. Positive intervention effects indicated that the scales were sensitive to change. However, these scales also seem specific to tertiary education and unfit for primary education.

There is not much research on motivation specifically with respect to data-related tasks and especially not in primary education. However, there are many scales that can be adapted from other domains. For example, Marsh's (1990) Self-Description Questionnaire I (SDQ I) and its German version (Arens et al., 2011) are used to assess self-concept. Arens et al. (2011) found evidence of the reliability and validity of this instrument in a sample of $N = 589$ third to sixth graders. Gaspard et al. (2015a, 2015b) developed and evaluated an instrument for assessing different value beliefs. These value beliefs were intrinsic value, attainment value, utility value, and cost in the domain of mathematics, all of which showed good internal consistencies ($.84 \leq \alpha \leq .94$) and model fit (Gaspard et al., 2015a). These well-established scales can easily be adapted to the domain of statistics and could offer an easier way to measure motivation in data-related tasks in intervention studies, rather than taking already existing instruments from the statistics domain, as most of these are long and designed for older students. However, evidence of the validity of any adapted scales is needed.

1.2 Promoting Primary School Children's Statistical Literacy

The development of statistical literacy in primary school children is essential for promoting their ability to make informed decisions in an increasingly data-driven world (Bodemer, 2014; Wallman, 1993). This subchapter explores how key aspects of statistical literacy (data-based argumentation, understanding the concept of variability, and motivation in data-related tasks) can be effectively promoted through targeted interventions. By identifying strategies that enhance children's comprehension and engagement with statistical concepts, in this section, I seek to provide insights into educational practices that support key aspects of statistical literacy in real-world contexts.

1.2.1 Promoting data-based argumentation

Being able to communicate interpretations of statistical data is one of the central aspects of statistical literacy (e.g., Gal, 2002; Schield, 1999). Therefore, learning to use data-based

argumentation is a key requisite for becoming statistically literate. A small number of studies have tried to examine how data-based argumentation can be promoted.

A study by Ben-Zvi (2006) aimed to promote data-based argumentation and informal inference among fifth graders through an interdisciplinary learning environment. Using TinkerPlots (Konold & Miller, 2005), a dynamic data visualization tool, students engaged in open-ended data analysis tasks that involved progressively larger sample sizes. This tool visualizes data and is supposed to help students understand statistical concepts without prior knowledge of conventional graphs. Ben-Zvi's (2006) scaffolded approach allowed the fifth graders to explore patterns, understand variability, and make informal statistical inferences. Cooperative peer work and class discussions promoted argumentative reasoning, where students supported their inferences with data-based arguments. The study demonstrated that even young students can develop informal reasoning akin to expert practices, though some struggled to shift from individual to aggregate data perspectives. Overall, the research highlighted the potential of tools such as TinkerPlots and experiences with real-world data for enhancing data-based argumentation skills in primary education.

A study by Paparistodemou and Meletiou-Mavrotheris (2008) investigated how third graders make data-based inferences by using a hands-on, project-based approach with the software TinkerPlots (Konold & Miller, 2005). The intervention involved 22 third-grade students in Cyprus who completed personal journals on health, nutrition, and safety habits, analyzed the data with TinkerPlots, and presented their findings over 4 weeks. Researchers observed and interacted with students to understand their reasoning processes. Results showed that students were highly engaged and used TinkerPlots to make data-based arguments and generalizations, often drawing on personal experiences. They expressed informal inference through data-based argumentation, generalization, and chance. Paparistodemou and Meletiou-Mavotheris concluded that an early introduction to statistical reasoning is beneficial and feasible with appropriate tools, emphasizing the importance of real data, personal interest, and context in motivating students.

A study by van Dijke-Droogers et al. (2024) aimed to evaluate the effects of an intervention for ninth-grade students' statistical literacy, focusing in particular on making inferences based on repeated sampling within an inquiry-based approach. The researchers used a pre-post design to measure the impact of the intervention on three postulated domains of statistical literacy (statistical inference; average and chance; and graphing and variation). The study found significant positive effects on students' statistical literacy for all three domains with an overall effect size of $d = 0.90$. The findings suggest that an early introduction to

inferential statistics, supported by an inquiry-based approach, can enhance data-based argumentation about inferences. However, the study did not include a randomized control group. Therefore, we cannot be sure whether the observed changes were due to the intervention.

Even though only a few studies have focused on promoting data-based argumentation, there is evidence that it can be successfully promoted in children. However, more evidence is needed to understand which approaches are most useful and whether it can also be promoted in primary school children. In particular, more studies with more robust designs are needed (e.g., randomized controlled trials).

1.2.2 Promoting the understanding of the concept of variability

It is crucial to promote the understanding of variability for enhancing children's statistical literacy because it helps them grasp the inherent fluctuations in data and recognize that not all outcomes are fixed or predictable (Wild & Pfannkuch, 1999). Understanding variability allows children to better interpret graphs, averages, and data trends. It also empowers them to evaluate uncertainty and make informed judgments, aspects that are essential for navigating real-world situations where variability is a constant factor. By developing this foundational concept, children can more accurately analyze data and draw meaningful conclusions, thereby building stronger statistical literacy.

Studies on the promotion of the understanding of variability are scarce. However, there are several models on the development of the understanding of variability (Peters, 2011; Reading & Shaughnessy, 2000, 2004; Shaughnessy et al., 1999; Torok & Watson, 2000; Watson et al., 2003). Fife et al. (2020) aggregated the models to represent how learning progresses. They proposed five levels of sophistication that portray how students learn about variability from below sixth grade to the university level. They range from a *naïve understanding of variability* (Level 1) to a *robust understanding of variability* (Level 5). At the lower end, students often rely on personal stories and unrelated factors in reasoning, and they misinterpret variability. At higher levels, however, students can compute measures of central tendency and spread, critically evaluate variability in studies, and use context to explore sources of variability.

Even though there have been many efforts to describe the theoretical development of the understanding of statistical variability, studies on how to promote it are scarce. Ben-Zvi (2004) wondered how students begin to reason about variability. A seventh-grade class engaged in an activity in which the lengths of surnames in a data set with their own 35 Hebrew names and 35 given English names were compared. Then two highly able seventh graders were

selected as participants for closer observation. They were videotaped and interviewed, and their notebooks were analyzed. In the beginning, the students had problems comparing the lengths of the two groups of surnames and focused, for instance, on details such as names beginning with “Mc,” but the author found several factors that helped the students learn to reason about variability in a more global way. Technological tools helped them explore the data in different forms of representation. Their interactions with the teacher helped them adopt a more statistical perspective. The context of the surname problem inspired them to include contextual reasons for the difference between the two groups. The more the students worked with the data, the more they were able to see the distribution as a whole instead of as a group of single numbers.

Bakker (2004) conducted a study to find out how reasoning about variability, sampling, and distributions develops in eighth-grade students with little statistical knowledge. In two sessions, 30 students reasoned about a growing sample and the shapes of distributions. The students went through cycles in which they had to construct hypothetical diagrams of weight distributions of eighth graders and then compare them with real data. The author found that, in the discussions between the cycles, the students discussed variability a great deal. In the session about the shapes of distributions, the students were confronted with different shapes (e.g., skewed distributions). The author found that the students began to reason about values of central tendency and spread in a very informal way. He concluded that his sessions were a good way to get students engaged in thinking about variability, sampling, and distributions, all of which can be built on with more formal training.

A key objective of most introductory statistics courses is to help students recognize the omnipresent nature of variability and to understand how to measure and explain it (Cobb, 1993). Garfield and Ben-Zvi (2008) conceptualized an intervention to improve students' understanding of variability by using a research-based activity in an introductory statistics course. In this intervention, students compared histograms on the sizes of their standard deviations, discussed their answers, entered data into software, and calculated standard deviations. This process was supposed to help them understand factors that influence the sizes of standard deviations, such as the distribution of data points relative to the mean, while identifying factors that did not affect it. The study emphasized the need for activities that integrated reasoning about variability into teaching plans, rather than treating it as an isolated topic. Garfield and Ben-Zvi (2008) concluded that students could develop a deeper understanding of variability by engaging in activities that encouraged informal reasoning about data distributions, leading to formal comparisons of variability measures. However, the authors

did not execute their intervention, so the question of whether it can promote reasoning about variability has yet to be answered.

In a study by Dierdorp et al. (2017), 13 students from the 12th grade learned about variability by engaging in real-world heart rate measurement tasks that highlighted inconsistencies and “noise” in data to help them understand variability as part of authentic contexts. They noticed and explained differences in data, proposed strategies for controlling variability, and saw how statistical tools such as regression help find the signal within the noise in the data. With teacher support, they moved beyond calculations to make meaningful interpretations of data, using trends to generalize results. Despite there being no pretests or posttests, the authors concluded that the tasks effectively promoted students' reasoning about variability.

Even though the research on promoting the understanding of statistical variability is scarce, there have been a few attempts that show preliminary evidence that letting students gather and work with actual data could enhance their understanding of variability. However, most of these studies have been rather qualitative, and better evidence is needed, preferably from randomized controlled trials with a pre- and posttest design. Also, more studies focusing on primary school education are needed.

1.2.3 Promoting motivation in data-related tasks

Motivation is seen as a crucial aspect of statistical literacy by many scholars (e.g., Gal, 2002; Wallman, 1993; Watson, 2006; Wild & Pfannkuch, 1999). To make use of their statistical literacy, children must be motivated to do so. Therefore, it is crucial to promote motivational aspects of statistical literacy. However, research on fostering statistics motivation is scarce and has predominantly centered on tertiary education (e.g., Gopal et al., 2018).

A study by Krause et al. (2009) examined the impact of cooperative learning and feedback on the performance and motivation of 137 university students in a statistics e-learning environment. Feedback significantly improved outcomes, especially for students with less prior knowledge, although it did not enhance their perceived competence or performance. However, cooperative learning did not improve individual test results but it did positively affect students' self-efficacy and perceived competence. This result suggests that whereas group work might not lead directly to better learning outcomes, it fosters motivation and confidence, which could benefit long-term engagement with the subject.

A study by Acee and Weinstein (2010) examined whether a value-reappraisal intervention could boost motivation and performance in an undergraduate statistics course. A

total of 82 students were assigned to either the intervention or the control group, with value-reappraisal participants receiving messages about the significance of statistics and engaging in activities that helped them relate its relevance to their personal goals. The value-reappraisal condition led to significant improvements in task value ($d = 0.54$), endogenous instrumentality ($d = 0.50$), and interest ($OR = 9.23$) compared with the control group. However, the impact on exam performance was significant only for students who were taught by one of the two instructors, suggesting that instructor-related factors may moderate intervention effectiveness. This finding highlights the potential for value reappraisal in motivational interventions but also indicates the need to consider instructor influence in future efforts to support value perceptions in educational contexts.

Gopal et al. (2018) conducted a quasi-experimental study with 50 first-year students, assigning them to either an experimental group that uses RStudio with traditional instruction or a control group that received only traditional instruction. Results showed that the experimental group had significantly higher motivation for learning statistics than the control group, particularly in areas such as self-efficacy, active learning strategies, and the perceived value of learning statistics. The findings suggest that integrating technology such as RStudio could effectively promote motivation for learning statistics, making it a valuable tool for enhancing students' statistical literacy. However, this quasi-experimental study compared only posttest values between the experimental and control groups. Thus, this conclusion needs further support, as the differences could have already existed before the intervention.

Even though research on promoting statistics motivation is scarce, there is some evidence that fostering statistics motivation is possible in various ways. However, more evidence is needed to find the most effective methods and to show whether fostering statistics motivation in primary school children is possible as well.

1.3 Development of a Statistical Literacy Intervention

1.3.1 Context and steps of development

Because statistical literacy helps citizens make evidence-based decisions and helps democracies function, it is important to support its development from early on (e.g., United Nations 2012; Wallman, 1993). Thus, this dissertation focuses on the development and evaluation of a statistical literacy intervention for primary school children. This intervention is part of an extracurricular enrichment program for gifted primary school children called the Hector Children's Academy Program (HCAP; for more information, see Trautwein et al., 2023).

The HCAP offers various Science, Technology, Engineering, and Mathematics (STEM) courses that the students can participate in voluntarily. To participate in the program, students are nominated by their teacher on the basis of individual motivation, interest, and performance. One role of the scientific support team of the HCAP is to provide so-called Hector Core Courses (HCCs). The HCCs are scientifically evaluated courses that are designed by either the members of the scientific support team themselves or practicing course instructors. Another goal of the scientific support team is to evaluate these HCCs.

In the context of the HCAP, several HCCs have been developed on the basis of scientific theory and evidence and evaluated in randomized controlled field trials (e.g., Herbein et al., 2018a, 2018b; Rebholz et al., 2022; Schiefer et al., 2021). The development of such interventions usually follows a six-step approach to gradually develop and evaluate the real-world effectiveness of an intervention. These steps are (1) detecting the needs of the target group, (2), developing the intervention on the basis of scientific theories and evidence, (3) running a pilot study, (4) conducting an efficacy study, (5) conducting an effectiveness study, and (6) running a scaling-up study (Herbein et al., 2018b; Humphrey et al., 2016; Lendrum & Wigelsworth, 2013). This trajectory ensures that the design and evaluation of an intervention are based on scientific evidence and real-world needs. The studies should be conducted as randomized controlled trials, which are regarded as the *gold standard* in educational and psychological research (Lendrum & Wigelsworth, 2013; Torgerson & Torgerson, 2013).

1.3.2 Potential of gifted primary school children

The target group of the intervention in this dissertation consisted of gifted children from the HCAP. Gifted children are assumed to be the decision-makers of the future (Lee et al., 2021). With the aim of eminence, it makes sense to promote gifted children (Subotnik et al., 2011). Decision-makers in high positions with a lack of statistical understanding could cause a lot of damage. However, when they are statistically literate, there are higher chances that they will make the right decisions for our societies. Therefore, it is important to strive for the early promotion of skills and interests in gifted children in the field of statistics. This kind of early exposure could form a basis and a stimulus for further developments in their individual lives.

In the past, research has placed a stronger focus on students from secondary and tertiary education than on students in primary education. However, findings suggest that there is already a foundation of statistical literacy in primary school students, and this foundation can be built on. For instance, Kuntze et al. (2015) tested primary and secondary school students' abilities to use models and representations in statistical contexts in Columbia and Germany.

The authors found that the students could reach the first competency level fairly easily, but the tasks on the second and third levels were harder to solve. This outcome means that the students could read single data points from graphs, but it was more difficult for them to get more information from comparing single data points and even more difficult to draw inferences from it.

The literature suggests that statistical concepts can be taught at a young age. For example, Martignon and Wassner (2005) found that young children already understand a sample's representativeness, which can be linked to their understanding of fairness. Already in 1973, Engel used simplified roulette games to teach students about probabilities with fractions. Martignon and Wassner (2005) repeated the experiment and found that young students were also able to learn strategies linked to the probabilities of success. Other studies have shown that even many adults have problems with statistical concepts (e.g., Fischbein & Schnarch, 1997; Konold et al., 1993; Shaughnessy, 1977). Thus, statistical literacy could be fostered early in education, in order for students to build a foundation throughout their school lives.

To the best of my knowledge, no intervention studies have focused on statistical literacy and giftedness. It appears to make sense to train highly gifted children first because statistical concepts are challenging to understand. Sproesser et al. (2018) showed that the ability to use models and representations in statistical contexts is moderately correlated with nonverbal and verbal cognitive abilities. Therefore, students with higher cognitive abilities may be able to grasp the concepts more quickly and benefit from a higher average class performance (Dumont et al., 2013). One way to lead to a deeper understanding of the underlying statistical concepts could be to develop extracurricular interventions that promote students' statistical literacy. Highly gifted children can be a good starting point to investigate because previous studies have shown that it is rather difficult for young children to achieve higher competence levels (Krummenauer & Kuntze, 2019; Kuntze et al., 2015; Watson, & Callingham, 2003), and statistical literacy is related to cognitive abilities (Sproesser et al., 2018).

1.3.3 Core components of the intervention

To promote statistical literacy in our target group, activities in the intervention had to be designed in such a way that they could be assumed to be effective. To do so, we designed core components that describe design principles that are believed to be effective for reaching the instructional goals (Blase & Fixsen, 2013; Nelson et al., 2012). This approach also offers the opportunity to assess the fidelity of the intervention, how well the core components have been implemented, and how their implementation is related to the effectiveness of the

intervention (Blase & Fixsen, 2013; Humphrey et al., 2016; Nelson et al., 2012). With this approach, the effectiveness of the core components can be examined and the findings can be used for further intervention development. Two core components were selected for the current intervention.

The first core component is the predict–observe–explain (POE) approach (Gunstone & White, 1981). The POE approach is an instructional strategy designed to enhance student understanding through active engagement in the learning process. It typically involves three stages. In the predict phase, students make predictions about the outcome of an experiment or observation. In the observe phase, students conduct the experiment or observe a phenomenon. Finally, in the explain phase, students explain the results, reflecting on their predictions and the observed outcomes. Because an assumption must be made before the experiment is carried out in this approach, it is possible to compare it with the outcome of the experiment, and it is possible that it will cause cognitive dissonance, which can be particularly fruitful in terms of learning success. This method encourages critical thinking and helps address misconceptions by allowing students to reassess their predictions on the basis of actual evidence (Radovanović & Sliško, 2013). A meta-analysis by Gustina et al. (2023) provided support for the effect of the POE approach on learning outcomes in mathematics and natural sciences with an effect size of $d = 0.66$.

The second core component is cooperative learning (e.g., Capar & Tarim, 2015). In a meta-study by Johnson et al. (2000) of 164 different studies, the authors found positive effects of different cooperative learning methods on school achievement ($0.18 < d < 1.03$) compared with individualistic approaches. In cooperative learning environments, students interact and communicate with each other, thereby helping each other construct meaning. They learn by exchanging views, predictions, and explanations. Cooperative learning was found to be more effective than traditional methods in improving both mathematics achievement and attitudes. A meta-analysis by Capar and Tarim (2015) found that cooperative learning had a medium, positive, significant effect of $d = 0.59$ on students' mathematics achievement compared with traditional methods. It also showed a small, positive, significant effect of $d = 0.16$ on students' attitudes toward mathematics. Similarly, a meta-analysis by Kalaian and Kasim (2014) found that cooperative learning methods significantly improved students' academic achievement in college-level statistics courses, with an overall effect size of $d = 0.60$.

In the current intervention, the children experienced the POE approach and cooperative learning in playful research activities. The connection of the two core components might be additionally fruitful for the intervention’s effectiveness. In the predict phase, the children were

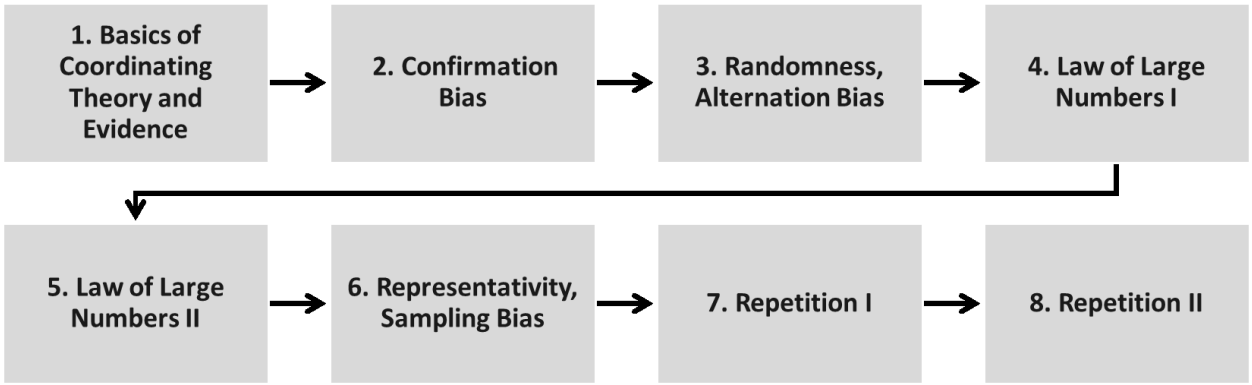
given the opportunity to exchange their views on the predictions that they were generating. In the observe phase, the children could work together to collect the data and therefore had a more motivating experience. In the explain phase, the children could learn together and exchange their views on the explanation of the observed phenomenon and communicate to each other anything that they learned from the adult course instructor. Overall, there are theoretical reasons and practical evidence that the two defined core components could support the effectiveness of the current statistical literacy intervention.

1.3.4 Contents of the intervention sessions

The statistical literacy intervention "Luck or Genius? Understanding Data and Making Predictions" consisted of eight sessions, each 90 min long. The sessions and their contents are described in the following (see Figure 1).

Figure 1

Contents of the Intervention "Luck or Genius? Understanding Data and Making Predictions"



Session 1: Basics of coordinating theory and evidence

When dealing with data, children often encounter requirements that are also present in scientific thinking (e.g., Kuhn, 2011; Zimmerman, 2007). For example, it is often necessary to check interpretations of data against the respective data or to interpret the data themselves, that is, to make statements about data that are consistent with them. Such activities are described in the literature as the coordination of theory and evidence (e.g., Kuhn, 2011).

In the first session of the course, the aim was to develop specific, sustainable strategies for coordinating theory and evidence to counter any misconceptions the children may have, as misconceptions can hinder their handling of data in a variety of ways. The strategies that should be developed and practiced in the first session included, in particular, being able to cognitively separate elements of theory (e.g., statements, one's own interpretations, and also one's own

ideas) from the available evidence (i.e., the data) and to treat each separately, to consider that theories can be wrong, and to reject or adapt them accordingly if counter-evidence is available (e.g., Morris et al., 2012; Zimmerman, 2007).

Concerning possible misconceptions, children tend to mix theory and evidence (e.g., Kuhn, 1989), for example, they often mix the available data with their ideas or accept statements prematurely instead of using the data to carefully consider alternatives. Children can also be expected not to reject statements of which they are convinced, even when counter-evidence is available in the data; and children can be expected to try to confirm statements and possibly ignore or reinterpret available counter-evidence. The first session was geared toward counteracting such tendencies by giving the students adequate strategies for coordinating theory and evidence.

To summarize, in Session 1, the children learned to separate theory and evidence, and they learned that evidence can be used to support theories. This goal was achieved by having the children participate in activities in which they had to generate hypotheses and check whether the given or gathered data supported the hypotheses or not.

Session 2: Confirmation bias

The second session was about making the children aware of the confirmation bias (Wason, 1968) and showing them what strategies they could use to counteract it. Successful theory and evidence coordination must integrate the evidence that is suitable for particular theories. The confirmation bias plays a significant role in the context of cognitive biases: People are more inclined to believe information if it supports their beliefs, expectations, and opinions (Wason, 1968). This process usually happens unconsciously and unintentionally and at the same time means that people are less likely to select information that is less in line with their opinions, beliefs, or expectations, pay less attention to it, and remember it to a lesser extent (Vedejová & Čavojová, 2022). Among other things, selection serves to maintain self-image and leads to the fact that the same situation can be interpreted completely differently by two different people.

The children learned about the confirmation bias by playing a game that was an adaption of Wason's (1960) 2-4-6 task. In the original experiment, participants were asked to find the hidden rule to sets of three numbers. The first example of a set following the hidden rule was 2-4-6. Then the participants would often guess other sets of numbers that ascended by 2 each time. However, they usually did not try to find sets of numbers that would disconfirm the belief that the hidden rule was that the numbers were ascending by 2 each time. The actual

hidden rule that it was any ascending set of numbers was often not guessed. In the game in the current intervention session, the children had to figure out a rule that determined which animals were allowed in an imaginary zoo. The secret rule in the first round was that the only animals that lived in the zoo were those that can lay eggs. Then the children were told about three animals that lived in that zoo: eagles, flamingos, and doves. The animals were chosen so that the children were misled to believe that only birds or animals with wings lived in this zoo. Then each child was allowed to ask whether two specific animals lived in the zoo. The children often chose a positive testing strategy (e.g., asking if other birds such as a blackbird lived in the zoo), which is in line with the confirmation bias (Wason, 1960, 1968). With this approach, in most of the groups, the children arrived at the wrong rule at the end of the game. This game was used to show them that they should adopt a falsifying testing strategy, which could then be used in the next round of the game. Further activities in the session were used to deepen the children's knowledge of the confirmation bias and a falsifying testing strategy.

Session 3: Randomness and the alternation bias

The third session was about uncovering and discussing the children's concepts of chance. Teaching primary school children about randomness is crucial for developing statistical literacy. Understanding randomness helps children grasp the concepts of variability and uncertainty, which are fundamental in interpreting data and making informed decisions (e.g., Wild & Pfannkuch, 1999). Early exposure to these concepts can enhance children's abilities to understand and interpret statistical information presented in various forms, such as graphs and tables, which are common in everyday life.

As a starting point for this topic, the children came up with a subjectively random sequence of coin tossing events and then compared it with randomly tossed sequences. In doing so, they discovered the alternation bias (e.g., Yu et al., 2018), people's tendency to have more alternations between different random outcomes than can be expected on average. After this demonstration, the children's various concepts of chance were discussed, and a common definition of randomness was established. The understanding of the alternation bias was then deepened with a small activity on combinatorics.

Sessions 4 and 5: The law of large numbers

The law of large numbers states that the probability of an event approaches the theoretical probability more and more as the number of repetitions of the same experimental situation increases (e.g., Blume & Royall, 2003). In terms of theory and evidence coordination, this law means that larger samples are more likely to represent the true answer than smaller

samples, and larger samples can be weighted more. The law of large numbers is an essential part of statistical thinking, as it smooths out inconsistent measurement results and thereby makes more reliable statements about the underlying probabilities. It is an essential way to deal with variability.

In the fourth session, the children learned about the law of large numbers in a playful way. Each child received a die, and they had to find out whether it was weighted or not. To figure out whether it was weighted, they threw the dice and gathered their dice results as data in a diagram. After 2, 5, and 10 min, they had to take a guess about whether they thought their die was weighted or not. Usually, in the beginning, it was not as clear, and they often thought that their die was weighted, even when it was not. For example, they often thought that rolling a six three times whilst no three was rolled meant that the die was weighted. However, the more often they threw their die, the more clearly they could see whether their die was really weighted. The weighting became especially clear when all the children compared their diagrams with each other. This activity was then used to introduce the law of large numbers. This understanding was further exercised through other smaller activities.

In the fifth session, the children deepened their understanding of the law of large numbers by doing a large experiment with paper planes. In the beginning, the children had to predict whether a paper plane with a sharp tip would fly farther than a paper plane with a flat tip, whether it would be the other way around, or whether the two planes would fly about the same distance. Then data were gathered. Children threw the paper planes in pairs and wrote down their observations. The more data were gathered, the clearer the result. This result was then talked about in a group discussion, highlighting again that more data lead to clearer results. Also, it was mentioned that different kinds of questions need different amounts of data because measures of central tendency and spread can vary, thus resulting in less or more clear differences between groups. Whereas it may have been clear after 10 throws of both paper planes that the one with the flat tip generally flew farther than the one with the sharp tip, it was not necessarily as clear after 10 throws whether Child A threw the planes a farther distance than Child B.

Session 6: Representativity and sampling bias

When drawing a sample, care must be taken to ensure that the population is represented as well as it possibly can be by the sample. The population is the totality of all units from which the sample is drawn. The sample is a subset of the population that is selected in order to draw inferences about the population. Ideally, sampling should be completely random so that all

cases have the same probability of occurring. We can then assume that the sample (if sufficiently large) represents the population well. However, sampling bias can also occur if, for example, one part of the population is not surveyed or is surveyed less or more than other parts (Ellenberg, 1994).

In the sixth session, the children learned about the concept of representativity by playing a game. In this game, the children helped a fictional mayor decide what should be built in a park: a slide, a kiosk, or a flower bed. The game was conceptualized so that the players most often surveyed children of this fictional city instead of adults and pensioners. In this way, the participating children thought that the mayor should build a slide, which was what the *fictional children* wanted most often. However, after the game, the answers of *all the fictional citizens* were disclosed, showing that the citizens actually wanted the flower bed the most. This task was used to talk about representativity and how the children could have come to the right conclusion (i.e., by drawing randomly or evenly across different groups of citizens).

Sessions 7 and 8: Repetition

In the seventh and eighth sessions, games were played to reinforce what the children had already learned. Also, all the terms they already learned were placed in a diagram in relation to each other. The last activity was a quiz game in which the children needed to develop data-based arguments based on fictional hypotheses and given data.

1.4 Research Questions of the Present Dissertation

Because our society needs more statistically literate citizens, and there is currently very little evidence-based research on statistical literacy, this dissertation addresses the question of how statistical literacy can be promoted. The population that is the focus of this dissertation is gifted primary school children. This special target group has high learning potential and makes it possible to explore the extent to which statistical literacy can be promoted. It also allows these children to be supported in the development of eminence, which is particularly important, as they are likely to be the decision-makers of the future in important positions (Lee et al., 2021; Subotnik et al., 2011).

As a first step, we developed a statistical literacy intervention based on evidence from existing studies (see Chapter 1.3). In Study 1 (*Evaluating the Efficacy of a Statistical Literacy Intervention*), we tested this intervention in a standardized way. In particular, academic staff from a university carried out the intervention to ensure adherence to the course manual. The goal of this intervention was to promote children's data-based argumentation, their understanding of variability, and their motivation in data-related tasks. As previously stated,

research on the promotion of these aspects of statistical literacy is scarce, especially in a primary school setting (see Chapter 1.2). However, we developed the intervention on the basis of two core components (Blase & Fixsen, 2013; Nelson et al., 2012) to create a suitable learning environment that promoted the abovementioned aspects of statistical literacy. First, the POE approach (Gunstone & White, 1981) created a setting in which students could reveal their assumptions and views about certain topics that could later be compared with actual data. This comparison helped them rethink their arguments on the basis of data and provided them with experience with handling variability. Second, cooperative learning methods (e.g., Capar & Tarim, 2015) provided a motivating learning environment by letting the children learn in teams in which they could exchange their ideas and views in their own language. Together, these core components were supposed to enhance children's data-based argumentation, understanding of variability, and motivation in data-related tasks. However, motivation could also decrease, as children's motivation in these courses tends to be high at baseline (e.g., Herbein et al., 2018a, 2018b; Rebholz et al., 2022; Schiefer et al., 2021) and because of the big-fish-little-pond effect (Marsh, 1987; Marsh & Parker, 1984). This effect describes the social comparison process in which learners compare themselves with their peers. And as the children in this study came from their regular classes into an environment with a higher average ability level, they might see themselves as less competent compared with their new peers.

Study 2 (*Validating an Adapted Version of the "Beads Task": Assessing Decision Thresholds in Talented Primary School Children Using Signal Detection Theory*) focused on the development and evaluation of a newly adapted measurement instrument for assessing the jumping to conclusions bias (Garety et al., 2005). This bias describes people's tendency to make hasty conclusions on the basis of too little evidence. This bias is usually assessed with the beads task (Dudley et al., 1997a, 1997b; Garety et al., 2005; Huq et al., 1988), and the measure that is used is called draws to decision. Fewer draws to decision are usually associated with a larger jumping to conclusions bias. However, even though the beads task has been proven useful in scientific research, there are several concerns about its reliability (Moritz et al., 2012), its comprehensibility (Balzan et al., 2017), and the linearity of its interpretation (see Chapter 1.1.4). Therefore, Moritz et al. (e.g., 2006, 2020) proposed the idea of a decision threshold, which describes the minimum probability of being right that a person needs to be certain enough to make a decision. This probability can be used as a better way to interpret the jumping to conclusions bias. In the study in Paper 2, my co-authors and I further developed this instrument to be used by children and to be based on objective probability values instead of subjective ones. We used signal detection theory (SDT; Green and Swets, 1966; Stanislaw

& Todorov, 1999) to create a more objective version of the decision threshold. This instrument was developed and evaluated for subsequent use in Study 3.

Study 3 (*Promoting Primary School Children's Statistical Literacy: Results From a Randomized Controlled Field Trial*) focused on determining the effectiveness of the previously developed statistical literacy intervention. Such effectiveness studies examine the intervention's effects under real-world conditions (Gottfredson et al., 2015). In this study, course instructors from the field were trained to conduct the intervention themselves instead of having academic staff from universities conduct the intervention. The training was conducted to ensure that the course instructors were aware of the content of each session, what the core components are, and how they are assumed to work. This change of course instructors raises the question of whether the effects measured in the efficacy study could still be observed in the effectiveness study. In addition, the course instructors documented the extents to which they were able to adhere to the course concept (Nelson et al., 2012), which made it possible to see whether the fidelity of implementation was successful and whether there were significant relationships with the effects of the intervention.

2

Study 1: Evaluating the Efficacy of a Statistical Literacy Intervention

Stark, L., Krummenauer, J., Jaggy, A.-K., Kremer, F., Kuntze, S., Nagengast, B., Trautwein, U., & Golle, J. (2025). *Evaluating the efficacy of a statical literacy intervention*. Manuscript in preparation.

The Hector Foundation II supported this work. Lucas Stark and Fabienne Kremer are doctoral students at the LEAD Graduate School & Research Network [GSC 1028], funded by the Baden-Württemberg Ministry of Science, Research and the Arts within the framework of sustainability funding for the projects of the Excellence Initiative II.

Abstract

Statistical literacy is crucial for data-based argumentation and informed decision-making but remains challenging due to common misconceptions and biases. We developed an 8-week statistical literacy program that focuses on data-based argumentation, dealing with variability, and motivation and combines the predict-observe-explain approach with cooperative learning. The intervention's efficacy was tested in a multisite randomized-controlled field trial with 53 third- and fourth-grade students who were randomized to the intervention or waitlist control group. Preregistered multilevel multiple regression analyses indicated positive effects of the intervention on data-based argumentation, views on variability, draws to decision, and self-concept in data-related tasks relative to the waitlist control group. The efficacy of the intervention showed that statistical literacy can already be promoted at primary school age.

Keywords: statistical literacy, randomized–controlled field trial, intervention study, primary school children, data-based argumentation

Evaluating the Efficacy of a Statistical Literacy Intervention

The interpretation of data is playing an increasingly important role in today's world, particularly against the background of the unprecedented growth of digital technologies and the availability of vast amounts of information (Mitchell et al., 2021). As failing to consider data is associated with superstitious or conspiracy-related beliefs (e.g., Kuhn et al., 2022), a crucial aspect of empowering citizens involves equipping them with the competencies needed to face statistical challenges in their daily lives, commonly referred to as being "statistically literate" (e.g., Franklin et al., 2005; PARIS21, 2020). Statistical literacy is associated with various abilities that enable individuals to comprehend and interpret statistical data, facilitating evidence-based decision-making (Ben-Zvi & Garfield, 2004; Callingham & Watson, 2003; Wallman, 1993) and data-based argumentation (Krummenauer & Kuntze, 2018).

As an understanding of statistical concepts and experience with data are key prerequisites for statistical literacy (e.g., Ben-Zvi & Garfield, 2004) that begin to develop during the primary school years, this phase presents a crucial opportunity to foster foundational statistical concepts (English & Watson, 2013; Piaget & Inhelder, 1975; Watson & Moritz, 2000). In this phase, learners can even produce some first data-based arguments (Krummenauer & Kuntze, 2018) by drawing on their individual statistical literacy prerequisites and motivational dispositions related to dealing with statistical data. However, much of the existing empirical research on statistical literacy interventions has relied heavily on qualitative approaches or has suffered from methodological shortcomings, such as a lack of a control group (e.g., Paparistodemou, & Meletiou-Mavrotheris, 2008; Vahey et al., 2010).

To systematically test whether statistical literacy including data-based argumentation can already be promoted in primary school, we developed a statistical literacy intervention for third- and fourth-graders with emphasis on data-based argumentation, experience related to statistical variability, and motivation to engage in data-related tasks. The intervention is based on the predict-observe-explain approach (Gunstone & White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015). Given varying cognitive backgrounds in the elementary grades, we strategically chose highly able students from an extracurricular STEM program called the Hector Children's Academy Program (HCAP, Trautwein et al., 2023) to address their special needs (Özdemir & Işıksal Bostan, 2021) and to explore upper-performance limits. We investigated the efficacy of the newly developed statistical literacy intervention using a cluster-randomized controlled field trial.

Key Aspects of Statistical Literacy

Statistical literacy has been associated with a wide range of cognitive, motivational, and behavioral skills in the literature, for example, interpreting data, evaluating statistical claims, skepticism, and motivation (e.g., Ben-Zvi & Garfield, 2004; Gal, 2002; Wallman, 1993; Watson & Callingham, 2003). Yet, there is a lack of consensus on the definition of statistical literacy and how it should be promoted (e.g., Garfield et al., 2015). In addition, models of statistical literacy often posit subfacets, such as *mathematical knowledge*, a *critical stance* (Gal, 2002), or a *need for data* (Moore, 1997), all of which are extensive and difficult to define themselves.

Differences in the conceptions of statistical literacy notwithstanding, there is a broad consensus that students should be able to "critically evaluate statistical results" (Wallman, 1993, p. 1). This aspect of competence is key for data-based argumentation (Krummenauer & Kuntze, 2018) and enables a person to participate in social discourse and become an active citizen (Weiland, 2017). Therefore, we define statistical literacy as the ability to understand and evaluate data in order to participate in data-based argumentation. To generate data-based arguments, evidence from data has to be connected consistently with an interpretation of data, and such a process requires various aspects of statistical literacy. Most importantly, children need to understand statistical concepts relevant to the problem at hand and they need the motivation to consider data instead of relying on personal opinions. In particular, for developing proficiency in data-based argumentation, children who start learning about statistics should learn to deal with statistical variability when interpreting data (e.g., McKenzie, 2004), and their motivation to engage in data-related tasks should be promoted.

Data-Based Argumentation

Data-based argumentation is considered an important component of statistical literacy (e.g., Krummenauer & Kuntze, 2018). It entails aligning interpretations of data (e.g., "It rains less in July than in September") with empirical data (e.g., weather data), a cognitive process termed *coordination of theory and evidence* (Kuhn, 2011) according to scientific reasoning approaches (e.g., Kuhn, 2011): children need to learn to differentiate between theory (claims or interpretations of data) and evidence (here: available statistical data) in data-based argumentation. Studies have shown that children begin to understand the concept of false beliefs by the age of 4 (e.g., Wellman et al., 2001). This means that they should be able to realize that theories can be true or false and therefore should be able to learn that theory and evidence are separate.

Kuhn (2011) proposed that the next step involves learning to identify evidence as a source of knowledge. In this context, and relevant for generating arguments on the basis of statistical data, young children tend to think that knowledge is certain (Conley et al., 2004) and therefore might be confident in their interpretations without evidence. However, over time, they realize that knowledge needs to be supported by empirical evidence. Six-year-olds already tend to be better at using evidence in explanations than four-year-olds (Kuhn & Pearsall, 2000). However, even though primary school children still experience difficulties in data-based argumentation, approximately one third of children in Grades 3 and 4 (Krummenauer & Kuntze, 2018) and more than half of middle school students (Ruiz-Primo et al., 2010) are able to use data as evidence in an argument.

Previous studies have shown that primary school students' data-based argumentation can be fostered by interventions. Paparistodemou and Meletiou-Mayrotheris (2008) found that using dynamic data-visualization statistical software helped 9-year-olds improve their data-based argumentation skills, and Vahey et al. (2010) found that integrating statistics education into real-world contexts enhanced students' data-based argumentation. These studies indicate that data-based argumentation can be promoted in primary school children by providing data-related learning opportunities. However, stronger evidence of causality is needed, as only a few studies have promoted data-based argumentation, and to our knowledge, no intervention studies have applied a randomized control field trial as a study design.

Views on Statistical Variability

Dealing with statistical variability has been identified as central for statistical literacy (e.g., McKenzie, 2004; Moore, 1990; Watson & Callingham, 2003), particularly for advanced abilities in data-based argumentation (Krummenauer & Kuntze, 2018). Statistical variability encompasses phenomena related to the spread and dispersion of data values, signifying the inherent diversity within statistical data (e.g., Garfield & Ben-Zvi, 2005). However, educational curricula usually do not make statistical variability a priority in comparison with algorithmic goals in statistics (Reading & Shaughnessy, 2004).

As a result, not just adults' but also children's views of statistical variability often fail to acknowledge variability-related phenomena (e.g., Engel & Sedlmeier, 2005). Such views can lead to biased interpretations. So-called *cognitive biases* (e.g., Tversky & Kahnemann, 1971) can be connected to misconceptions about variability, such as *jumping to conclusions* (Garety et al., 2005), a bias that describes overconfident decisions that are based on insufficient data. This bias can be viewed as an underestimation or negligence of statistical variability.

Jumping to conclusions has previously been found to be related to superstitious or conspiracy-related beliefs (e.g., Kuhn et al., 2022), and early interventions have been called for (Gregersen, et al. 2022).

A small number of intervention studies have been aimed at changing misconceptions related to variability. Ben-Zvi and Sharett-Amir (2005) qualitatively analyzed how second-graders can learn to reason about distributions by engaging in exploratory data analysis activities. Ben-Zvi (2006) showed that an exploratory data visualization tool helped three classes of fifth-graders better understand variability-related phenomena and the value of large samples to improve their reasoning. However, the study did not include a control group. These findings suggest that it is possible to help students develop their views of variability through educational interventions. However, these studies need to be supplemented by systematic quantitative approaches.

Motivation

To use data in an argument, a person must be motivated to do so. Motivation has been linked to achievement in several subjects (e.g., Nasser, 2004) and has often been considered a key component of statistical literacy (e.g., Gal, 2002; Wild & Pfannkuch, 1999). According to Eccles' expectancy-value theory (Eccles & Wigfield, 2020), individuals are more likely to engage in a behavior when they feel competent and perceive sufficient value in it or its potential outcomes. Consequently, it is crucial to nurture motivational beliefs such as *self-concept*, *attainment value*, and *intrinsic value* in young learners. Developing these beliefs early on can enable them to effectively apply their statistical literacy skills. This might be even more relevant for girls as they tend to have lower motivational beliefs in math compared to boys (e.g., Lindberg et al., 2013).

Promoting early motivation for data-related tasks could be crucial for enhancing statistical literacy. Even though the exact interrelationships are not yet clear, self-concept and achievement in a domain seem to have reciprocal influences on each other (Marsh & Martin, 2011). There is evidence that cooperative learning can be used in statistics courses to help students overcome misconceptions and enhance their learning and motivation (e.g., Giraud, 1997; Jones, 1991). Research on fostering statistics motivation has predominantly centered on tertiary education (e.g., Gopal et al., 2018). This research suggests that active involvement in data handling is crucial for motivation. However, to establish whether this finding also holds true for primary school children, additional intervention studies are needed.

Given our target group of highly able primary school children, there might be challenges in enhancing motivation. Highly able children are often highly motivated (Gottfried & Gottfried, 1996). This makes it difficult to make them even more motivated. Additionally, reference group effects could reduce their motivation when moving from mixed classrooms to groups with only highly able children (Preckel et al., 2010; Zeidner & Schleyer, 1999). However, by enhancing children's statistical knowledge reciprocal positive effects on their motivation could be observable (Marsh & Martin, 2011). Therefore, it is possible, but difficult to enhance highly able children's motivation.

The Present Study

There is a need for early statistical literacy interventions (e.g., Franklin et al., 2005). The goal of the present study was to test the efficacy of an intervention that aimed to foster primary school students' data-based argumentation skills, along with their statistical literacy prerequisites, to further develop their views on statistical variability and to improve their motivation in data-related tasks. To systematically investigate the efficacy of the intervention, we implemented a randomized controlled trial (RCT) with a waitlist control group.

We had the following preregistered hypotheses:

In the first hypothesis, we expected positive effects of the intervention on students' statistical literacy and more specifically on their data-based argumentation and understanding of variability. Prior evidence indicates that active involvement in data exploration can positively influence children's argumentation skills (Paparistodemou & Meletiou-Mayrotheris, 2008; Vahey et al, 2010) and their views on statistical variability (Ben-Zvi, 2006; Ben-Zvi and Sharett-Amir, 2005)

In the second hypothesis, we expected that there would be no negative effects of the intervention on students' motivational beliefs about data-related tasks. Because children in previous studies within the same enrichment program were already highly motivated at baseline (e.g., Schiefer et al., 2021) and reference group effects (Preckel et al., 2010; Zeidner & Schleyer, 1999) might counteract the positive effects of the intervention, we assumed that an increase in motivation was possible but might be difficult to observe.

In the third hypothesis, our aim was to explore differential intervention effects. This investigation sought to ensure that the intervention could meet the diverse needs of learners, address disparities in learning outcomes, and identify any unintended consequences that might disproportionately affect specific groups of learners. Ultimately, the goal is to promote

equitable and effective educational practices. Therefore, we explored differential effects of prior knowledge, fluid intelligence, gender, and grade level.

Method

Transparency and Openness

We adhere to the Journal Article Reporting Standards (JARS; Kazak, 2018). All data and analysis codes for the main analyses are available on the Open Science Framework (OSF; https://osf.io/p4t53). Additional materials are available upon request. We preregistered the hypotheses and analysis and the analysis plan of this study on the *Registry of Efficacy and Effectiveness Studies* (REES; https://sreereg.icpsr.umich.edu/sreereg/subEntry/21020/pdf?action=view). We have largely adhered to the preregistration and point out below where and why we had to deviate from the preregistration. The study protocol was approved by the Ethics Committee of the Faculty of Economics and Social Sciences at the University of Tübingen, granting ethical clearance for the research (Approval No. A2.5.4-191_ns).

Procedure and Participants

The intervention was implemented at the HCAP, an extracurricular enrichment program in the German federal state of Baden-Württemberg, in the first term of the 2021/2022 school year while COVID-19 pandemic restrictions were still in place. The HCAP is hosted at 68 different local sites and is tailored to talented, interested, and motivated primary school children (for more information about the HCAP, see Trautwein et al., 2023). Children are nominated by their teachers for the program. The nominated children can choose from a variety of courses with a focus on Science, Technology, Engineering, and Mathematics (STEM) subjects.

The intervention was offered as a course called “Luck or genius? Understanding data and making predictions” in the course programs of five local sites of the HCAP. An announcement provided information about the intervention’s contents and goals as well as some initial information about the study. The intervention was open to all third- and fourth-graders nominated for the program with written parental consent to participate in the study. For each site, up to 20 children could be admitted to the course.

A total of 58 third- and fourth-graders enrolled in the course, and 53 of them participated in the study (79.2% boys, M_{age} = 9.10, SD_{age} = 0.67, see Table 1). To evaluate the efficacy of the intervention, we used a repeated-measures (pre-post measurement) RCT

(Friedman et al., 2010). The pretest measurement took place 1 week before the intervention began (see Figure 1).

Table 1

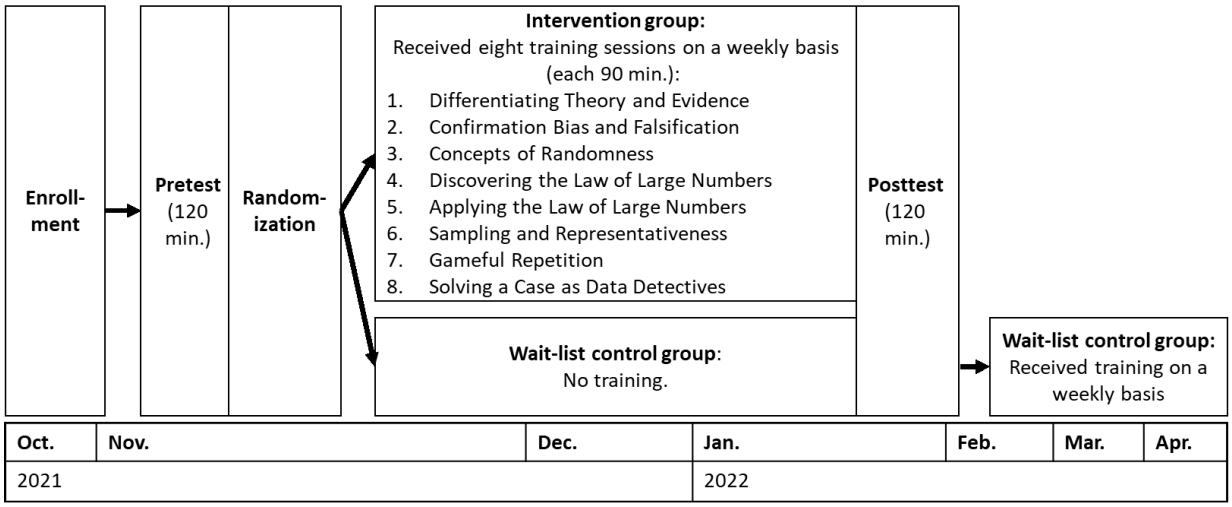
Description of the Sample

Group	*N*	Boys	Age	Grade 3	Grade 4
Intervention Group	25	72%	*M* = 9.13 (*SD* = .79)	13	12
Waitlist Control Group	28	86%	*M* = 9.07 (*SD* = 56)	16	12

Note. *N* = number of participating children; *M* = mean; *SD* = standard deviation.

Figure 1

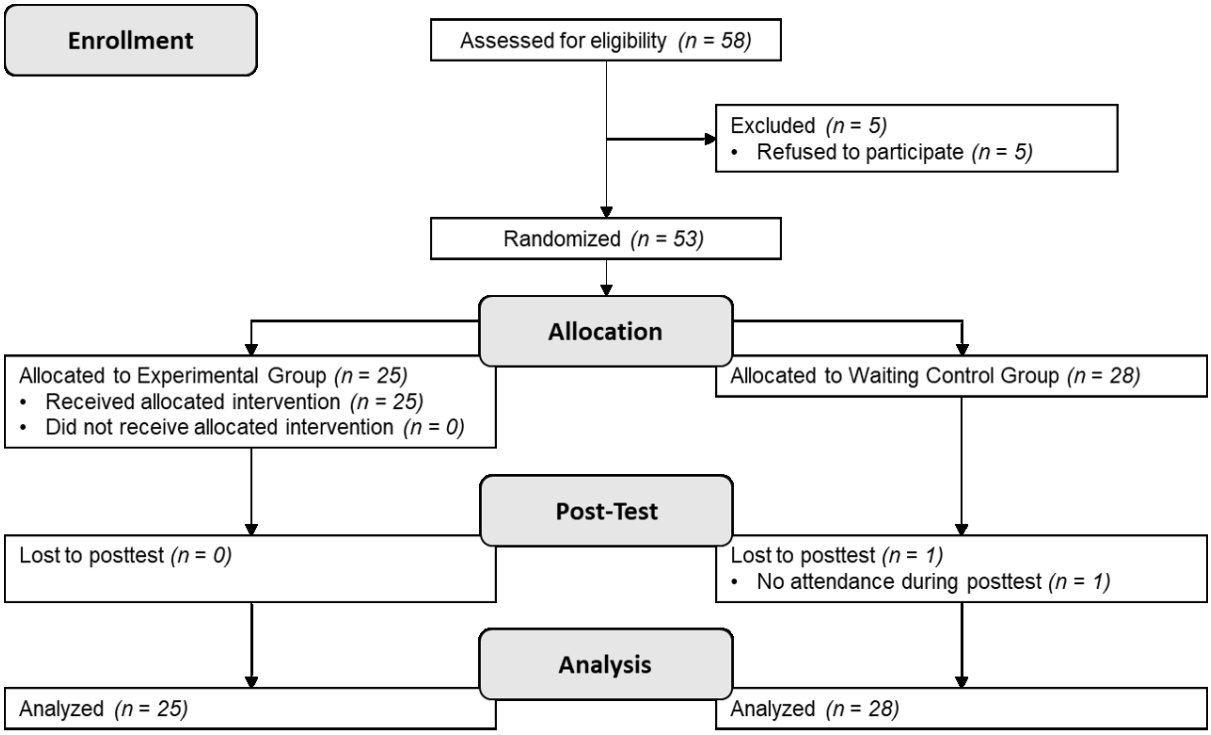
Study Procedure and the Contents of the Intervention



After the pretest, the children were randomly assigned to one of the two conditions. Children in the intervention condition (n = 25) received weekly training for a total of eight sessions, each 90 min long. Children in the waitlist control group (n = 28) received the training after the posttest. It was possible for the children from the intervention condition and the waitlist control condition to also participate in other courses. At one of the sites, there were not enough children to form groups in both conditions, so they were all randomly assigned to one group (waitlist control group). After the eight sessions were taught in the intervention group, the children from both groups participated in the posttest. Additionally, parents were asked to fill out a questionnaire on family background and to provide additional information about the children. Attrition data is displayed in Figure 2.

Figure 2

Attrition Flow Chart



Intervention: Training in Statistical Literacy

The intervention ("Luck or genius? Describing data and making predictions") consisted of eight weekly 90-min course sessions and was designed to teach basic statistical concepts like randomness, variability, and representativeness (see Fig. 1). It was based on two didactical core components (Fixsen et al., 2009): the predict-observe-explain approach (Gunstone & White, 1981) and cooperative learning (e.g., Capar & Tarim, 2015), both of which were expected to make the intervention effective.

The predict-observe-explain approach (Gunstone & White, 1981) is a three-step approach that can be linked to the coordination of theory and evidence (Kuhn, 2011) and is similar to structured inquiry learning (Banchi & Bell, 2008). In the predict phase, children state their theories (e.g., "Most of the children in this course play a musical instrument"). In the observe phase, evidence is gathered (e.g., all course participants complete a survey). Eventually, in the explain phase, children need to coordinate theory and evidence by asking themselves whether their previous prediction was correct or not and how any discrepancies could be explained (e.g., "My prediction was not correct, but I had based my prediction on the fact that I play a musical instrument myself"). Through this process, the children can see that theory and

evidence are not equivalent and that evidence can contradict or support a theory in data-based argumentation.

Cooperative learning methods include collaborative tasks in which individuals interact to enhance their understanding and learning outcomes (e.g., Capar & Tarim, 2015). This intervention provides many opportunities for the children to interact and exchange arguments in a playful way. Children can therefore exchange ideas in their own age-appropriate language and help each other to learn new concepts. This social exchange enhances motivation and performance (Capar & Tarim, 2015; Gürdoğan-Bayır & Bozkurt, 2018).

The intervention was designed so that all genders would be equally interested in and knowledgeable about the contents. All sessions included two main characters (Tim and Tamara), who were designed as children with which the participants could identify, and served as a framework for the stories inside the intervention. Each session was built on the same general blueprint:

A unit (except for the first) usually started with the repetition of word memory cards from the previous sessions, which were important for the current session. These memory word cards were used to store the children's knowledge visually and make the learning gains obvious. For example, the concept of randomness was captured as following: "Something is random if you cannot know or influence beforehand what the result will be."

Then, students would be participating in an activity based on the predict-observe-explain approach. In the first session, students secretly make predictions about characteristics of the whole class (e.g. the most liked school subject, the most mentioned favorite color, etc.) as part of a get-to-know game. Then they share information about themselves in front of the class while sorting their data points into bar graphs. When all students revealed information about themselves, the graphs were talked about in an instructor-led discussion. Children would then say what they secretly predicted and why, and if necessary revised their prediction. As part of the discussion, children were often stating underlying theories, which could not be supported by their data (e.g. "This is a math course, so the most liked school subject should be mathematics.", or "Boys like the color blue. Because most of us are boys, I thought the most liked color would be blue."). Instructors were encouraged to tone down such theories. If the data did not suit them, this was easy to do. If the data did fit, the instructors pointed out that the result did fit, but that this did not necessarily mean that the theory was correct. Alternative explanations were addressed.

The predict-observe-explain phases were designed so that children were likely to make predictions that did not match with the results observed afterwards to make them aware of their

misconceptions, and then come to a new mutual understanding through playful experiences, social exchange, and discussion. For example, in course unit 5, every child had to make a prediction about which out of two paper plane models can fly further. Usually, children expected the model with a sharp tip to fly further than the model with a flat tip. Then, in pairs they gathered data by throwing the paper planes multiple times and making notes of the distance. During this activity, children got to discuss their predictions with their partners. Then, they reflected on the results in a group discussion, led by the course instructor. In this case, the children learned that their assumption that the paper plane with a sharp tip would fly further than the one with a flat tip was false. Also, they internalized that the more data they gathered, the clearer the result, especially when the results of all children were aggregated. Whereas the predict-observe-explain approach was assumed to be the main trigger for such learning experiences, cooperative learning was assumed to be its catalyst.

After these predict-observe-explain phases, usually children deepened their knowledge about the present content of the session. For example, in the first session, children were playing a game of data dominoes, were they would match displayed evidence with the corresponding theory. In other sessions, there were also similar activities such as playing guessing games or discussing exemplary statements about presented or gathered data.

Most sessions concluded with children completing a "My Learning Progress" worksheet. This document revisited the content of the session's new word memory cards and prompted children to identify real-life instances related to the material. For example, in the session centered on the concept of randomness, children were prompted to recognize instances of randomness in their daily lives.

Measures

We assessed the three key aspects of statistical literacy mentioned above (data-based argumentation, dealing with and developing views related to statistical variability, and motivation in data-related tasks). We aimed to assess these aspects with several measurement instruments (see Table 2).

Data-Based Argumentation

We used 14 items with increasing difficulty based on tasks as described in Krummenauer and Kuntze (2018) to assess data-based argumentation. For each item, children had to evaluate whether a statement was true and to justify their decision on the basis of the available data. For

Table 2

Descriptive Statistics of Dependent Variables and Fluid Intelligence: Means, Standard Deviations, Reliabilities, and Number of Items

Construct			T1				T2			
	N items	Group	*N*	*M*	*SD*	α/*KR-20*	*N*	*M*	*SD*	α/*KR-20*
Data-based argumentation*	14	All	51	2.78	2.14	.70	52	3.85	2.30	.68
		IG	25	2.96	2.24		25	4.80	2.18	
		CG	26	2.62	2.06		27	2.96	2.07	
Views on variability	5	All	49	2.05	0.45	.76	51	2.21	0.51	.83
		IG	24	2.09	0.42		25	2.48	0.36	
		CG	25	2.01	0.49		26	1.95	0.51	
Draws to decision	2	All	51	10.00	5.66	.79	52	12.19	5.32	.81
		IG	25	11.00	5.89		25	14.94	4.39	
		CG	26	9.04	5.36		27	9.65	4.87	
Self-concept	6	All	49	3.13	0.62	.86	51	3.25	0.63	.92
		IG	24	3.17	0.69		25	3.44	0.61	
		CG	25	3.09	0.55		26	3.07	0.60	
Intrinsic value	6	All	49	3.01	0.68	.90	52	3.18	0.75	.94
		IG	25	2.96	0.70		25	3.11	0.82	
		CG	24	3.06	0.66		27	3.24	0.67	
Attainment value	3	All	48	3.35	0.61	.72	52	3.44	0.55	.66
		IG	23	3.25	0.67		25	3.44	0.45	
		CG	25	3.44	0.55		27	3.44	0.63	
Fluid Intelligence*	16	All	51	9.98	3.25	.82				
		IG	25	10.16	2.91					
		CG	26	9.81	3.59					

Note. *N* = number of participating children; *M* = mean; *SD* = standard deviation; IG = intervention group; CG = control group.

* For instruments with binary items the Kuder-Richardson coefficient was calculated instead of Cronbach's Alpha.

Measurement points: T1 = November 2021, T2 = January to February 2022.

the easiest items, students had to name only one data point to justify their evaluation. For example, children had to read off a value from a graph to tell whether less than 120.000 cinema tickets had been sold in a certain year. For more difficult items, test takers had to identify multiple data points, variability, and differences in samples, if any, to solve the task correctly. For example, children had to use multiple data points from product reviews to decide whether a certain bike is the best to buy from a small selection. Answers were rated by three raters (0 = *false*, 1 = *correct*). The rating that had the highest level of agreement among the raters was selected as the final rating for each item. Then a sum score was calculated. To check for interrater reliability, we calculated Fleiss' Kappa (Fleiss & Cohen, 1973), which confirmed good reliability (K = .863).

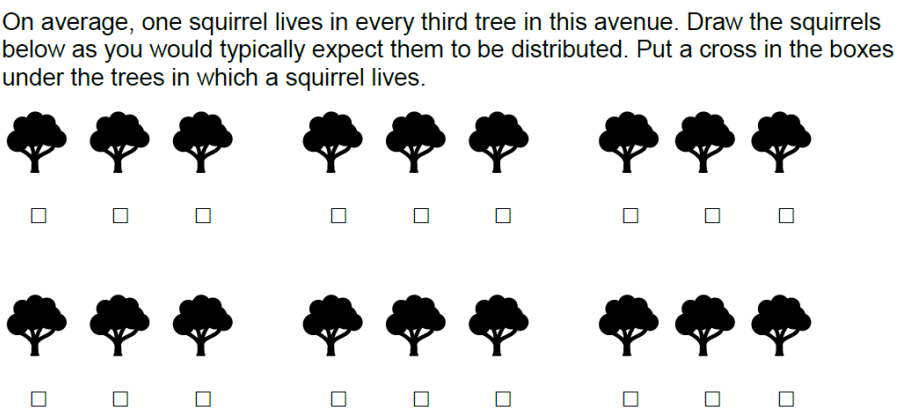
The internal consistency of the scale was assessed as sufficient by applying the Kuder-Richardson coefficient for binary variables (*KR*-20; Kuder & Richardson, 1937, $KR\text{-}20_{T1}$ = .70, $KR\text{-}20_{T2}$ = .68) and computing EAP reliability (Andrich, 1982; EAP_{T1} = .71, EAP_{T2} = .69).

Dealing With Variability

As a first indicator of whether the students developed an understanding of variability, we assessed what the children would expect variability to look like in distributions. In five items, they were asked to generate distributions on the basis of prompts, such as "On one street, a squirrel lives in every third tree on average." Children were requested to tick the boxes to create a distribution that looked typical to them (see Fig. 3). The children's answers were then sorted into three levels; (1) *no variability* (i.e., one squirrel in every third tree), (2) *little variability* (i.e., one squirrel in every group of three trees), and (3) *full variability* (i.e., approximately one squirrel per three trees on average, but the squirrels show no regularity). Then a mean value for their views on variability was created (α_{T1} = .76, α_{T2} = .83).

Figure 3

Sample Item for Views on Variability



As a second indicator of the children's understanding of variability, the jumping to conclusions bias was assessed with the 85:15 and 60:40 versions of the beads task (Dudley, John, Young, & Over, 1997a, 1997b; Garety et al., 2005). In the first of two items, children saw two different jars, one with 85 yellow beads and 15 black beads and one with 15 yellow beads and 85 black beads. In the second item, the ratios were 60 to 40 and 40 to 60 respectively. They were then told that one of the two jars would be chosen at random, and beads would be drawn from this jar in an order that had been determined beforehand. After each draw, the children could decide whether they wanted to see another bead or if they already knew which jar the beads were drawn from. All drawn beads were visible at any time to aid memory. Draws to decision were then used as an outcome measure with fewer draws suggesting a greater jumping to conclusions bias. We created a mean value for each participant. Because there were only two items, we calculated the correlation between the items to judge the scale's reliability as sufficient (r_{T1} = .66, r_{T2} = .70; α_{T1} = .79, α_{T2} = .81).

Motivation

We assessed three domain-specific motivational variables. First, self-concept was measured with six adapted items (α_{T1} = .86, α_{T2} = .92; based on Arens et al, 2011, and Gaspard et al., 2015; e.g., "Everything that has to do with data comes easy to me"). Second, intrinsic value was assessed with six items (α_{T1} = .90, α_{T2} = .94; based on Stalder 2013, e.g., "I like to do everything that has to do with data"). Third, attainment value was measured with three adapted items (α_{T1} = .72, α_{T2} = .68; Ramm et al., 2013, e.g., "Everything that has to do with data is important to me"). All items were rated on a scale ranging from 1 (*not true at all*) to 4 (*exactly right*). Mean values were calculated for each participant.

Fluid Intelligence

Fluid intelligence was measured as a covariate at pretest with 16 figural items from an age-adapted version of the BEFKI (Berlin Test of Fluid and Crystallized Intelligence; see, e.g., Schroeders et al., 2020), appropriate for primary school children (Schroeders et al., 2016). In each item, children needed to select one of three possible figures for every fourth and fifth unit within a sequence to complete it. Time was limited to 15 min. Answers were coded as binary variables (0 = *false*, 1 = *correct*), and a sum score was calculated. Internal consistency was high (*KR*-20_{T1} = .82).

Excluded Variables

Due to unexpected disruptions during the COVID-19 pandemic, three measures (epistemic beliefs, confirmation bias, anthropomorphism) could not be completed by participants at some sites, resulting in a large number of missing values at pretest (43%–57% NA). Additionally, internal consistencies and confirmatory factor analyses indicated insufficient reliabilities. As such, these measures were excluded from the analysis. All analyses were still calculated and are reported in the Appendix.

Analyses

The analysis plan was preregistered. Cases in which we had to deviate from the preregistration are mentioned below. Intention-To-Treat (ITT) effects, which are considered a "strict test" (Fisher et al., 1990) of effects, were estimated for each outcome variable separately to identify treatment effects. By including all participants, regardless of whether they complied with the treatment or dropped out, the ITT analysis can account for potential biases that can arise when analyzing only those who completed and complied with the treatment. The pretest score was included in each model to increase power (e.g., Aiken et al., 2003).

To check for any differences between the intervention and control groups, we preregistered that we would check for nonsignificant baseline differences to assess baseline equivalence of all pretest values of dependent variable, gender, grade level, and fluid intelligence. However, to be in accordance with the guidelines of the What Works Clearinghouse (2022), if there were any mean value differences between the two groups on the pretest with an effect size $d > 0.05$ (even if this difference was not statistically significant) despite randomization, these variables had to be included as covariates even when nonsignificant. In line with their recommendations for measures of baseline equivalence, we used *Hedges' g* (Hedges, 1981) for continuous variables and the Cox index (Cox, 1970) for binary variables. The treatment variable was dummy-coded (1 = *intervention*, 0 = *waitlist control group*).

Due to the standardization of the dependent and predictor variables, the regression coefficients can be interpreted as Cohen's *d* (Cohen, 1988). To assess differential effects (e.g., for children's pretest scores), the cross-product of the corresponding variables with the intervention condition was included as an additional predictor variable in additional models.

Missing values ranged from 1.89% to 9.43% on the primary variables. To deal with missing data due to students' absence at pretest or posttest or due to nonresponse for individual scales, we used the full information maximum likelihood approach (Enders, 2001). To account

for the nesting of students in local sites, multiple linear regressions were computed with a design-based correction of standard errors (Asparouhov & Muthén, 2006). We used the *lavaan* (Rosseel, 2012) package to estimate the regression models. Contrary to the preregistration, we did not use the package lavaan.survey (Oberski, 2014) because the standard errors were already adjusted due to the cluster argument in lavaan. For directed hypothesis (1), one-tailed tests were applied. For undirected hypotheses (2 and 3), we applied two-tailed tests. Finally, we used the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to control the false discovery rate due to multiple testing within each research question. All analyses were conducted in R (R Core Team, 2019) with a significance level of $p < .05$.

Results

In a preliminary analysis, we checked for baseline equivalence in accordance with the What Works Clearinghouse (2022) guidelines. All covariates had nonsignificant baseline differences greater than an effect size of $d = 0.05$ (see Table 3). Using stricter guidelines for baseline differences (What Works Clearinghouse, 2022) than the preregistered requirement of only nonsignificant results, we identified three variables (gender, draws to decision, and attainment value of data-related tasks) with baseline differences above $d = 0.25$, indicating that baseline equivalence was not satisfied. To address this issue, we included all pretest variables as covariates in the regression analyses to control for whether any of the posttest values could be better explained by these baseline differences in contrast to any effects of the intervention. Since grade level and age are closely intertwined, we exclusively employed grade level as a predictor, as it appeared better suited to represent knowledge acquired throughout one's educational journey.

We investigated the intercorrelations among all variables to understand their relationships and identify potential underlying patterns (see Table 4). All dependent variables showed a significant correlation between pretest and posttest, speaking for the reliability and consistency of the measurements over time. The three statistical literacy measures (data-based argumentation, views on variability, and draws to decision) were more highly correlated at posttest than at pretest, which speaks for the increased coherence and integration of these skills following the intervention. All motivational variables (self-concept, intrinsic value, and attainment value) showed significant intercorrelations at pre and posttest. This is plausible as these measures are closely related. Full intercorrelations with confidence intervals of included and excluded variables are reported in Table A4 in the Appendix.

Table 3

Baseline Differences

Variable	Baseline Differences			
	g	*t*	*df*	*p*
Age	0.08	-0.29	42.54	.770
Gender (1 = girls)	0.34			
Grade Level (1 = grade level 4)	0.10			
Data-based Argumentation	0.16	-0.57	48.24	.571
Views on Variability	0.16	-0.57	46.41	.570
Jumping to Conclusions	0.35	-1.24	48.12	.220
Self-concept	0.12	-0.43	43.95	.672
Intrinsic Value	-0.15	0.52	46.98	.605
Attainment Value	-0.32	1.09	42.76	.282
Fluid Intelligence	0.11	-0.39	47.67	.701

Note. Instead of Hedges' g, the Cox index is displayed for binary variables.

Table 4

Correlations Between the Dependent Variables and Covariates on Pretest (Below the Diagonal) and on Posttest (Above the Diagonal)

Variable	1	2	3	4	5	6	7	8	9	10
1. Data-based Argumentation	.68***	.35*	.24	.13	.06	-.01	.08	.20	.35*	.37**
2. Views on Variability	.00	.61***	.37**	.08	-.09	-.00	.08	.08	.11	.09
3. Draws to Decision	-.06	.16	.44**	.03	-.08	-.05	.02	.07	-.09	-.02
4. Self-concept	.09	.13	-.12	.34*	.39**	.52***	-.14	-.11	.24	.26
5. Intrinsic Value	.21	.23	.12	.43**	.51***	.45***	.07	.00	.08	.07
6. Attainment Value	-.11	-.11	-.33*	.40**	.32*	.29*	.22	.09	.10	.16
7. Fluid Intelligence	.23	-.05	-.02	.22	.21	-.21				
8. Gender (1 = girls)	.12	-.04	.00	-.12	-.02	-.10	-.30			
9. Grade Level (1= fourth)	.33*	.26	.11	.29*	.25	.12	.17	.10		
10. Age	.34*	.14	.00	.30*	.22	.18	.04	.13	.84***	

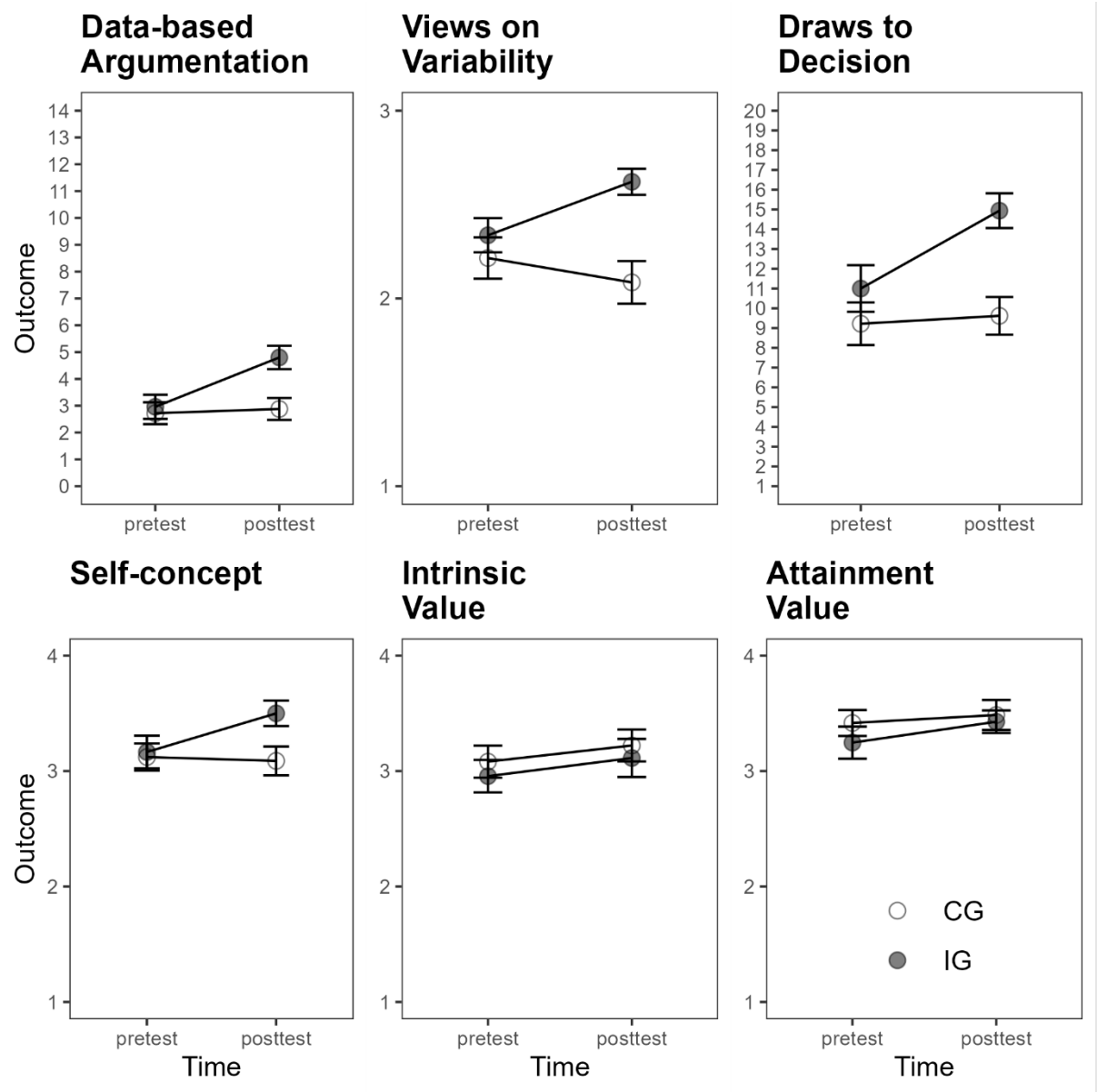
Note. The diagonal displays the correlations between pre-test and post-test. Fluid intelligence was only assessed on pre-test. Correlations between fluid intelligence, gender, grade level and age are only reported once because the correlations don't vary between pre-test and post-test. *$p < .05$., **$p < .01$., ***$p < .001$.

Hypothesis 1 stated that there were positive effects of the intervention on chosen aspects of students' statistical literacy. Figure 4 shows the mean values on the pretest and posttest for all dependent variables, indicating intervention effects on all three variables (results on excluded

variables are included in the Appendix). Tables 5 and 6 present the results of the regression analysis for all target variables. The full tables with all predictors and additional analyses are in the Appendix. Statistically significant intervention effects on the posttest values were found for data-based argumentation ($B = 0.72$, 95% CI [0.61, 0.83], $p < 001$), views on variability ($B = 0.81$, 95% CI [0.52, 1.10], $p < .001$) and draws to decision ($B = 0.80$, 95% CI [0.51, 1.09], $p < .001$).

Figure 4

Descriptive Mean Values With Standard Errors for the Dependent Variables at Pretest and Posttest



Note. CG = Control Group; IG = Intervention Group. This graphic was produced by using the ggplot2 package (Wickham, 2016)

Table 5

Intervention Effects on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention	0.72***	0.06	[0.61, 0.83]	<.001	<.001	0.81***	0.15	[0.52, 1.10]	<.001	<.001	0.80***	0.15	[0.51, 1.09]	<.001	<.001
Pretest Value	0.62***	0.10	[0.42, 0.82]	<.001		0.65***	0.10	[0.44, 0.85]	<.001		0.36*	0.15	[0.07, 0.65	<.001	
R^2	.66					.58					.54				

Note. For the pretest score and the intervention, one-tailed significance levels are reported because directional hypotheses were tested. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure. The complete table with the effects of baseline differences is reported in the Appendix.

* p < .05. ** p < .01. *** p < .001.

Table 6

Intervention Effects on Motivation for Data-related Tasks Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention	0.68***	0.14	[0.41, 0.95]	<.001	<.001	0.02	0.20	[-0.37, 0.42]	.906	.906	-0.02	0.17	[-0.35, 0.31]	.900	.906
Pretest Value	0.12	0.22	[-0.31, 0.55]	.593		0.77***	0.09	[0.60, 0.95]	<.001		0.47*	0.19	[0.11, 0.83]	.011	
R^2	0.32					0.45					0.20				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure. The complete table with the effects of baseline differences is reported in the Appendix.

* p < .05. ** p < .01. *** p < .001.

Hypothesis 2 postulated that there were no negative effects of the intervention on students' motivational beliefs concerning data-related tasks. The mean values in Figure 2 suggest an intervention effect only on self-concept. The results of the regression analysis (Table 6) indeed showed a positive and significant course effect on self-concept in data-related tasks (B = 0.68, 95% CI [0.41, 0.95], p < .001) but not on intrinsic value (B = 0.02, 95% CI [-0.37, 0.42], p = .906) and attainment value (B = -0.02, 95% CI [-0.35, 0.31], p = .900).

Hypothesis 3 explored whether there were differential intervention effects based on previous knowledge, gender, grade level, and fluid intelligence (see Tables A13 to A28 in the Appendix). We found three significant interaction effects out of all 24 models. Children in grade level 4 had larger gains in their views on variability (B = 0.80, 95% CI [0.37; 1.22], p < .001) and their self-concept (B = 0.95, 95% CI [0.52; 1.38], p < .001) than children in grade level 3. Children with higher fluid intelligence had larger gains in their data-based argumentation (B = 0.39, 95% CI [0.14; 0.63], p = .016).

Discussion

In this study, we examined the efficacy of an 8-week statistical literacy intervention called "Luck or genius? Understanding data and making predictions" in enhancing the statistical literacy of highly able primary school children in an extracurricular enrichment program. The intervention focused on improving third- and fourth-graders' data-based argumentation, understanding of variability, and motivation when engaging in data-related tasks. The efficacy of the intervention was assessed with a randomized-controlled field trial involving 53 children.

Overall, the results indicated the efficacy of the intervention. Evidence showed that the intervention significantly enhanced students' data-based argumentation skills, views on variability, draws to decision, and self-concept in data-related tasks, but not intrinsic and attainment value in data-related tasks. The analysis of the differential effects suggested that there might be greater benefits on data-based argumentation for children with higher fluid intelligence and greater benefits on views on variability and self-concept for fourth-graders. However, these differential effects should be interpreted with caution because of the relatively small sample size.

Effects of the Intervention on Aspects of Statistical Literacy

The study benefits the research field by conducting a statistical literacy intervention in a randomized-controlled field trial to provide robust evidence for the malleability of statistical

literacy in primary school children. Intervention studies of this kind are rather scarce. Additionally, it offers insights into how statistical literacy can be promoted.

Kuhn (2011) suggested that children must first distinguish between theory and evidence, and then employ evidence as a basis for scientific thinking. In our intervention, the predict-observe-explain approach (Gunstone & White, 1981) labels theory and evidence as prediction and observation, fostering awareness of their distinctness. As expected, the intervention had a significant medium-sized positive effect on data-based argumentation. This finding was anticipated, given the intervention's focus on predict-observe-explain activities. It implies that primary school children can improve their data-based argumentation and suggests the effectiveness of distinguishing between these constructs, a topic that warrants deeper exploration. Future studies may shed more light on how the distinct labeling of theory and evidence can help enhance data-based argumentation by comparing an intervention with the predict-observe-explain approach with an intervention without it.

Given variability's pivotal role in statistics (e.g., McKenzie, 2004), we anticipated that the intervention would shift children's expectations toward more variability due to real data exposure. In line with our hypothesis, a notable effect on children's views on variability was found. Also, the intervention significantly bolstered students' draws to decision. These findings align with our expectations, as the intervention introduced the concept of randomness and the law of large numbers, aiding students' abilities to deal with and describe phenomena related to statistical variability. So, the children seemed to learn about variability and that larger samples yield less variable and thus more accurate outcomes.

Finally, as anticipated, children's motivation was not reduced relative to the control group. The intervention did not significantly alter children’s intrinsic and attainment values regarding data-related tasks. In contrast, a significant positive effect on students' self-concept was found. These findings suggest that despite initially high scores on pretests, children in the intervention group gained confidence in completing the tasks. The predict-observe-explain approach (Gunstone & White, 1981) might have contributed to these findings, as predictions that were incongruent with the data likely prompted the students to improve their interpretations. Cooperative learning (e.g., Capar & Tarim, 2015), linked to motivation (e.g., Giraud, 1997; Jones, 1991), was incorporated, featuring engaging games and experiments. Additionally, at the end of each session, students reflected on what they had learned, further enhancing their self-concepts.

Practical Implications

Given the demonstrated effectiveness of the statistical literacy intervention, we can derive some practical implications. First, based on our defined core components, we can assume that the predict-observe-explain approach (Gunstone & White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015) might be effective in promoting statistical literacy. We assume that making predictions and comparing them to gathered data within a context of cooperative social exchange drives a deeper understanding of statistical concepts. Therefore, these or similar approaches might be used by teachers and educators to promote statistical literacy in other contexts.

Second, we found that our intervention was successfully implemented in a sample of highly able primary school children from an extracurricular enrichment program. This serves as a positive example of evidence-based interventions that are needed for talent development (Subotnik et al., 2011). Also, on the one hand, because of our highly able target group we expected less variable and larger effects than in the general population, as our sample was highly motivated and able to learn more complex topics. But on the other hand, our results are less generalizable. This view is enhanced by the fact that almost 80% of the sample were boys. To check whether specific children had greater benefits from the intervention than others, we conducted multiple regression analyses and computed interactions between the intervention condition and gender disparities, pretest scores, grade level, and fluid intelligence. The intervention did not show significant differential effects on gender disparities, or previous knowledge. However, a significant interaction between the intervention and fluid intelligence was found on data-based argumentation, which indicates a more beneficial effect of the intervention on more intelligent children. Additionally, significant interactions between the intervention and grade level was found on draws to decision and self-concept in data-related tasks, which indicated a more beneficial effect of the intervention on fourth-graders. Altogether, only three out of 24 differential effects were significant. Because there were so little differential effects, the intervention may benefit a broader range of children. However, with only 53 participating children, our sample size was relatively small. Therefore, this study lacks statistical power to detect effects, especially interaction effects.

Third, this study provides evidence that the promotion of constructs in the context of extracurricular STEM (Science, Technology, Engineering, and Mathematics) education is possible. This study adds to the research body of similar studies from the same enrichment program with positive effects (e.g., Rebholz & Golle, 2017; Herbein et al., 2018; Schiefer et

al., 2020). This first evidence of efficacy under rather standardized conditions should be followed up with an effectiveness study with conditions that are closer to the real implementation of the intervention (e.g., course instructors from field instead of university staff; e.g., Carroll et al., 2007, Greenberg et al., 2005). This way, greater evidence for the real-world effectiveness of the intervention could be found. In a final step, the intervention should be scaled up to a more diverse target group with a larger sample, to test whether the intervention is also effective for a broader group of children (Herbein et al., 2018; Gottfredson et al., 2015; Humphrey et al., 2016).

Strengths, Limitations, and Outlook

One of the main strengths of this study lies in its adoption of a randomized controlled field trial, which is widely recognized as the gold standard for establishing causality in psychological and educational interventions (Lendrum & Wigelsworth, 2013; Torgerson & Torgerson, 2013). This rigorous experimental design provides robust evidence of the efficacy of the intervention in enhancing statistical literacy in the selected population. To further enhance the design, further studies should use larger samples than the relatively small sample of 53 children. Also, by having the posttest 1 week after the intervention, we detected short-term effects of the intervention. However, this assessment did not provide insights into the long-term durability of the intervention's effects. To address this limitation, future studies could adopt a longitudinal approach with a longer time frame to examine how long the improvements in statistical literacy skills persist over time.

Another strength of this study is the evidence-based conceptualization of the intervention. The study's core components, such as the predict-observe-explain approach (Gunstone and White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015), have shown promise in enhancing statistical literacy. By applying this evidence-based conceptualization of the intervention, we wanted to add to this body of literature. However, the experimental design used in this study did not allow us to conclusively demonstrate that these specific components are responsible for the observed effects. Future research could address this limitation by comparing the efficacy of such interventions that employ these core components against similar interventions that do not.

A final strength of this study lies in the high impact of our designed intervention. As preregistered, we interpreted our effect sizes in accordance with Cohen's *d* (Cohen, 1988). We found significant effect sizes between $d = 0.69$ and $d = 0.81$. According to Cohen, these effect sizes can be interpreted as medium to large. However, Kraft (2020) analyzed and compared the

effect sizes specifically of educational interventions. Based on that he proposes that effect sizes above 0.20 are already large in an educational context. Therefore, the intervention effect sizes of this study can be interpreted as exceptionally large.

Conclusions

In conclusion, our study offers valuable evidence of the positive effects of a statistical literacy intervention on primary school children's data-based argumentation, views on variability, draws to decision, and self-concept in data-related tasks; intrinsic value and attainment value remained unchanged. This evidence-based intervention serves as a first step in exploring the connections between the predict-observe-explain approach, cooperative learning methods, and statistical literacy. Future studies should aim to use larger samples, more diverse participants, and longitudinal assessments to deepen our understanding of the impact of such interventions.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process.

During the preparation of this work the authors used ChatGPT (OpenAI, 2024) in order to improve style and grammar of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Aiken, L. S., West, S. G., & Pitts, S. C. (2003). *Multiple linear regression*. In Handbook of Psychology (pp. 481–507). John Wiley & Sons, Inc. https://doi.org/10.1051/eas/1466005

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives*, *9*(1), 95-104. https://rasch.org/erp7.htm

Arens, A. K., Trautwein, U., & Hasselhorn, M. (2011). Erfassung des Selbstkonzepts im mittleren Kindesalter: Validierung einer deutschen Version des SDQ I [Assessment of self-concept in middle childhood: validation of a German version of the SDQ I]. *Zeitschrift für Pädagogische Psychologie*. https://doi.org/10.1024/1010-0652/a000030

Asparouhov, T., & Muthén, B. O. (2006). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meeting*, 2718–2726. https://doi.org/10.1111/j.1467-985X.2006.00426.x

Banchi, H. & Bell, R. (2008). The many levels of inquiry. *Science and Children*, *46*(2), 26-29. https://www.michiganseagrant.org/lessons/wp-content/uploads/sites/3/2019/04/The-Many-Levels-of-Inquiry-NSTA-article.pdf

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman, & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics* [CD-ROM]. Voorburg, The Netherlands: International Association for Statistics Education. https://www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/2D1_BENZ.pdf

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004). The challenge of developing statistical literacy, reasoning and thinking (pp. 3-16). Dordrecht: Kluwer academic publishers. http://dx.doi.org/10.1007/1-4020-2278-6

Ben-Zvi, D., & Sharett-Amir, Y. (2005). *How do primary school students begin to reason about distributions?* In K. Makar (Ed.), Proceedings of SRTL-4. University of Queensland. https://www.academia.edu/976792

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Capar, G., & Tarim, K. (2015). Efficacy of the cooperative learning method on mathematics achievement and attitude: A meta-analysis research. *Educational Sciences: Theory and Practice*, *15*(2), 553-559. http://dx.doi.org/102738/estp.2015.2.2098

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implement Science*, *2*(1), 40-49. https://doi.org/10.1186/1748-5908-2-40

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Edition). Hillsdale, NJ: Lawrence Earlbaum Associates. https://www.utstat.toronto.edu/brunner/oldclass/378f16/readings/CohenPower.pdf

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, *29*(2), 186-204. https://doi.org/10.1016/j.cedpsych.2004.01.004

Cox, D. R. (1970). Analysis of binary data. New York: Chapman & Hall/CRC.

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997a). The effect of self-referent material on the reasoning of people with delusions. *British Journal of Clinical Psychology, 36,* 575–584. https://doi.org/10.1111/j.2044-8260.1997.tb01262.x

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997b). Normal and abnormal reasoning in people with delusions. *British Journal of Clinical Psychology, 36,* 243–258. https://doi.org/10.1111/j.2044-8260.1997.tb01410.x

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. Contemporary educational psychology, 61, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61*(5), 713–740. https://doi.org/10.1177/00131640121971482

Engel, J., & Sedlmeier, P. (2005). On middle-school students' comprehension of randomness and chance variability in data. ZDM, 37, 168-177. http://dx.doi.org/10.1007/s11858-005-0006-4

English, L., & Watson, J. (2013). Beginning inference in fourth grade: Exploring variation in measurement. In V. Steinle, L. Ball, & C. Bardini (Eds.), *Mathematics education: Yesterday, today and tomorrow* (pp. 274–281). Melbourne, Australia: MERGA. https://files.eric.ed.gov/fulltext/ED572843.pdf

Fisher, L. D., Dixon, D. O., Herson, J., Frankowski, R. F., & Pearce, K. E. (1990). Analysis of randomized clinical trials: Intention to treat. *Statistical Issues in Drug Research and Development*, *331*, 331-345.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*(5), 531-540. http://dx.doi.org/10.1177/1049731509335549

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613-619. https://doi.org/10.1177/001316447303300309

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R. (2005). Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework, Alexandria, VA: American Statistical Association. https://www.amstat.org/asa/files/pdfs/gaise/gaiseprek-12_full.pdf

Friedman, L. M., Furberg, C., & DeMets, D. L. (2010). *Fundamentals of clinical trials* (4th ed.). New York, NY: Springer. https://link.springer.com/book/10.1007/978-3-319-18539-2

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1-25. http://dx.doi.org/10.1111/j.1751-5823.2002.tb00336.x

Garety, P. A., Freeman, D., Jolley, S., Dunn, G., Bebbington, P. E., Fowler, D. G., ... & Dudley, R. (2005). Reasoning, emotions, and delusional conviction in psychosis. *Journal of Abnormal Psychology*, *114*(3), 373. https://doi.org/10.1037/0021-843x.114.3.373

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92-99. https://doi.org/10.52041/serj.v4i1.527

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics*, *88*, 327-342. http://dx.doi.org/10.1007/s10649-014-9541-7

Gaspard, H., Dicke, A. L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, *51*(9), 1226. https://doi.org/10.1037/dev0000028

Greenberg, M. T., Domitrovich, C. E., Graczyk, P. A., & Zins, J. E. (2005). *The study of implementation in school-based preventive interventions: Theory, practice and research*. Washington, DC: Department of Health and Human Services.

https://www.researchgate.net/publication/253475340_The_Study_of_Implementation_in_School-Based_Preventive_Interventions_Theory_Research_and_Practice

Gregersen, M., Rohd, S. B., Jepsen, J. R. M., Brandt, J. M., Søndergaard, A., Hjorthøj, C., ... & Hemager, N. (2022). Jumping to conclusions and its associations with psychotic experiences in preadolescent children at familial high risk of schizophrenia or bipolar disorder-The Danish high risk and resilience study, VIA 11. *Schizophrenia Bulletin*, *48*(6), 1363-1372. https://doi.org/10.1093/schbul/sbac060

Giraud, G. (1997). Cooperative learning and statistics instruction. *Journal of Statistics Education*, *5*(3). https://doi.org/10.1080/10691898.1997.11910598

Gopal, K., Salim, N. R., & Ayub, A. F. M. (2018, October). RStudio as a tool to motivate students to learn statistics: A study in a Malaysian public university. In *AIP Conference Proceedings* (Vol. 2013, No. 1). AIP Publishing. http://dx.doi.org/10.1063/1.5054226

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science: The Official Journal of the Society for Prevention Research*, *16*(7), 893–926. https://link.springer.com/article/10.1007/s11121-015-0555-x

Gottfried, A. E., & Gottfried, A. W. (1996). A longitudinal study of academic intrinsic motivation in intellectually gifted children: Childhood through early adolescence. *Gifted Child Quarterly*, *40*(4), 179-183. https://doi.org/10.1177/001698629604000402

Gunstone, R. F., & White, R. T. (1981). Understanding of gravity. *Science Education*, *65*(3), 291-299. http://dx.doi.org/10.1002/sce.3730650308

Gürdoğan-Bayır, Ö., & Bozkurt, M. (2018). Effectiveness of cooperative learning approaches used in the course of social studies in Turkey: A meta-analysis study. *European Journal of Education Studies*, *4*(10), 172. http://dx.doi.org/10.5281/zenodo.1313863

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107-128. https://doi.org/10.2307/1164588

Herbein, E., Golle, J., Tibus, M., Zettler, I., & Trautwein, U. (2018). 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. https://doi.org/10.1016/j.cedpsych.2018.09.003

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in education settings: A*

synthesis of the literature. Retrieved from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/IPE_Review_Final.pdf?v=1721031784

Jones, L. V. (1991). *Using cooperative learning to teach statistics*. LL Thurstone Psychometric Laboratory, University of North Carolina.

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253. https://doi.org/10.3102/0013189X20912798

Krummenauer, J., & Kuntze, S. (2018). Primary student's data-based argumentation – an empirical reanalysis. In Bergqvist, E., Österholm, M., Granberg, C., & Sumpter, L. (Eds.), *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 251–258). Umeå, Sweden: PME.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160.

Kuhn, D. (2011). *What is scientific thinking and how does it develop?* In U. Goswami (Eds.), The Wiley-Blackwell handbook of childhood cognitive development (S. 497–523). Wiley-Blackwell. http://dx.doi.org/10.1002/9781444325485.ch19

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development*, *1*(1), 113-129. https://doi.org/10.1207/S15327647JCD0101N_11

Kuhn, S. A. K., Lieb, R., Freeman, D., Andreou, C., & Zander-Schellenberg, T. (2022). Coronavirus conspiracy beliefs in the German-speaking general population: endorsement rates and links to reasoning biases and paranoia. *Psychological Medicine*, *52*(16), 4162-4176. https://doi.org/10.1017/S0033291721001124

Lendrum, A., & Wigelsworth, M. (2013). The evaluation of school-based social and emotional learning interventions: Current issues and future directions. *Psychology of Education Review*, *37*, 70-76. https://research.manchester.ac.uk/en/publications/the-evaluation-of-school-based-social-and-emotional-learning-inte

Lindberg, S., Linkersdörfer, J., Ehm, J. H., Hasselhorn, M., & Lonnemann, J. (2013). Gender Differences in Children's Math Self-Concept in the First Years of Elementary School. *Journal of Education and Learning*, *2*(3), 1-8. https://doi.org/10.5539/jel.v2n3p1

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology*, *81*(1), 59-77. http://dx.doi.org/10.1348/000709910X503501

McKenzie, J. D., Jr. (2004). Conveying the Core Concepts. In ASA Section on Statistical Education. (pp. 2755 - 2757). http://www.statlit.org/pdf/2004mckenzieasa.pdf

Mitchell, J., Ker, D., Lesher, M. (2021). Measuring the economic value of data, Going Digital Toolkit Note, 20. https://goingdigital.oecd.org/data/notes/No20_ToolkitNote_MeasuringtheValueofData.pdf

Moore, D. S. (1990). Uncertainty. *On the shoulders of giants: New approaches to numeracy*, 95-137. https://files.eric.ed.gov/fulltext/ED334084.pdf#page=104

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123-137. https://doi.org/10.1111/j.1751-5823.1997.tb00390.x

Nasser, F. M. (2004). Structural model of the effects of cognitive and affective factors on the achievement of Arabic-speaking pre-service teachers in introductory statistics. *Journal of Statistics Education*, *12*(1). http://dx.doi.org/10.1080/10691898.2004.11910717

Oberski, D. (2014). lavaan.survey: An R package for complex survey analysis of structural equation models. *Journal of Statistical Software*, 57(1), 1-27. http://dx.doi.org/10.18637/jss.v057.i01

OpenAI. (2024). *ChatGPT* (April 10 version) [Large language model]. https://chat.openai.com/chat

Özdemir, D. A., & Işiksal Bostan, M. (2021). Mathematically gifted students' differentiated needs: what kind of support do they need?. *International Journal of Mathematical Education in Science and Technology*, *52*(1), 65-83. https://doi.org/10.1080/0020739X.2019.1658817

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, *7*(2), 83-106. http://dx.doi.org/10.52041/serj.v7i2.471

PARIS21. (2020). Guidelines for Developing Statistical Capacity-A Roadmap for Capacity Development 4.0. Retrieved from: https://paris21.org/sites/default/files/inline-files/UNV003_Guidelines%20for %20Capacity%20Development%20PRINT_0.pdf.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. Oxford, England: Norton.

Preckel, F., Götz, T., & Frenzel, A. (2010). Ability grouping of gifted students: Effects on academic self-concept and boredom. *British Journal of Educational Psychology*, *80*(3), 451-472. https://doi.org/10.1348/000709909X480716

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., & Leutner, D. (2006*). Pisa 2003: Dokumentation der Erhebungsinstrumente [PISA 2003: Documentation of assessment instruments]*. Waxmann. https://hdl.handle.net/11858/00-001M-0000-0025-815F-1

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201–226). Kluwer. https://doi.org/10.1007/1-4020-2278-6_9

Rebholz, F., & Golle, J. (2017). Förderung mathematischer Fähigkeiten in der Grundschule - Die Rolle von Schülerwettbewerben am Beispiel der Mathematik-Olympiade [Fostering mathematical skills in elementary school – the role of academic competions using the example of the Mathematical Olympiad]. In Trautwein, U. & Hasselhorn, M. (Ed.), *Tests und Trends - Jahrbuch der pädagogisch-psychologischen Diagnostik, Band 15. Begabungen und Talente*. (pp. 213–228). Göttingen: Hogrefe. https://doi.org/10.1026/02846-000

Rosseel Y (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02.RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA http://dx.doi.org/10.18637/jss.v048.i02

Ruiz-Primo, M. A., Li, M., Tsai, S. P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *47*(5), 583-608. http://dx.doi.org/10.1002/tea.20356

Schiefer, J., Golle, J., Tibus, M., Herbein, E., Gindele, V., Trautwein, U., & Oschatz, K. (2020). Effects of an extracurricular science intervention on elementary school children's epistemic beliefs: A randomized controlled trial. *British Journal of Educational Psychology*, *90*(2), 382-402. https://doi.org/10.1111/bjep.12301

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. https://doi.org/10.1037/edu0000630

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2020). Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 5. Bis 7. Klasse (BEFKI 5-7). Hogrefe Verlag. https://www.testzentrale.de/shop/berliner-test-zur-erfassung-fluider-und-kristalliner-intelligenz-fuer-die-5-bis-7-jahrgangsstufe.html

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? Development of fluid and crystallized intelligence in 3rd grade. *Intelligence*, *59*, 84–95. https://doi.org/10.1016/j.intell.2016.08.00

Stalder, U. M. (2013). *Leselust in Risikogruppen: Gruppenspezifische Wirkungszusammenhänge [Reading pleasure in risk groups: Group-specific interdependencies]*. Springer Fachmedien. https://doi.org/10.1007/978-3-658-01701-9

Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest*, *12*(1), 3-54. https://doi.org/10.1177/1529100611418056

Torgerson, C. J., & Torgerson, D. J. (2013). *Randomised trials in education: An introductory handbook.* London: EEF. https://www.researchgate.net/publication/273421357_Randomised_trials_in_education_An_introductory_handbook

Trautwein, U., Golle, J., Jaggy, A.-K., 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. https://doi.org/10.1111/nyas.15074

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105. https://doi.org/10.1037/h0031322

Vahey, P., Rafanan, K., Swan, K., van't Hooft, M. A., Annette Kratcoski, R. C. E. T., Stanford, T., & Patton, C. (2010, May). Thinking with data: A cross-disciplinary approach to teaching data literacy and proportionality. In *Annual Conference of the American Educational Research Association.* http://dx.doi.org/10.1007/s10649-012-9392-z

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, *88*(421), 1-8. https://doi.org/10.2307/2290686

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46. https://psycnet.apa.org/record/2009-03902-001

Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *The Journal of Mathematical Behavior*, *19*(1), 109-136. http://dx.doi.org/10.1016/S0732-3123(00)00039-0

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. Educational Studies in Mathematics, 96(1), 33-47. https://link.springer.com/article/10.1007/s10649-017-9764-5

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

What Works Clearinghouse. (2022). What Works Clearinghouse procedures and standards handbook, version 5.0. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). Retrieved from https://ies.ed.gov/ncee/wwc/Handbooks.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-248. https://iase-web.org/documents/intstatreview/99.Wild.Pfannkuch.pdf

Zeidner, M., & Schleyer, E. J. (1999). 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. https://doi.org/10.1006/ceps.1998.0985

Appendix

Text A1: Description of Excluded Measurement Instruments

A1.1 Epistemic Certainty Beliefs

Epistemic beliefs are beliefs about the nature of knowledge and knowing (Hofer & Pintrich, 1997). One dimension of those beliefs focuses on the certainty of knowledge (Conley et al., 2004). Sophisticated *epistemic certainty beliefs* contain the idea that knowledge is almost never certain, but some evidence is better than other to determine its certainty. A better understanding of the concept of variability could relate to a person's epistemic certainty beliefs as variability reflects uncertainty.

We assessed epistemic certainty beliefs with six adapted items of the certainty dimension by Conley et al. (2004; German version: Urhahne & Hopf, 2004). 6 items (e.g., „All scientific questions have exactly one answer. ") were rated on a scale from 1 ("Agree") to 4 ("Do not agree"). Then a mean value was calculated for each participant. Because there were many missing on pretest (49.1 %) and internal consistency was not sufficient (α_{T1} = -.16, α_{T2} = .28), we excluded this variable.

A1.2 Anthropomorphism

Anthropomorphism can be described as the tendency to allocate human characteristics to gods, animals or objects (e.g. Waytz et al., 2010). As human behavior is often associated with predictability (Ayache et al., 2022), people with higher expressions of anthropomorphism could more often associate random events with agency rather than variability. This could result in a lower consideration of variability in data-based arguments.

We assessed anthropomorphism with the *individual differences in anthropomorphism questionnaire* (adapted from Waytz et al., 2010). It consists of 15 self-assessment items (e.g., "To what extent does a car have free will?") which were rated from 1 ("Not at all") to 5 ("Entirely"). Then a mean value was created (α_{T1} = 0.60, α_{T2} = 0.79). 52.8% of data were missing at pretest. Therefore, we excluded this variable from the main analyses.

A1.3 Confirmation Bias

The *confirmation bias* describes the human tendency to look only for evidence that confirms one's own beliefs (Wason, 1968). While forming data-based arguments, people should also look out for contradicting evidence, so the interpretation isn't biased. An increased appreciation of variability could decrease confirmation bias by keeping in mind that preliminary evidence could vary from population parameters.

We assessed the confirmation bias with an adapted version of Muris and colleagues' (2009) modified Wason Selection Task. In two different scenarios, children read information about fictitious animals. One scenario was included threatening information ("If you stroke a suksuk, it will bite you"), the other was non-threatening information ("If you stroke a tingtong, it will lick your hand") about the animal. In each scenario, there were two items which assessed confirmative expectations (e.g. "Andy stroke a suksuk. Do you think Andy was bitten in his hand?") and two items which assessed falsifying expectations (e.g. "An animal scratched Ron's hand. Do you think Ron petted a tingtong?"), which were answered on a scale from "No, for sure not." to "Yes, I am sure". Internal consistency was sufficient for confirmative items (α_{T1} = .65, α_{T2} = .76) and low for falsifying items (α_{T1} = .54, α_{T2} = .42). The missing rate at pretest was 45.3 % for both scales. For those reasons, we excluded these variables from the main analysis.

Table A2

Descriptive Statistics of Excluded Variables: Means, Standard Deviations, Reliabilities, ICCs and Number of Items

Construct			T1					T2				
	N items	Group	*N*	*M*	*SD*	α	*ICC*	*N*	*M*	*SD*	α	*ICC*
Epistemic Certainty Beliefs	6	All	27	3.06	0.35	-.16	.00	51	3.19	0.39	.28	.00
		IG	11	3.14	0.3		.66	24	3.22	0.39		.04
		CG	16	3	0.38		.07	27	3.17	0.4		.00
Anthropomorphism	15	All	25	2.95	0.52	.60	.00	51	2.92	0.63	.79	.00
		IG	11	3.05	0.56		.00	24	2.81	0.66		.06
		CG	14	2.88	0.49		.32	27	3.01	0.61		.00
Confirmation Bias (C)	4	All	29	3.18	0.78	.65	.00	51	3.22	0.76	.76	.00
		IG	12	3.12	0.98		.70	25	3.19	0.71		.00
		CG	17	3.22	0.64		.24	26	3.26	0.83		.00
Confirmation Bias (F)	4	All	29	2.81	0.76	.54	.38	51	2.64	0.55	.42	.00
		IG	12	2.88	0.96		.60	24	2.56	0.52		.00
		CG	17	2.76	0.62		.30	27	2.7	0.57		.00

Note. *N* = number of participating children; *M* = mean; *SD* = standard deviation; IG = intervention group; CG = control group.
Measurement points: T1 = November 2021, T2 = January to February 2022.

Table A3

Baseline Differences of Excluded Variables

	Baseline Differences			
Variable	*g*	*t*	*df*	*p*
Epistemic Certainty Beliefs	-0.40	-1.05	24.51	0.306
Anthropomorphism	-0.33	-0.77	20.09	0.450
Confirmation Bias (C)	0.12	0.28	17.51	0.782
Confirmation Bias (F)	-0.15	-0.35	17.47	0.730

Table A4

Correlations with Confidence Intervals Between the Dependent Variables and Fluid Intelligence on Pretest (Below the Diagonal) and on Posttest (Above the Diagonal)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Data-based Argumentation	.68***	.35*	.24	.13	.06	-.01	.08	-.24	.07	-.10	.27	.20	.35*	.37**
	[.50, .81]	[.09, .57]	[-.04, .48]	[-.15, .39]	[-.22, .32]	[-.28, .27]	[-.20, .35]	[-.49, .03]	[-.21, .34]	[-.36, .19]	[-.01, .51]	[-.08, .45]	[.09, .57]	[.11, .59]
2. Views on Variability	.00	.61***	.37**	.08	-.09	-.00	.08	-.04	.12	-.06	-.05	.08	.11	.09
	[-.29, .29]	[.38, .76]	[.11, .59]	[-.20, .34]	[-.36, .19]	[-.28, .27]	[-.20, .35]	[-.31, .24]	[-.16, .39]	[-.33, .22]	[-.32, .23]	[-.20, .34]	[-.17, .37]	[-.18, .36]
3. Draws to Decision	-.06	.16	.44**	.03	-.08	-.05	.02	-.05	-.05	-.24	-.11	.07	-.09	-.02
	[-.33, .22]	[-.13, .43]	[.18, .64]	[-.25, .30]	[-.35, .20]	[-.32, .22]	[-.26, .29]	[-.32, .23]	[-.32, .23]	[-.48, .04]	[-.38, .18]	[-.21, .33]	[-.35, .19]	[-.30, .25]
4. Self-concept	.09	.13	-.12	.34*	.39**	.52***	-.14	-.10	.02	-.36**	.22	-.11	.24	.26
	[-.19, .36]	[-.17, .41]	[-.38, .17]	[.06, .57]	[.13, .60]	[.28, .69]	[-.40, .15]	[-.37, .18]	[-.26, .29]	[-.58, -.09]	[-.07, .47]	[-.37, .17]	[-.04, .48]	[-.01, .50]
5. Intrinsic Value	.21	.23	.12	.43**	.51***	.45***	.07	-.12	.06	.00	.08	.00	.08	.07
	[-.07, .46]	[-.07, .50]	[-.17, .39]	[.17, .64]	[.26, .69]	[.20, .64]	[-.21, .34]	[-.38, .16]	[-.21, .33]	[-.28, .28]	[-.20, .35]	[-.27, .27]	[-.20, .34]	[-.21, .33]
6. Attainment Value	-.11	-.11	-.33*	.40**	.32*	.29*	.22	-.03	.02	-.24	.04	.09	.10	.16
	[-.38, .18]	[-.39, .20]	[-.56, -.05]	[.13, .61]	[.04, .56]	[.01, .53]	[-.06, .47]	[-.30, .25]	[-.26, .29]	[-.49, .03]	[-.24, .31]	[-.19, .35]	[-.17, .37]	[-.12, .42]
7. Epistemic Certainty Beliefs	-.11	.46*	.14	.23	.13	-.03	.41*	-.05	-.24	.15	-.23	-.10	-.15	-.14
	[-.47, .28]	[.07, .73]	[-.25, .50]	[-.16, .56]	[-.27, .49]	[-.41, .35]	[.01, .69]	[-.32, .23]	[-.48, .04]	[-.13, .42]	[-.48, .05]	[-.37, .18]	[-.41, .13]	[-.40, .14]
8. Anthropomorphism	-.37	.29	-.10	.26	-.00	.31	.18	.72***	.21	-.14	.04	-.30*	-.02	.20
	[-.66, .03]	[-.15, .64]	[-.48, .31]	[-.15, .59]	[-.40, .40]	[-.10, .63]	[-.23, .54]	[.43, .87]	[-.07, .46]	[-.40, .14]	[-.24, .32]	[-.53, -.03]	[-.29, .26]	[-.08, .45]

Table A4

(continued)

9. Confirmation Bias (C)	.01	.21	-.14	.58***	.21	.43*	.46*	.17	.25	-.15	.32*	.06	.46***	.36**
	[-.36, .37]	[-.20, .56]	[-.48, .24]	[.27, .78]	[-.17, .54]	[.08, .69]	[.10, .72]	[-.24, .53]	[-.14, .57]	[-.41, .13]	[.05, .55]	[-.22, .33]	[.21, .65]	[.09, .58]
10. Confirmation Bias (F)	.01	-.40*	-.07	.29	.23	.26	.13	.01	.39*	-.08	-.25	.03	-.15	-.19
	[-.36, .37]	[-.69, -.01]	[-.43, .30]	[-.09, .59]	[-.16, .56]	[-.11, .57]	[-.27, .48]	[-.39, .40]	[.03, .66]	[-.44, .30]	[-.49, .04]	[-.24, .31]	[-.41, .13]	[-.44, .09]
11. Fluid Intelligence	.23	-.05	-.02	.22	.21	-.21	-.02	.20	.04	-.38*				
	[-.05, .48]	[-.33, .25]	[-.30, .26]	[-.07, .47]	[-.07, .47]	[-.47, .08]	[-.40, .36]	[-.21, .55]	[-.33, .40]	[-.66, -.02]				
12. Gender (1 = female)	.12	-.04	.00	-.12	-.02	-.10	-.30	-.39	-.48**	-.44*	-.03			
	[-.16, .38]	[-.33, .25]	[-.28, .28]	[-.39, .17]	[-.30, .27]	[-.37, .19]	[-.61, .09]	[-.68, .01]	[-.72, -.14]	[-.69, -.09]	[-.30, .25]			
13. Grade Level (1= fourth)	.33*	.26	.11	.29*	.25	.12	.17	.32	.38*	.07	.21	.10		
	[.06, .56]	[-.03, .51]	[-.17, .38]	[.01, .53]	[-.03, .50]	[-.17, .39]	[-.23, .52]	[-.09, .63]	[.01, .65]	[-.31, .42]	[-.07, .46]	[-.18, .36]		
14. Age	.34*	.14	.00	.30*	.22	.18	.04	.46*	.27	.09	.22	.13	.84***	
	[.07, .56]	[-.16, .41]	[-.27, .28]	[.02, .54]	[-.06, .47]	[-.11, .44]	[-.35, .41]	[.08, .72]	[-.10, .58]	[-.29, .44]	[-.06, .46]	[-.15, .38]	[.73, .90]	

Note. The diagonal displays the correlations between pre-test and post-test. Values in square brackets indicate the 95% confidence interval for each correlation. Fluid intelligence was only assessed on pre-test. Correlations between fluid intelligence, gender, grade level and age are only reported once because the correlations don't vary between pre-test and post-test *$p < .05$. **$p < .01$. ***$p < .001$.

Table A5

Intervention Effects on Statistical Literacy Outcomes

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Intervention	0.83***	0.22	[0.41; 1.26]	<.001	<.001	0.90***	0.15	[0.60; 1.21]	<.001	<.001	1.01***	0.18	[0.65; 1.37]	<.001	<.001
R^2	.17					.21					.26				
Model 2															
Intervention	0.75***	0.14	[0.46; 1.03]	<.001	<.001	0.87***	0.21	[0.46; 1.27]	<.001	<.001	0.89***	0.19	[0.52; 1.25]	<.001	<.001
Pretest Value	0.67***	0.05	[0.58; 0.77]	<.001		0.57***	0.10	[0.36; 0.77]	<.001		0.37**	0.11	[0.15; 0.59]	.001	
R^2	.60					.54					.35				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A6

Intervention Effects on Motivation for Data-related Tasks

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Intervention	0.59***	0.14	[0.32; 0.86]	<.001	<.001	-0.16	0.34	[-0.82; 0.50]	.635	.953	0.00	0.21	[-0.41; 0.42]	.996	.996
R^2	.09					.01					.00				
Model 2															
Intervention	0.56*	0.22	[0.12; 0.99]	.012	.036	-0.10	0.37	[-0.82; 0.62]	.789	.789	0.07	0.19	[-0.31; 0.45]	.713	.789
Pretest Value	0.32	0.17	[-0.02; 0.65]	.063		0.50***	0.11	[0.29; 0.71]	<.001		0.28	0.21	[-0.13; 0.68]	.179	
R^2	.18					.26					0.08				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A7

Intervention Effects on Epistemic Certainty Beliefs and Anthropomorphism

Variable	Epistemic Certainty Beliefs					Anthropomorphism				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1										
Intervention	0.11	0.16	[-0.22; 0.43]	.521	.608	-0.35	0.20	[-0.75; 0.05]	.084	.147
R^2	.00					.03				
Model 2										
Intervention	-0.01	0.27	[-0.54; 0.51]	.964	.964	-0.37	0.21	[-0.78; 0.05]	.081	.142
Pretest Value	0.46***	0.09	[0.28; 0.63]	<.001		0.65*	0.28	[0.10; 1.20]	.022	
R^2	.22					.38				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A8

Intervention Effects on Confirmation Bias

Variable	Confirmation Bias (C)					Confirmation Bias (F)				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1										
Intervention	-0.08	0.31	[-0.69; 0.53]	.803	.803	-0.25	0.34	[-0.92; 0.42]	.464	.608
R^2	.00					.02				
Model 2										
Intervention	-0.02	0.27	[-0.54; 0.50]	.932	.964	-0.25	0.34	[-0.92; 0.42]	.466	.652
Pretest Value	0.31**	0.10	[0.11; 0.51]	.003		0.01	0.09	[-0.17; 0.19]	.926	
R^2	.10					.02				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A9

Intervention Effects on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.72***	0.06	[0.61; 0.83]	<.001	<.001	0.81***	0.15	[0.52; 1.10]	<.001	<.001	0.80***	0.15	[0.51; 1.09]	<.001	<.001
Data-based Argumentation	0.62***	0.10	[0.42; 0.82]	<.001		0.05	0.12	[-0.19; 0.28]	.697		0.23*	0.09	[0.05; 0.42]	.015	
Views on Variability	-0.03	0.12	[-0.27; 0.21]	.805		0.65***	0.10	[0.44; 0.85]	<.001		-0.08***	0.02	[-0.11; -0.04]	<.001	
Draws to Decision	0.01	0.11	[-0.21; 0.22]	.960		-0.02	0.08	[-0.18; 0.14]	.811		0.36*	0.15	[0.07; 0.65]	.014	
Self-concept	-0.01	0.04	[-0.09; 0.08]	.859		-0.05	0.11	[-0.26; 0.17]	.671		0.17	0.15	[-0.11; 0.46]	.237	
Intrinsic Value	-0.08	0.09	[-0.25; 0.09]	.378		-0.16	0.11	[-0.38; 0.06]	.156		-0.16**	0.06	[-0.27; -0.05]	.005	
Attainment Value	0.09	0.07	[-0.05; 0.24]	.219		-0.01	0.10	[-0.21; 0.19]	.895		-0.24	0.15	[-0.54; 0.05]	.109	
Gender (1=female)	0.19	0.20	[-0.21; 0.58]	.349		0.21*	0.09	[0.04; 0.39]	.017		-0.06	0.07	[-0.20; 0.07]	.372	
Grade Level (1 = fourth)	0.30	0.16	[-0.02; 0.62]	.067		-0.06	0.12	[-0.29; 0.17]	.613		-0.25**	0.08	[-0.40; -0.10]	.001	
Fluid Intelligence	0.13**	0.05	[0.05; 0.22]	.003		0.02	0.05	[-0.08; 0.13]	.691		-0.18*	0.08	[-0.33; -0.03]	.019	
R^2	.66					.58					.54				

Note. For the intervention condition, one-tailed significance levels are reported because directional hypotheses were tested. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A10

Intervention Effects on Motivation for Data-related Tasks Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.68***	0.14	[0.41; 0.95]	<.001	<.001	0.02	0.20	[-0.37; 0.42]	.906	.906	-0.02	0.17	[-0.35; 0.31]	.900	.906
Data-based Argumentation	-0.18	0.13	[-0.43; 0.07]	.164		-0.13	0.13	[-0.39; 0.13]	.338		0.04	0.07	[-0.10; 0.17]	.604	
Views on Variability	-0.08	0.17	[-0.40; 0.25]	.647		-0.16	0.16	[-0.48; 0.15]	.312		0.25	0.22	[-0.19; 0.68]	.264	
Draws to Decision	0.03	0.12	[-0.21; 0.27]	.784		-0.02	0.15	[-0.31; 0.27]	.899		0.15	0.11	[-0.06; 0.35]	.160	
Self-concept	0.12	0.22	[-0.31; 0.55]	.593		-0.39**	0.14	[-0.67; -0.11]	.007		-0.22	0.17	[-0.55; 0.11]	.192	
Intrinsic Value	0.09	0.13	[-0.18; 0.35]	.521		0.77***	0.09	[0.60; 0.95]	<.001		0.05	0.09	[-0.13; 0.23]	.592	
Attainment Value	0.21	0.17	[-0.13; 0.55]	.219		-0.17**	0.06	[-0.28; -0.06]	.003		0.47*	0.19	[0.11; 0.83]	.011	
Gender (1=female)	-0.29	0.52	[-1.32; 0.73]	.572		-0.11	0.37	[-0.83; 0.60]	.761		0.28	0.42	[-0.55; 1.11]	.509	
Grade Level (1 = fourth)	0.39	0.39	[-0.37; 1.15]	.317		0.20	0.37	[-0.53; 0.93]	.594		-0.05	0.39	[-0.82; 0.72]	.894	
Fluid Intelligence	0.18	0.12	[-0.05; 0.41]	.118		-0.02	0.23	[-0.47; 0.42]	.920		0.17	0.12	[-0.07; 0.41]	.169	
R^2	.32					.45					.20				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A11

Intervention Effects on Epistemic Certainty Beliefs and Anthropomorphism Including Baseline Differences

Variable	Epistemic Certainty Beliefs					Anthropomorphism				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.06	0.31	[-0.68; 0.55]	.841	.901	-0.04	0.28	[-0.59; 0.52]	.901	.901
Pretest Value	0.35	0.25	[-0.14; 0.84]	.162		-0.76*	0.33	[-1.42; -0.11]	.022	
Data-based Argumentation	-0.11	0.07	[-0.26; 0.03]	.120		-0.19	0.18	[-0.54; 0.16]	.287	
Views on Variability	-0.15	0.09	[-0.34; 0.03]	.107		-0.60	0.31	[-1.21; 0.01]	.055	
Draws to Decision	-0.04	0.17	[-0.38; 0.30]	.837		-0.42***	0.11	[-0.63; -0.20]	<.001	
Self-concept	-0.12	0.17	[-0.44; 0.21]	.474		0.08*	0.04	[0.00; 0.17]	.050	
Intrinsic Value	0.21	0.11	[-0.02; 0.43]	.074		0.32	0.19	[-0.05; 0.70]	.090	
Attainment Value	-0.09	0.17	[-0.41; 0.24]	.609		-0.69**	0.23	[-1.15; -0.24]	.003	
Gender (1=female)	0.42	0.36	[-0.27; 1.12]	.233		-0.99*	0.40	[-1.77; -0.20]	.014	
Grade Level (1 = fourth)	0.71**	0.27	[0.19; 1.23]	.007		0.70	0.40	[-0.09; 1.49]	.084	
Fluid Intelligence	0.23	0.13	[-0.01; 0.48]	.063		-0.90**	0.30	[-1.49; -0.31]	.003	
R^2	.35					.35				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A12

Intervention Effects on Confirmation Bias Including Baseline Differences

Variable	Confirmation Bias (C)					Confirmation Bias (F)				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.35	0.37	[-1.07; 0.37]	.342	.599	-0.03	0.22	[-0.47; 0.40]	.878	.901
Pretest Value	0.77**	0.24	[0.31; 1.23]	.001		0.57***	0.14	[0.30; 0.84]	<.001	
Data-based Argumentation	0.02	0.31	[-0.59; 0.64]	.942		0.32***	0.10	[0.13; 0.50]	.001	
Views on Variability	-0.19***	0.05	[-0.29; -0.10]	<.001		0.07	0.19	[-0.29; 0.44]	.703	
Draws to Decision	0.04	0.18	[-0.30; 0.38]	.821		0.06	0.13	[-0.18; 0.31]	.613	
Self-concept	-0.18	0.22	[-0.61; 0.25]	.420		-0.42**	0.15	[-0.72; -0.12]	.006	
Intrinsic Value	0.00	0.22	[-0.42; 0.42]	.995		0.06	0.11	[-0.15; 0.27]	.580	
Attainment Value	-0.01	0.13	[-0.27; 0.24]	.932		0.38***	0.11	[0.17; 0.60]	<.001	
Gender (1=female)	-0.30	0.28	[-0.85; 0.25]	.289		0.04	0.16	[-0.27; 0.36]	.779	
Grade Level (1 = fourth)	0.02	0.26	[-0.49; 0.52]	.953		-0.67	0.39	[-1.44; 0.09]	.086	
Fluid Intelligence	-0.05	0.15	[-0.35; 0.25]	.728		-0.06	0.10	[-0.26; 0.14]	.556	
R^2	.47					.52				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A13

Differential Intervention Effects of Previous Knowledge on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.73***	0.06	[0.62; 0.83]	<.001		0.83***	0.18	[0.47; 1.18]	<.001		0.82***	0.15	[0.52; 1.12]	<.001	
Pretest Value × Intervention Condition	0.11	0.19	[-0.26; 0.48]	.570	.801	-0.19	0.18	[-0.53; 0.15]	1.000	1.000	-0.25	0.27	[-0.78; 0.28]	.351	.585
Data-based Argumentation	0.55***	0.09	[0.39; 0.72]	<.001		0.06	0.12	[-0.17; 0.30]	.835		0.23*	0.10	[0.03; 0.43]	.021	
Views on Variability	-0.04	0.16	[-0.34; 0.27]	.820		0.72***	0.11	[0.51; 0.93]	.126		-0.09**	0.03	[-0.15; -0.03]	.002	
Draws to Decision	0.00	0.11	[-0.21; 0.21]	.974		-0.02	0.09	[-0.20; 0.15]	.992		0.48**	0.18	[0.14; 0.82]	.006	
Self-concept	-0.03	0.06	[-0.15; 0.09]	.610		-0.07	0.11	[-0.28; 0.15]	.670		0.17	0.15	[-0.12; 0.47]	.252	
Intrinsic Value	-0.05	0.06	[-0.18; 0.07]	.421		-0.16	0.11	[-0.38; 0.06]	.801		-0.15*	0.06	[-0.26; -0.03]	.011	
Attainment Value	0.09	0.07	[-0.05; 0.22]	.211		-0.01	0.10	[-0.19; 0.18]	.780		-0.28	0.19	[-0.66; 0.09]	.139	
Gender (1=female)	0.21	0.18	[-0.13; 0.56]	.227		0.22**	0.08	[0.06; 0.37]	.763		-0.08	0.05	[-0.18; 0.03]	.136	
Grade Level (1 = fourth)	0.31*	0.16	[0.00; 0.62]	.050		-0.04	0.11	[-0.27; 0.18]	.384		-0.29***	0.04	[-0.37; -0.21]	<.001	
Fluid Intelligence	0.14**	0.04	[0.05; 0.22]	.002		0.03	0.04	[-0.05; 0.11]	.365		-0.18*	0.09	[-0.35; -0.01]	.033	
R^2	.66					.59					.56				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A14

Differential Intervention Effects of Previous Knowledge on Motivation for Data-related Tasks Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.74***	0.17	[0.41; 1.07]	<.001		0.02	0.19	[-0.35; 0.40]	.901		0.00	0.20	[-0.39; 0.39]	.994	
Pretest Value × Intervention condition	-0.64	0.26	[-1.16; -0.13]	.014	.070	0.05	0.28	[-0.50; 0.61]	.853	.956	-0.21	0.42	[-1.04; 0.62]	.623	.814
Data-based Argumentation	-0.11*	0.05	[-0.21; -0.01]	.038		-0.12	0.15	[-0.42; 0.17]	.402		0.02	0.10	[-0.18; 0.22]	.835	
Views on Variability	-0.11	0.12	[-0.35; 0.12]	.346		-0.16	0.17	[-0.49; 0.18]	.353		0.25	0.21	[-0.17; 0.67]	.248	
Draws to Decision	-0.01	0.12	[-0.25; 0.24]	.962		-0.02	0.15	[-0.31; 0.28]	.909		0.09	0.09	[-0.09; 0.27]	.314	
Self-concept	0.45*	0.22	[0.02; 0.88]	.040		-0.39**	0.15	[-0.68; -0.10]	.010		-0.19	0.19	[-0.55; 0.18]	.316	
Intrinsic Value	0.13	0.09	[-0.04; 0.30]	.125		0.75***	0.12	[0.52; 0.97]	<.001		0.10	0.13	[-0.15; 0.34]	.442	
Attainment Value	0.30*	0.12	[0.06; 0.53]	.015		-0.18**	0.07	[-0.31; -0.05]	.007		0.55*	0.25	[0.07; 1.04]	.025	
Gender (1=female)	-0.35	0.53	[-1.38; 0.69]	.512		-0.10	0.38	[-0.85; 0.66]	.804		0.20	0.37	[-0.52; 0.92]	.591	
Grade level (1 = fourth)	0.44	0.31	[-0.17; 1.05]	.154		0.18	0.42	[-0.64; 1.01]	.660		-0.01	0.43	[-0.85; 0.83]	.980	
Fluid Intelligence	0.12	0.08	[-0.03; 0.27]	.106		-0.03	0.23	[-0.47; 0.42]	.912		0.16	0.13	[-0.09; 0.41]	.204	
R^2	.39					.46					.22				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A15

Differential Intervention Effects of Previous Knowledge on Epistemic Certainty Beliefs and Anthropomorphism Including Baseline Differences

Variable	Epistemic Certainty Beliefs					Anthropomorphism				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.47***	0.10	[-0.66; -0.28]	<.001		-0.20	0.29	[-0.76; 0.36]	.489	
Pretest Value × Intervention condition	-0.11	0.07	[-0.24; 0.03]	.133	.410	0.21	0.31	[-0.40; 0.82]	.493	.758
Pretest Value	2.17***	0.10	[1.97; 2.37]	<.001		1.08**	0.40	[0.30; 1.87]	.007	
Data-based Argumentation	-0.29	0.17	[-0.61; 0.04]	.085		-0.15	0.12	[-0.38; 0.08]	.215	
Views on Variability	0.40***	0.07	[0.27; 0.53]	<.001		-0.25***	0.07	[-0.39; -0.12]	<.001	
Draws to Decision	-0.11	0.11	[-0.32; 0.10]	.311		-0.04	0.13	[-0.29; 0.22]	.781	
Self-concept	-0.47***	0.11	[-0.69; -0.25]	<.001		-0.21	0.13	[-0.46; 0.03]	.090	
Intrinsic Value	-0.10	0.13	[-0.35; 0.15]	.446		-0.22	0.22	[-0.65; 0.21]	.324	
Attainment Value	0.05	0.16	[-0.26; 0.36]	.737		-0.03	0.09	[-0.21; 0.14]	.710	
Gender (1=female)	0.80***	0.13	[0.55; 1.04]	<.001		-0.46*	0.23	[-0.92; 0.00]	.049	
Grade Level (1 = fourth)	-0.61*	0.29	[-1.19; -0.04]	.037		0.11	0.12	[-0.12; 0.34]	.343	
Fluid Intelligence	-0.06	0.05	[-0.16; 0.04]	.247		-0.05	0.06	[-0.16; 0.06]	.349	
R^2	.80					.64				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A16

Differential Intervention Effects of Previous Knowledge on Intervention Effects on Confirmation Bias Including Baseline Differences

Variable	Confirmation Bias (C)					Confirmation Bias (F)				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.23	0.30	[-0.82; 0.36]	.445		-0.03	0.23	[-0.49; 0.43]	.896	
Pretest Value × Intervention condition	-0.24	0.22	[-0.67; 0.20]	.284	.541	-0.68	0.40	[-1.45; 0.10]	.088	.352
Pretest value	1.23*	0.50	[0.25; 2.21]	.014		-0.18	0.26	[-0.68; 0.32]	.487	
Data-based Argumentation	-0.11	0.08	[-0.26; 0.04]	.161		-0.17	0.18	[-0.53; 0.18]	.343	
Views on Variability	0.27	0.14	[-0.01; 0.54]	.055		-0.61	0.31	[-1.22; 0.01]	.053	
Draws to Decision	0.19*	0.08	[0.03; 0.35]	.022		-0.41***	0.10	[-0.60; -0.22]	<.001	
Self-concept	0.05	0.20	[-0.34; 0.45]	.788		0.11*	0.05	[0.01; 0.21]	.035	
Intrinsic Value	-0.18	0.18	[-0.54; 0.19]	.343		0.31	0.20	[-0.07; 0.69]	.112	
Attainment Value	0.01	0.16	[-0.31; 0.32]	.974		-0.68**	0.22	[-1.12; -0.24]	.002	
Gender (1=female)	1.39*	0.60	[0.22; 2.57]	.020		-1.01**	0.37	[-1.73; -0.29]	.006	
Grade Level (1 = fourth)	0.09	0.36	[-0.62; 0.80]	.806		0.71	0.40	[-0.08; 1.50]	.080	
Fluid Intelligence	0.45**	0.15	[0.15; 0.75]	.003		-0.90**	0.30	[-1.49; -0.31]	.003	
R^2	.44					.36				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A17

Differential Intervention Effects of Gender on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.83***	0.07	[0.68; 0.97]	<.001		0.82***	0.22	[0.39; 1.25]	<.001		0.62***	0.08	[0.46; 0.78]	<.001	
Gender × Intervention Condition	-0.56	0.22	[-1.00; -0.12]	.013	.070	-0.03	0.55	[-1.10; 1.04]	.956	.981	0.91	0.41	[0.11; 1.71]	.027	.120
Data-based Argumentation	0.59***	0.10	[0.40; 0.78]	<.001		0.04	0.11	[-0.17; 0.25]	.714		0.29**	0.09	[0.11; 0.47]	.002	
Views on Variability	-0.02	0.14	[-0.29; 0.24]	.859		0.65***	0.10	[0.44; 0.85]	<.001		-0.09	0.07	[-0.22; 0.05]	.211	
Draws to Decision	-0.01	0.11	[-0.22; 0.21]	.942		-0.02	0.08	[-0.19; 0.14]	.774		0.39**	0.14	[0.12; 0.66]	.004	
Self-concept	0.00	0.04	[-0.08; 0.09]	.909		-0.05	0.12	[-0.28; 0.19]	.708		0.15	0.15	[-0.15; 0.44]	.329	
Intrinsic Value	-0.08	0.08	[-0.24; 0.07]	.299		-0.16	0.12	[-0.39; 0.07]	.166		-0.15*	0.07	[-0.29; -0.01]	.032	
Attainment Value	0.06	0.06	[-0.06; 0.18]	.323		-0.02	0.12	[-0.26; 0.23]	.896		-0.18	0.12	[-0.42; 0.06]	.151	
Gender (1=female)	0.53*	0.22	[0.09; 0.96]	.017		0.23	0.29	[-0.35; 0.81]	.432		-0.62***	0.16	[-0.94; -0.30]	<.001	
Grade Level (1 = fourth)	0.28	0.15	[-0.01; 0.57]	.059		-0.06	0.12	[-0.28; 0.17]	.608		-0.22***	0.06	[-0.34; -0.11]	<.001	
Fluid Intelligence	0.14***	0.04	[0.06; 0.22]	.001		0.02	0.05	[-0.08; 0.13]	.669		-0.18*	0.09	[-0.35; -0.01]	.038	
R^2	.68					.58					.57				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A18

Differential Intervention Effects of Gender on Motivation for Data-related Tasks Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.65***	0.16	[0.34; 0.97]	<.001		0.03	0.18	[-0.31; 0.38]	.846		-0.04	0.18	[-0.40; 0.32]	.837	
Gender × Intervention Condition	0.10	0.33	[-0.54; 0.74]	.766	.928	-0.06	0.30	[-0.65; 0.53]	.843	.956	0.07	0.40	[-0.70; 0.84]	.860	.956
Data-based Argumentation	-0.16	0.15	[-0.45; 0.13]	.282		-0.13	0.14	[-0.41; 0.15]	.357		0.04	0.07	[-0.10; 0.19]	.544	
Views on Variability	-0.06	0.17	[-0.40; 0.27]	.715		-0.16	0.13	[-0.40; 0.09]	.205		0.25	0.20	[-0.15; 0.65]	.219	
Draws to Decision	0.05	0.11	[-0.18; 0.27]	.671		-0.02	0.13	[-0.28; 0.25]	.888		0.15	0.11	[-0.07; 0.37]	.189	
Self-concept	0.10	0.22	[-0.33; 0.53]	.637		-0.39**	0.14	[-0.66; -0.12]	.005		-0.22	0.16	[-0.54; 0.10]	.181	
Intrinsic Value	0.07	0.14	[-0.20; 0.35]	.597		0.77***	0.10	[0.58; 0.97]	<.001		0.05	0.09	[-0.13; 0.23]	.603	
Attainment Value	0.25	0.19	[-0.12; 0.61]	.184		-0.17***	0.04	[-0.26; -0.08]	<.001		0.48**	0.17	[0.14; 0.82]	.006	
Gender (1=female)	-0.34	0.42	[-1.17; 0.49]	.420		-0.08	0.30	[-0.66; 0.51]	.802		0.25	0.35	[-0.44; 0.93]	.480	
Grade Level (1 = fourth)	0.37	0.38	[-0.38; 1.13]	.331		0.19	0.36	[-0.50; 0.89]	.586		-0.06	0.38	[-0.81; 0.70]	.883	
Fluid Intelligence	0.19	0.11	[-0.02; 0.40]	.078		-0.02	0.23	[-0.46; 0.42]	.933		0.17	0.12	[-0.07; 0.42]	.168	
R^2	.32					.45					.21				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A19

Differential Intervention Effects of Gender on Epistemic Certainty Beliefs and Anthropomorphism Including Baseline Differences

Variable	Epistemic Certainty Beliefs					Anthropomorphism				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.31	0.28	[-0.86; 0.24]	.270		-0.48	0.36	[-1.18; 0.22]	.182	
Gender × Intervention Condition	1.13	0.45	[0.25; 2.01]	.012	.070	0.64	0.62	[-0.58; 1.85]	.303	.551
Pretest Value	0.63**	0.20	[0.24; 1.02]	.001		0.78***	0.21	[0.35; 1.20]	<.001	
Data-based Argumentation	0.39**	0.13	[0.14; 0.64]	.002		0.06	0.32	[-0.57; 0.68]	.861	
Views on Variability	0.02	0.13	[-0.23; 0.28]	.854		-0.21***	0.06	[-0.32; -0.09]	.001	
Draws to Decision	0.09	0.13	[-0.17; 0.35]	.515		0.06	0.18	[-0.29; 0.42]	.736	
Self-concept	-0.43**	0.14	[-0.71; -0.15]	.002		-0.18	0.23	[-0.62; 0.26]	.427	
Intrinsic Value	0.10	0.08	[-0.05; 0.25]	.196		0.01	0.22	[-0.42; 0.44]	.965	
Attainment Value	0.45***	0.09	[0.28; 0.62]	<.001		0.01	0.10	[-0.18; 0.21]	.901	
Gender (1=female)	-0.49	0.35	[-1.18; 0.19]	.157		-0.68	0.39	[-1.44; 0.07]	.077	
Grade level (1 = fourth)	-0.67	0.34	[-1.34; 0.01]	.052		0.04	0.27	[-0.48; 0.57]	.868	
Fluid Intelligence	-0.05	0.11	[-0.27; 0.17]	.666		-0.06	0.15	[-0.36; 0.24]	.706	
R^2	.55					.49				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A20

Differential Intervention Effects of Gender on Confirmation Bias Including Baseline Differences

Variable	Confirmation Bias (C)					Confirmation Bias (F)				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.26	0.26	[-0.25; 0.77]	.323		0.13	0.45	[-0.75; 1.01]	.774	
Gender × Intervention Condition	-1.74***	0.48	[-2.69; -0.80]	<.001	<.001	-0.76	0.61	[-1.96; 0.44]	.215	.430
Pretest Value	0.10	0.16	[-0.21; 0.41]	.529		-0.77*	0.34	[-1.44; -0.11]	.023	
Data-based Argumentation	-0.19*	0.09	[-0.37; -0.01]	.039		-0.24	0.23	[-0.69; 0.20]	.289	
Views on Variability	-0.10	0.05	[-0.20; 0.00]	.051		-0.58	0.31	[-1.19; 0.04]	.065	
Draws to Decision	-0.07	0.18	[-0.42; 0.28]	.693		-0.44**	0.15	[-0.73; -0.16]	.002	
Self-concept	0.03	0.14	[-0.24; 0.30]	.816		0.10	0.07	[-0.03; 0.23]	.128	
Intrinsic Value	0.17	0.11	[-0.05; 0.38]	.127		0.30	0.18	[-0.06; 0.66]	.100	
Attainment Value	-0.18	0.22	[-0.62; 0.26]	.419		-0.74**	0.28	[-1.29; -0.18]	.009	
Gender (1=female)	1.27**	0.42	[0.45; 2.09]	.002		-0.59***	0.11	[-0.81; -0.36]	<.001	
Grade Level (1 = fourth)	0.76*	0.33	[0.12; 1.40]	.019		0.70	0.43	[-0.15; 1.56]	.106	
Fluid Intelligence	0.22	0.12	[-0.01; 0.45]	.060		-0.92**	0.31	[-1.53; -0.30]	.003	
R^2	.42					.35				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A21

Differential Intervention Effects of Grade Level on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.71***	0.12	[0.47; 0.95]	<.001		0.40*	0.20	[0.01; 0.79]	.047		0.64***	0.16	[0.32; 0.96]	<.001	
Grade Level × Intervention Condition	0.03	0.21	[-0.39; 0.44]	.899	.972	0.80***	0.22	[0.37; 1.22]	<.001	<.001	0.33	0.20	[-0.06; 0.73]	.099	.360
Data-based Argumentation	0.62***	0.10	[0.42; 0.82]	<.001		0.06	0.11	[-0.16; 0.28]	.588		0.23*	0.09	[0.05; 0.42]	.013	
Views on Variability	-0.03	0.12	[-0.26; 0.20]	.799		0.66***	0.08	[0.49; 0.82]	<.001		-0.07*	0.03	[-0.13; -0.01]	.017	
Draws to Decision	0.00	0.11	[-0.21; 0.21]	.996		0.05	0.07	[-0.09; 0.20]	.455		0.38*	0.15	[0.09; 0.66]	.010	
Self-concept	-0.01	0.05	[-0.10; 0.08]	.824		-0.06	0.13	[-0.32; 0.19]	.621		0.17	0.15	[-0.13; 0.47]	.274	
Intrinsic Value	-0.07	0.09	[-0.25; 0.11]	.470		-0.23*	0.10	[-0.44; -0.03]	.026		-0.18***	0.05	[-0.28; -0.08]	.001	
Attainment Value	0.09	0.07	[-0.05; 0.23]	.213		0.01	0.11	[-0.21; 0.24]	.911		-0.23	0.15	[-0.53; 0.06]	.120	
Gender (1=female)	0.19	0.19	[-0.19; 0.57]	.329		0.32**	0.10	[0.12; 0.52]	.002		-0.02	0.06	[-0.14; 0.09]	.664	
Grade level (1 = fourth)	0.29	0.23	[-0.17; 0.74]	.215		-0.47**	0.15	[-0.78; -0.17]	.002		-0.43*	0.19	[-0.81; -0.06]	.024	
Fluid Intelligence	0.13**	0.04	[0.04; 0.22]	.004		0.08	0.06	[-0.04; 0.19]	.182		-0.16	0.09	[-0.33; 0.01]	.063	
R^2	.66					.62					.54				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A22

Differential Intervention Effects of Grade Level on Motivation for Data-related Tasks Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.18	0.16	[-0.14; 0.50]	.264		-0.18	0.17	[-0.51; 0.15]	.280		-0.35	0.38	[-1.10; 0.40]	.361	
Grade Level × Intervention Condition	0.95***	0.22	[0.52; 1.38]	<.001	<.001	0.39	0.28	[-0.17; 0.95]	.168	.430	0.62	0.47	[-0.31; 1.54]	.191	.430
Data-based Argumentation	-0.17	0.13	[-0.42; 0.07]	.173		-0.13	0.14	[-0.39; 0.14]	.341		0.05	0.07	[-0.09; 0.19]	.466	
Views on Variability	-0.08	0.17	[-0.42; 0.26]	.635		-0.16	0.17	[-0.48; 0.17]	.342		0.26	0.24	[-0.21; 0.73]	.285	
Draws to Decision	0.11	0.14	[-0.18; 0.39]	.459		0.00	0.14	[-0.26; 0.27]	.974		0.21	0.13	[-0.04; 0.45]	.107	
Self-concept	0.10	0.21	[-0.32; 0.52]	.629		-0.39**	0.14	[-0.65; -0.12]	.004		-0.23	0.16	[-0.55; 0.09]	.153	
Intrinsic Value	-0.01	0.10	[-0.21; 0.20]	.948		0.74***	0.11	[0.53; 0.95]	<.001		-0.01	0.12	[-0.24; 0.22]	.929	
Attainment Value	0.25	0.19	[-0.13; 0.63]	.196		-0.17**	0.06	[-0.28; -0.05]	.003		0.50**	0.19	[0.13; 0.88]	.009	
Gender (1=female)	-0.20	0.45	[-1.08; 0.68]	.654		-0.07	0.34	[-0.73; 0.59]	.833		0.34	0.40	[-0.45; 1.13]	.400	
Grade Level (1 = fourth)	-0.09	0.39	[-0.85; 0.68]	.818		0.01	0.37	[-0.70; 0.73]	.975		-0.37	0.56	[-1.46; 0.73]	.513	
Fluid Intelligence	0.25*	0.11	[0.03; 0.47]	.023		0.01	0.22	[-0.43; 0.44]	.978		0.22	0.16	[-0.09; 0.53]	.168	
R^2	.36					.46					.22				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A23

Differential Intervention Effects of Grade Level on Epistemic Certainty Beliefs and Anthropomorphism Including Baseline Differences

Variable	Epistemic Certainty Beliefs					Anthropomorphism				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.14	0.37	[-0.86; 0.58]	.707		-0.78*	0.34	[-1.44; -0.12]	.020	
Grade Level × Intervention Condition	0.20	0.42	[-0.62; 1.02]	.631	.814	0.91	0.68	[-0.42; 2.24]	.181	.430
Pretest Value	0.56***	0.13	[0.30; 0.83]	<.001		0.75***	0.22	[0.32; 1.19]	.001	
Data-based Argumentation	0.33**	0.10	[0.13; 0.53]	.001		0.02	0.33	[-0.62; 0.66]	.952	
Views on Variability	0.08	0.17	[-0.26; 0.42]	.656		-0.17**	0.06	[-0.28; -0.06]	.002	
Draws to Decision	0.08	0.12	[-0.16; 0.31]	.528		0.12	0.19	[-0.25; 0.48]	.534	
Self-concept	-0.43**	0.17	[-0.76; -0.11]	.009		-0.21	0.24	[-0.68; 0.25]	.371	
Intrinsic Value	0.05	0.14	[-0.23; 0.33]	.723		-0.09	0.22	[-0.52; 0.35]	.698	
Attainment Value	0.39**	0.12	[0.14; 0.63]	.002		0.04	0.20	[-0.35; 0.44]	.826	
Gender (1=female)	0.06	0.15	[-0.25; 0.36]	.715		-0.23	0.20	[-0.61; 0.16]	.243	
Grade Level (1 = fourth)	-0.74	0.48	[-1.68; 0.19]	.118		-0.45	0.28	[-1.01; 0.10]	.111	
Fluid Intelligence	-0.05	0.09	[-0.21; 0.12]	.581		0.06	0.24	[-0.40; 0.53]	.784	
R^2	.51					.54				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A24

Differential Intervention Effects of Grade Level on Confirmation Bias Including Baseline Differences

Variable	Confirmation Bias (C)					Confirmation Bias (F)				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.14	0.33	[-0.78; 0.50]	.669		0.39	0.60	[-0.78; 1.56]	.513	
Grade Level × Intervention Condition	0.15	0.28	[-0.39; 0.70]	.581	.801	-0.85	0.68	[-2.17; 0.48]	.209	.430
Pretest Value	0.35	0.26	[-0.15; 0.85]	.171		-0.72*	0.34	[-1.39; -0.06]	.032	
Data-based Argumentation	-0.12	0.08	[-0.27; 0.03]	.116		-0.20	0.19	[-0.58; 0.18]	.300	
Views on Variability	-0.15	0.10	[-0.35; 0.05]	.133		-0.56	0.32	[-1.20; 0.07]	.083	
Draws to Decision	-0.03	0.17	[-0.37; 0.31]	.863		-0.48**	0.16	[-0.79; -0.16]	.003	
Self-concept	-0.12	0.17	[-0.45; 0.21]	.466		0.09	0.06	[-0.02; 0.21]	.113	
Intrinsic Value	0.21	0.15	[-0.09; 0.51]	.176		0.38	0.26	[-0.14; 0.89]	.152	
Attainment Value	-0.09	0.16	[-0.41; 0.23]	.584		-0.70**	0.23	[-1.16; -0.25]	.002	
Gender (1=female)	0.44	0.37	[-0.29; 1.17]	.241		-1.01*	0.49	[-1.97; -0.04]	.042	
Grade Level (1 = fourth)	0.64*	0.32	[0.01; 1.26]	.046		1.08	0.73	[-0.36; 2.52]	.141	
Fluid Intelligence	0.25	0.15	[-0.05; 0.54]	.104		-0.92**	0.32	[-1.55; -0.29]	.004	
R^2	.35					.37				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A25

Differential Intervention Effects of Fluid Intelligence on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.71***	0.09	[0.54; 0.87]	<.001		0.81***	0.16	[0.49; 1.13]	<.001		0.81***	0.11	[0.59; 1.04]	<.001	
Fluid Intelligence × Intervention condition	0.39*	0.12	[0.14; 0.63]	.002	.016	-0.02	0.20	[-0.42; 0.38]	.935	.981	-0.26	0.20	[-0.66; 0.14]	.318	.553
Data-based Argumentation	0.60***	0.10	[0.42; 0.79]	<.001		0.05	0.13	[-0.20; 0.29]	.699		0.24*	0.10	[0.04; 0.44]	.154	
Views on Variability	-0.05	0.10	[-0.25; 0.14]	.595		0.65***	0.09	[0.46; 0.83]	<.001		-0.05*	0.02	[-0.10; -0.01]	.117	
Draws to Decision	-0.01	0.11	[-0.23; 0.21]	.949		-0.02	0.08	[-0.18; 0.14]	.826		0.37**	0.14	[0.11; 0.64]	.172	
Self-concept	0.02	0.06	[-0.11; 0.15]	.760		-0.05	0.10	[-0.24; 0.15]	.642		0.15	0.15	[-0.13; 0.44]	.422	
Intrinsic Value	-0.09	0.06	[-0.22; 0.03]	.136		-0.17	0.11	[-0.38; 0.05]	.131		-0.15**	0.06	[-0.26; -0.04]	.144	
Attainment Value	0.06	0.08	[-0.09; 0.20]	.448		-0.01	0.10	[-0.21; 0.19]	.923		-0.22	0.13	[-0.47; 0.04]	.219	
Gender (1=female)	0.19	0.21	[-0.21; 0.60]	.352		0.21*	0.09	[0.05; 0.38]	.012		-0.06	0.07	[-0.21; 0.08]	.882	
Grade Level (1 = fourth)	0.39*	0.20	[0.00; 0.78]	.047		-0.05	0.13	[-0.30; 0.20]	.676		-0.31**	0.10	[-0.51; -0.11]	.264	
Fluid Intelligence	-0.04	0.09	[-0.21; 0.14]	.697		0.03	0.13	[-0.22; 0.27]	.834		-0.06	0.16	[-0.39; 0.26]	.789	
R^2	.69					.59					.56				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A26

Differential Intervention Effects of Fluid Intelligence on Motivation for Data-related Tasks Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.68***	0.15	[0.39; 0.97]	<.001		0.03	0.19	[-0.34; 0.39]	.888		-0.01	0.18	[-0.37; 0.34]	.935	
Fluid Intelligence × Intervention Condition	-0.11	0.26	[-0.62; 0.40]	.672	.840	-0.12	0.17	[-0.46; 0.21]	.475	.758	-0.29	0.23	[-0.74; 0.16]	.209	.430
Data-based Argumentation	-0.17	0.11	[-0.39; 0.05]	.126		-0.12	0.13	[-0.37; 0.12]	.326		0.05	0.08	[-0.10; 0.20]	.506	
Views on Variability	-0.05	0.21	[-0.47; 0.36]	.804		-0.15	0.16	[-0.46; 0.17]	.371		0.27	0.22	[-0.16; 0.71]	.219	
Draws to Decision	0.04	0.14	[-0.22; 0.31]	.757		-0.02	0.15	[-0.31; 0.27]	.918		0.16	0.11	[-0.05; 0.37]	.141	
Self-concept	0.10	0.23	[-0.35; 0.56]	.653		-0.40**	0.14	[-0.68; -0.12]	.005		-0.24	0.15	[-0.53; 0.05]	.105	
Intrinsic Value	0.08	0.14	[-0.20; 0.37]	.558		0.78***	0.08	[0.62; 0.94]	<.001		0.06	0.10	[-0.14; 0.25]	.583	
Attainment Value	0.23	0.17	[-0.10; 0.55]	.167		-0.16*	0.07	[-0.29; -0.03]	.016		0.51**	0.18	[0.16; 0.86]	.005	
Gender (1=female)	-0.29	0.52	[-1.31; 0.74]	.582		-0.11	0.37	[-0.82; 0.61]	.772		0.28	0.43	[-0.56; 1.12]	.517	
Grade Level (1 = fourth)	0.35	0.38	[-0.40; 1.10]	.355		0.16	0.37	[-0.56; 0.88]	.666		-0.11	0.39	[-0.87; 0.64]	.768	
Fluid Intelligence	0.24	0.22	[-0.20; 0.67]	.286		0.04	0.23	[-0.42; 0.49]	.876		0.30	0.20	[-0.09; 0.68]	.129	
R^2	.32					.46					.22				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A27

Differential Intervention Effects of Fluid Intelligence on Epistemic Certainty Beliefs and Anthropomorphism Including Baseline Differences

Variable	Epistemic Certainty Beliefs					Anthropomorphism				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.08	0.16	[-0.24; 0.40]	.636		-0.29	0.43	[-1.14; 0.55]	.498	
Fluid Intelligence × Intervention Condition	0.53	0.34	[-0.13; 1.19]	.114	.380	0.48*	0.15	[0.18; 0.78]	.002	.016
Pretest Value	0.59**	0.19	[0.22; 0.96]	.002		0.73***	0.21	[0.32; 1.14]	<.001	
Data-based Argumentation	0.40*	0.20	[0.00; 0.81]	.049		0.01	0.29	[-0.56; 0.58]	.973	
Views on Variability	0.05	0.16	[-0.27; 0.37]	.749		-0.21***	0.03	[-0.28; -0.15]	<.001	
Draws to Decision	0.05	0.13	[-0.20; 0.30]	.686		0.02	0.17	[-0.32; 0.35]	.928	
Self-concept	-0.32	0.22	[-0.74; 0.10]	.134		-0.14	0.22	[-0.56; 0.29]	.529	
Intrinsic Value	-0.08	0.25	[-0.57; 0.40]	.740		-0.03	0.19	[-0.41; 0.35]	.877	
Attainment Value	0.40***	0.11	[0.18; 0.62]	<.001		-0.03	0.14	[-0.31; 0.25]	.832	
Gender (1=female)	0.13	0.29	[-0.45; 0.71]	.662		-0.33	0.26	[-0.84; 0.19]	.214	
Grade Level (1 = fourth)	-0.64	0.48	[-1.59; 0.31]	.187		0.11	0.23	[-0.35; 0.56]	.649	
Fluid Intelligence	-0.25**	0.09	[-0.42; -0.07]	.005		-0.23	0.13	[-0.48; 0.02]	.072	
R^2	.52					.52				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A28

Differential Intervention Effects of Fluid Intelligence on Confirmation Bias Including Baseline Differences

Variable	Confirmation Bias (C)					Confirmation Bias (F)				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.02	0.30	[-0.61; 0.57]	.944		-0.02	0.25	[-0.51; 0.47]	.943	
Fluid Intelligence × Intervention Condition	0.34	0.25	[-0.14; 0.83]	.164	.430	0.19	0.30	[-0.40; 0.78]	.536	.794
Pretest Value	0.45	0.26	[-0.07; 0.96]	.089		-0.73	0.39	[-1.48; 0.03]	.059	
Data-based Argumentation	-0.09	0.09	[-0.26; 0.08]	.316		-0.16	0.21	[-0.58; 0.26]	.443	
Views on Variability	-0.21	0.16	[-0.52; 0.09]	.174		-0.58	0.36	[-1.28; 0.13]	.110	
Draws to Decision	-0.05	0.18	[-0.40; 0.30]	.795		-0.41***	0.11	[-0.64; -0.19]	<.001	
Self-concept	-0.09	0.16	[-0.41; 0.22]	.568		0.12	0.07	[-0.02; 0.26]	.101	
Intrinsic Value	0.16	0.11	[-0.06; 0.38]	.163		0.27	0.23	[-0.19; 0.72]	.257	
Attainment Value	-0.11	0.16	[-0.42; 0.20]	.485		-0.68**	0.23	[-1.13; -0.23]	.003	
Gender (1=female)	0.57	0.37	[-0.17; 1.30]	.131		-0.89	0.49	[-1.85; 0.07]	.068	
Grade Level (1 = fourth)	0.71**	0.27	[0.19; 1.23]	.008		0.67	0.51	[-0.32; 1.66]	.186	
Fluid Intelligence	0.09	0.19	[-0.29; 0.46]	.656		-0.94*	0.43	[-1.79; -0.10]	.028	
R^2	.38					.34				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Appendix References

Ayache, J., Connor, A. M., Marks, S., Sumich, A., & Heym, N. (2022, December). Humanness lies in unpredictability: Role of Theory of Mind on anthropomorphism in human-computer interactions. In *Proceedings of the 10th International Conference on Human-Agent Interaction* (pp. 306-308). https://doi.org/10.1145/3527188.3563920

Conley, A. M., Pintrich, P. R., Vekiri, I. & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, *29*, 186-204. https://doi.org/10.1016/j.cedpsych.2004.01.004

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, *67*(1), 88-140. https://doi.org/10.2307/1170620

Muris, P., Rassin, E., Mayer, B., Smeets, G., Huijding, J., Remmerswaal, D., & Field, A. (2009). Effects of verbal information on fear-related reasoning biases in children. *Behaviour Research and Therapy*, *47*(3), 206-214. https://doi.org/10.1016/j.brat.2008.12.002

Urhahne, D., & Hopf, M. (2004). Epistemologische Überzeugungen in den Naturwissenschaften und ihre Zusammenhänge mit Motivation, Selbstkonzept und Lernstrategien. *Zeitschrift für Didaktik der Naturwissenschaften*, *10*, 71–87. https://doi.org/10.25656/01:31598

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273-281. https://doi.org/10.1080/14640746808400161

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219-232. https://doi.org/10.1177/1745691610369336

3

Study 2: Assessing Decision Thresholds in Primary School Students Using Signal Detection Theory: Validating an Adapted Version of the Beads Task

Stark, L., Goecke, B., Jaggy, A.-K., Krummenauer, J., Kuntze, S., Golle, J., Nagengast, B. (2025). *Assessing decision thresholds in primary school students using signal detection theory: Validating an adapted version of the beads task*. Manuscript in preparation.

The Hector Foundation II supported this work. Lucas Stark is a doctoral student at the LEAD Graduate School & Research Network [GSC 1028], funded by the Baden-Württemberg Ministry of Science, Research and the Arts within the framework of sustainability funding for the projects of the Excellence Initiative II.

Abstract

The decision threshold describes the minimum probability a person needs to be certain enough to make a decision. Lower decision thresholds lead people to jump to conclusions and are associated with superstitious and conspiracy-related beliefs. To shed light on the role of the decision threshold in the development of these phenomena, it is crucial to study decision thresholds at an early age. However, current measures of decision thresholds have only been administered in adult samples and are based on subjective probability estimates. The current study validates a new instrument that was adapted to measure the decision thresholds of children and is based on the objective probability values of a situation by using a signal detection analytical approach. We assessed decision thresholds in a sample of $N = 299$ children from an extracurricular enrichment program for talented primary school students. Receiver operating characteristic curves indicated that the new measurement instrument has high accuracy. Furthermore, its construct validity was supported by positive associations with children's performance in the original beads task, a measure of confirmation bias, age, grade level, and math performance. However, no associations were found with a measure of alternation bias, epistemic certainty beliefs, self-concept in data-related tasks, general self-efficacy, or gender. This study introduces a new instrument for measuring the decision threshold in children using objective probability values and gives insights into decision-making processes and their implications for cognitive biases and related constructs.

Keywords: decision threshold, beads task, jumping to conclusions bias, signal detection theory, cognitive bias

Assessing Decision Thresholds in Primary School Students Using Signal Detection Theory: Validating an Adapted Version of the Beads Task

Decision-making under uncertainty is a common experience in everyday life. There are many situations in which a person has to base a decision on only ambiguous evidence. In such situations, it is common for people to jump to conclusions, which may lead to false beliefs such as superstitious or conspiracy-related beliefs (Georgiou et al., 2021; Kuhn et al., 2022; Sanchez & Dunning, 2021). For instance, a person could hear about a politician who has been accused of misconduct and immediately assume the politician's guilt or innocence without examining the evidence or considering alternative explanations. Such jumping to conclusions could also be directed against whole groups (e.g., refugees) and may result in prejudice. In these times of fake news and artificial intelligence, it is crucial for people to be more thoughtful about developing their beliefs and any corresponding behaviors.

The level of certainty a person needs to make a decision can be described as a decision threshold. More specifically, the decision threshold is the minimum probability an individual needs to feel confident enough to make a decision given the available information (Moritz et al., 2006). Moritz et al. found that healthy individuals need between around 81% and 93% certainty to make a decision when confronted with probabilistic reasoning tasks (Moritz et al., 2006; Moritz et al., 2007; Moritz et al, 2016; Moritz et al., 2020; Veckenstedt et al., 2011). However, in these studies, the observed decision thresholds were based solely on subjective probabilities estimated by the participants, and they must therefore be deemed prone to error (Moritz et al., 2007). Moreover, the available studies have involved only adult participants and have focused on clinical settings. Whereas it has been shown that fostering more cautious decision-making in younger individuals (e.g., primary school children) is possible (Stark et al., 2025), measurement instruments that are sufficient for empirically determining decision thresholds specifically in children in the context of statistics education have yet to be developed.

With the present study, we aimed to validate a newly adapted version of *the beads task* (Dudley et al., 1997a, 1997b; Garety et al., 2005; Huq et al., 1988), namely, the *decision threshold beads task* (DTBT), which can determine primary school children's decision thresholds on the basis of objective probability values by using a signal detection analytical approach (Green & Swets, 1966).

The Role of Jumping to Conclusions in Decision-Making

The Jumping to Conclusions Bias and Related Constructs

The concept of the decision threshold originates from research on the *jumping to conclusions* (JTC) *bias*. This cognitive bias is defined as making decisions based on a feeling of certainty, even though the available evidence is not strong enough to justify the decision (Garety et al., 2005). Previous research on the JTC bias has predominantly focused on clinical settings, especially in patients with schizophrenia (e.g., Garety et al., 2005; Huq et al., 1988). Meta-analytical evidence has consistently revealed that the JTC bias is positively related to delusional beliefs in patients with schizophrenia (g = 0.53, Dudley et al., 2015; g = 0.28, McLean et al., 2016; r = .10, Ross et al., 2015). However, the JTC bias is not limited to clinical populations; in fact, it has also been found in nonclinical groups (Freeman et al. 2008; Ross et al., 2015) and has been found to be associated with superstitious or conspiracy-related beliefs (Georgiou et al., 2021; Kuhn et al., 2022; Sanchez & Dunning, 2021).

Although the JTC bias has been studied broadly, research has yet to identify the factors that contribute to its development. There are several factors that have been found to be related to the JTC bias or for which it makes theoretical sense for them to be connected to it in young children in the context of statistics education. We address some of these cognitive biases, cognitive outcome variables, motivational constructs, and demographic variables next.

Cognitive biases such as the JTC bias can be viewed as misunderstandings of statistical concepts (e.g., Ben-Zvi & Garfield, 2008), with the JTC bias being a manifestation of the failure to account for variability. Interestingly, Moritz et al. (2016) even stated that "patients [with schizophrenia] may reason like 'bad statisticians'" (p. 93), implying that the JTC bias may arise from an inadequate understanding of a situation's statistical properties. Other cognitive biases are related to the JTC bias, perhaps because they also represent misunderstandings of statistical concepts. For instance, there is evidence that the JTC bias is related to the bias against disconfirming evidence (Balzan et al., 2013; McLean et al., 2016; Moritz et al., 2020; Veckenstedt et al., 2011). Therefore, it might also be connected to other biases such as the confirmation bias and the alternation bias (Kahneman & Tversky, 1972; Tversky & Kahneman, 1971). Whereas the JTC bias results in quick, often unfounded conclusions (Fine et al., 2007), the confirmation bias and the bias against disconfirming evidence help sustain these conclusions by selectively focusing on information that supports them (Balzan et al., 2013; Gagliardi, 2023). The alternation bias describes a person's tendency to expect more alternations between random events than probability laws imply (Kahneman &

Tversky, 1972; Tversky & Kahneman, 1971). Just like the JTC bias, the alternation bias might be related to a person's understanding of the concept of variability. This view that cognitive biases such as the JTC bias are related to misunderstandings of statistical concepts found support from a study in which an extracurricular intervention for primary school children in the context of statistics education significantly reduced the JTC bias (Stark et al., 2025).

The JTC bias can also be viewed as overconfidence when insufficient evidence is given. Young children show a strong tendency to be overconfident (e.g., Lipko et al., 2009) and frequently overestimate their own performances (Finn & Metacalfe, 2014; Goecke et al., 2022). Thus, the JTC bias might be stronger in younger age groups and lower grade levels. Also, younger children tend to see knowledge as certain (Conley et al., 2004). Thus, the JTC bias might also be related to less sophisticated epistemic certainty beliefs.

Another potential factor contributing to the JTC bias is intelligence, as there is evidence that IQ is negatively associated with the JTC bias in both adults (e.g., Freeman et al., 2008; Ross et al., 2016) and children (Gregersen et al., 2022). Similarly, McLean et al. (2020) found that individuals who are better at understanding and using odds are less likely to jump to conclusions. Furthermore, mathematics competency is related to higher statistical literacy (Lai et al., 2011). As the JTC bias can be viewed as a misunderstanding of statistical concepts, this bias might also be related to lower mathematics competency. This relationship could imply that individuals who tend to jump to conclusions might lack the mathematical knowledge needed to process and interpret probabilities effectively.

The fact that Möller et al. (2020) showed a link between self-concept in mathematics and achievement raises the question of whether a similar relationship exists between self-concept in data-related tasks and the JTC bias. Whereas there is no direct support for this idea yet, it warrants investigation because higher confidence in one's own conclusions could be related to a higher JTC bias. Interestingly, Moritz et al. (2020) found no significant correlation between the JTC bias and general self-esteem, suggesting that the bias might not be related to broad self-concept measures. Thus, it is possible that the JTC bias is more closely connected to domain-specific motivational constructs, such as statistics-related self-concept, which could offer an important direction for future research.

Lastly, there is conflicting evidence on whether there are gender differences in overconfidence (Bandiera et al., 2022). To our knowledge, no research has specifically focused on gender and the JTC bias. Thus, whether there is a relationship between the two remains unknown.

Measuring the JTC Bias

The *beads task* (Dudley et al., 1997a, 1997b; Garety et al., 2005; Huq et al., 1988) is widely used for measuring the JTC bias. In this probabilistic reasoning task, participants draw beads from one of two jars with different proportions of colored beads without knowing which jar they are drawing from. For example, one jar may contain 85 orange and 15 black beads, whereas the other jar contains 15 orange and 85 black beads. After each bead is drawn, participants have to decide whether they are certain enough to identify which of the two jars the beads were drawn from—or they can draw an additional bead. If they decide to draw an additional bead, the previously drawn bead is returned to the jar. They can draw up to 20 beads. Although the sequence of the beads is described as random to the participants as a cover story, it is held constant across participants. The most commonly used metric in this task is called *draws to decision* (e.g., Garety et al., 2005), which refers to the number of beads a person draws before deciding. Garety et al. (2005) interpreted a higher number of draws to decision as a more cautious decision-making process.

Even though the beads task has proven useful for researching the JTC bias, it has several shortcomings. First, the JTC bias is often assessed with only a single item, which can lead to low reliability estimates (Moritz et al., 2012). For instance, Moritz et al. (2013, 2017) found that the 4-week test-retest reliability of the classical beads task amounted to only $r = .20$. Second, more draws to decision do not always indicate a higher probability of being correct. For instance, three orange beads provide stronger evidence for the predominantly orange jar than three orange beads and one black bead. Finally, whereas more draws to decision are typically seen as indicating more cautious and desirable decision-making, there is also the risk that excessive draws reflect an overly cautious, inefficient decision process.

The Decision Threshold.

To address the shortcomings of the beads task in measuring the JTC bias, Moritz et al. (2006) introduced the decision threshold. The decision threshold is the minimum probability required for a person to feel confident enough to make a decision in a probabilistic reasoning task. For instance, a person with a decision threshold of 95% would accept hypotheses as true only when they perceived at least a 95% certainty that it was true. As the JTC bias has mostly been studied in adult patients and not in children, we present findings from the adult patient group below. In their study, Moritz et al. (2006) presented 20 multiple-choice knowledge questions with four response alternatives in the style of the *Who Wants to Be a Millionaire* television game show to 32 patients with schizophrenia and 38 healthy controls. For each

question, participants were asked to rate the plausibility of each possible answer with values ranging from 0% to 100% (in steps of 10%). Participants were also given the option to rate the answer alternatives as true or false. Moritz et al. (2006) found that patients with schizophrenia rated response alternatives as true based on significantly lower probability ratings (85.9%) than healthy controls did (92.7%). Consecutive studies also found confirming evidence of a lower decision threshold in patients with schizophrenia (71% – 85%) than healthy controls (81% – 93%; Moritz et al., 2007; Moritz et al., 2016; Moritz et al., 2020; Veckenstedt et al., 2011). Thus, the authors concluded that a lower decision threshold speaks of a more pronounced liberal acceptance of hypotheses. However, these estimates of the decision threshold always relied on the participants' subjectively rated probability values. These subjective estimations have been found to underestimate the objective probabilities in the beads task (Moritz et al., 2007). Using objective probability values could make the measurements of the decision threshold more accurate. To do so, in this study, we expanded the conventional beads task to derive decision thresholds using a signal detection analytical approach. In this approach, multiple items with different objective probability values can be used to determine the threshold below which a person does not make a decision but above which they do make a decision.

Expanding the Beads Task: The Decision Threshold Beads Task.

To our knowledge, no studies have measured the decision threshold in children. However, this age group is particularly relevant for studying the development of decision-making processes and the JTC bias at an early stage. As the JTC bias is related to superstitious beliefs, conspiracy-related beliefs, and mental health (Dudley et al., 2015; Georgiou et al., 2021; Kuhn et al., 2022; McLean et al., 2016; Ross et al., 2015; Sanchez & Dunning, 2021), studying the decision threshold in children may provide information on the development of the JTC bias and could add to knowledge of how to design early interventions to reduce the JTC bias. Such interventions have been called for (Gregersen et al., 2022).

For the present study, we adapted the beads task (Dudley et al., 1997a, 1997b; Garety et al., 2005; Huq et al., 1988) to measure the decision threshold with objective probabilities that were specifically tailored for primary school children. We named it the *decision threshold beads task* (DTBT). As in the original beads task, children have to decide whether a certain collection of beads have been drawn from one of two jars. Instead of letting the participants draw the beads themselves, we present items in which fictional participants already drew a certain number of beads. By doing so, we could control the probabilities for each item.

We used van der Leer et al.'s (2015) formula to calculate the probabilities:

$$\Pr\big((\mathrm{n}_a, \mathrm{n}_b)\big|\mathrm{A}\big) = \frac{1}{1+\rho^{n_b-n_a}}\text{, where } \rho = \frac{p}{1-p} > 1 \quad (1)$$

In this formula, the probability of jar A is calculated as depending on the number of beads with the more common color in that jar (n_a), the number of beads with the less common color in that jar (n_b), and the probability of drawing a bead with the more common color in that jar (p). In this way, we could calculate the probabilities for the more probable jar for all items. When choosing the more probable jar, the probabilities of being right ranged from 50% to 99.9%. For instance, for Item 1, the probability of being right is 85%; for Item 7, it is 99.5%; and for Item 12, it is 70%. By having many different items with different probabilities, we could use a signal detection analytical approach to assess each individual's decision threshold.

Signal detection theory (SDT; Green & Swets, 1966; Stanislaw & Todorov, 1999) is a framework that is used to analyze decision-making under uncertainty. It distinguishes between a signal (a meaningful stimulus) and noise (random background stimuli) and helps determine the ability to detect signals amidst noise. In our case, the signal is the certainty of having identified the correct jar. The stronger the evidence (probability) of being right, the stronger the signal. With the help of SDT, we can then calculate the response bias, which is the tendency to lean toward seeing the signal as present or absent. In our case, we need to calculate the decision threshold, which describes the probability that needs to be met for a person to be confident enough to make a decision.

Aims of The Present Study

In the present study, we evaluated the adapted version of the beads task (i.e., the DTBT) to measure the decision threshold. As current measurement instruments of the decision threshold rely on subjective probabilities, we wanted to add an approach that relied on objective probabilities to determine decision thresholds to develop a more accurate measurement instrument, using the beads task as the basis.

For Research Question 1, we first investigated whether our new measurement instrument was suitable for measuring decision thresholds in primary school children. To this end, we tested the accuracy of the measurement instrument by using an SDT approach. Additionally, we checked whether the resulting values of the decision threshold were plausible by comparing them with the range we expected them to be in and by checking overall whether higher objective probabilities resulted in more decision-making.

For Research Question 2, we examined the construct validity of the decision threshold. We investigated the correlations with the measure of draws to decision. The decision threshold should be positively related to draws to decision, as they are both indicators of the JTC bias. We expected partial convergence between the two constructs because they both assess the JTC bias but from different perspectives. Therefore, their correlation should be between .40 and .70 (Carlson & Herdman, 2012). Additionally, we investigated the correlations with other related constructs. We expected the decision threshold to be related to other variables to a similar but not identical degree to what previous research found for their relationships with the JTC bias. For variables that have not yet been research in connection to the JTC bias, we expected them to be related to the decision threshold as argued in the theoretical background. More specifically, we expected that the decision threshold would be negatively related to the confirmation bias and the alternation bias. We expected that the decision threshold would be positively related to more sophisticated epistemic certainty beliefs, math grade, self-concept in data-related tasks, age, and grade level. Finally, we did not expect the decision threshold to be correlated with general self-efficacy or gender.

The research objectives of the present study were not preregistered. All necessary materials, data, and analysis scripts are available on the Open Science Framework (OSF, https://osf.io/b6kfj/).

Method

Participants and Procedure

The participants in this study were primary school children enrolled in the Hector Children's Academy Program, an extracurricular enrichment program for talented primary school children in Baden-Württemberg, Germany (for more information, see Trautwein et al., 2023). Children were recruited via the program's online learning platform, where children could access various educational activities. To incentivize participation, the children were informed they could win one of 15 card games valued at around 10€ each. The survey began with a child-friendly introduction. In the story, the website's mascot *Hasel* went to a friend's house and found a note saying that their friend had gone to the forest and Hasel could join them. By assessing the numbers of hazelnuts and walnuts left behind by the friend, Hasel had to answer the question of whether he had enough information to decide whether his friend had gone to the forest with more hazelnut trees or the one with more walnut trees. This problem introduced the children to the topic of decision-making under uncertainty.

Data collection took place from December 12, 2022, to January 13, 2023. A total of N = 299 children from first to fourth grade voluntarily participated in the study with informed consent from a parent or legal guardian. The sample included children from diverse backgrounds, representing different age groups (M = 8.69; SD = 1.04; Range: 6.30–12.13), grade levels (n_{first} = 12; n_{second} = 112; n_{third} = 92, n_{fourth} = 76; $n_{missing}$ = 7), gender groups (55.18% boys, 44.48% girls; 0.33% missing), and mother tongues (83.61% only German; 11.37% German and another language; 5.01% another language).

This study was approved by the Ethics Committee of the Faculty of Economics and Social Sciences at the University of Tübingen (A2.5.4-252_bi).

Measures

Beads Task

The JTC bias was assessed using the 85:15 and 60:40 versions of the beads task, as they are the most common versions that are tested (Dudley et al., 1997a, 1997b; Garety et al., 2005). In each item, children saw two different jars: one with 85 orange beads and 15 black beads and one with 15 orange beads and 85 black beads (or 60:40 and 40:60 respectively; for an example item, see Garety et al., 2011). They were told that one of the two jars was chosen at random and that beads were then drawn at random with replacement from that jar. The order of the beads was predetermined. After each draw, the children could decide whether they wanted to see another bead or if they already knew which jar the beads were drawn from. All drawn beads were visible at any time to aid memory. Draws to decision was then used as the outcome measure. Fewer draws suggest a greater JTC bias. We created a mean value for each participant. Additionally, we calculated the correlation between the two items to judge the beads task's reliability (r = .56, p < .001; α = .72).

Decision Threshold Beads Task

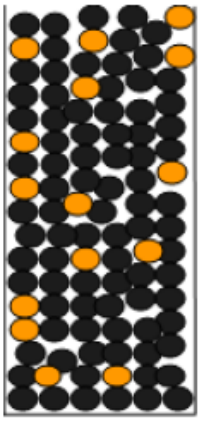
We adapted the original beads task (Dudley et al., 1997a, 1997b; Garety et al., 2005) to assess the decision threshold on the basis of objective probabilities with a signal detection analytical approach. We developed three sets of eight items each. In each set, two jars with different ratios of beads (85:15 and 15:85; 70:30 and 30:70; 60:40 and 40:60) were shown (see Figure 1). The participants were told that each child from a group of children randomly drew beads from one of the jars. Each item varied in how many beads were drawn and which proportion of colors they had, resulting in different probabilities of being right for each item. For each item, the participating children had to decide whether they were sure enough that they

Figure 1

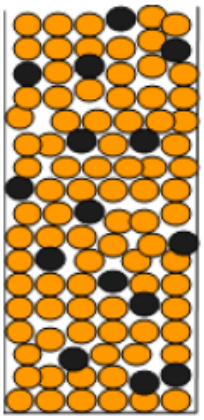
Sample Items From the Decision Threshold Beads Task

Set 1

Each child in a class draws beads at random from one of two jars. After drawing a bead, they put it back and draw another bead from the same jar until they want to stop. There are 85 black and 15 orange beads in the first jar. In the second jar there are 15 black and 85 orange beads. Can you find out for each child whether you can tell which jar they have drawn from? You decide for yourself when you are sure enough.



Jar 1
(85 black; 15 orange)



Jar 2
(15 black; 85 orange)

	Drawn beads:	I am sure enough: Jar 1.	I am not sure enough.	I am sure enough: Jar 2.
Alex:	●	☐	☐	☐
Bettina:	●●●	☐	☐	☐
Christine:	●●●●●●●	☐	☐	☐
Dennis:	●●●●●●●	☐	☐	☐
Erik:	●●●	☐	☐	☐
Florian:	●●	☐	☐	☐
Gabriel:	●●●●●●●●●●●	☐	☐	☐
Inge:	●●●●●●●●●●●	☐	☐	☐

could say which jar the beads were drawn from and indicate that jar or whether they were not sure enough. There were no incentives to make the children's behavior tend toward deciding or not deciding. Therefore, the instrument measures the decision threshold on the basis of the

children's personal reasons for making the decisions. For example, for the first item, Alex drew one black bead, resulting in an 85% probability for the mainly black jar being right. For the fifth item, Erik drew three black beads. The probability that the beads were drawn from the primarily black jar was 99.5%. Therefore, a person with a 95% decision threshold should not be certain enough for the first item to decide which jar the bead came from, whereas they should be certain enough for the fifth item to choose the primarily black jar. Further information on the calculation of the decision threshold is presented in the results section.

Confirmation Bias

We assessed children's confirmation bias with six items from a self-assessment questionnaire (e.g., "My first impression usually seems to be correct"; Rassin, 2008). The items were rated on a scale ranging from 1 (*that's not true at all*) to 5 (*that's exactly true*). Internal consistency was low ($\alpha = .50$).

Alternation Bias

We assessed the alternation bias with a self-developed instrument. Participants were told to create three sequences of 20 coin-throw results that looked random to them. Therefore, values could range from 0 (no alternations) to 19 (only alternations). For example, the sequence HTTTHTHHHTHTHTHTTHHT has 13 alternations. The closer the children were to a mean value of 9.5 alternations across all three items, the lower the alternation bias. Internal consistency was low ($\alpha = .52$).

Epistemic Certainty Beliefs

We assessed children's epistemic certainty beliefs with nine adapted items (e.g., "If you have found something out, it is certainly true"; Conley et al., 2004; Urhahne & Hopf, 2004). The items were rated on a scale ranging from 1 (*that's not true at all*) to 5 (*that's exactly true*). Internal consistency was low ($\alpha = .52$).

Self-Reported Mathematics Grade

Math performance was assessed by asking the participants "What grade did you get in math on your last report card?" Answer options ranged from 1 (*the highest grade*) to 6 (*the lowest grade*). For ease of interpretation, grades were reverse-scored so that higher grades represented higher levels of achievement. Also, it was possible to select "I haven't yet gotten a math grade" or "I don't know anymore," both of which were coded as missing.

Self-Concept in Data-Related Tasks

We assessed children's domain-specific self-concept with six items (e.g., "Anything that has to do with data is easy for me"; Arens et al., 2011; Gaspard et al., 2015). The items were rated on a scale ranging from 1 (*that's not true at all*) to 5 (*that's exactly true*). Internal consistency was low (α = .47).

General Self-Efficacy

We assessed general self-efficacy with the short form of the self-efficacy scale from Beierlein et al. (2012). In three items, children were asked to rate statements (e.g., "I can rely on my skills in difficult situations") on a scale ranging from 1 (*that's not true at all*) to 5 (*that's exactly true*). Internal consistency was acceptable (α = .65).

Descriptive statistics for all variables are presented in Table A1 in the Appendix.

Statistical Analyses

All analyses were computed in R (version 4.4.1, R Core Team, 2024). As a first analysis, we calculated the descriptive statistics for all DTBT items.

For Research Question 1, we calculated the decision threshold using a signal detection analytical approach with the *psycho* package (Makowski, 2018). We determined hits, false alarms, misses, and correct rejections on the basis of whether the children decided on a jar or were not sure enough and whether the objective probability was below or above the mean of all items' objective probabilities (79.5%). For example, choosing the correct jar on an item with a probability of 85% would be a hit, whereas not deciding or choosing the wrong jar would be a miss. To compute the response bias *c*, we used the following formula by Macmillan (1993):

$$c = -1 * \frac{\phi^{-1}(\mathrm{H}) + \phi^{-1}(\mathrm{F})}{2} \quad (2)$$

In this formula, the inverse phi function (ϕ^{-1}) converts probabilities into z scores. Therefore, the z score for the hit rate and the z score for the false alarm rate are added up and averaged by dividing the sum by 2. Then this average is multiplied by -1, so negative values of *c* reflect a bias toward deciding early and positive values reflect a bias toward not deciding. To calculate the decision threshold (*dt*), we multiplied the response bias (*c*) by the standard deviation of all items' objective probabilities (17.4%) and added this product to the mean probability (79.5%). This approach ensures that *dt* integrates both the individual's tendency to decide early or late (as captured by *c*) and the variability in the objective probabilities of the items (standard deviation). The mean anchors the threshold in the central tendency of the

distribution, whereas the standard deviation scales the bias to account for the spread of the probabilities. This combination provides a context-sensitive threshold that reflects both the individual's response tendencies and the characteristics of the decision environment:

$$dt = c * 0.174 + 0.795 \quad (3)$$

Additionally, we calculated the decision threshold using a range of cut-off values rather than relying solely on the mean probability (79.5%). This approach was taken to examine whether the calculated decision threshold values depended on the placement of the cut-off value separating signal and noise trials. By systematically varying the cut-off values from 55% to 95% in increments of 5%, we aimed to test the sensitivity of the decision threshold to different cut-off values for signal versus noise. This approach allowed us to account for the possibility that the specific choice of a cut-off value (e.g., the mean) might introduce bias or limit generalizability, as the distribution of probabilities in the data set might interact with the decision-making process differently at various cut-off values.

Furthermore, to follow Paulhus and Petrusic's (2010) and Goecke et al.'s (2022) approach, we calculated an aggregated mean decision threshold across all these decision threshold values to derive a more robust and stable estimate of the decision threshold. This robust estimate minimizes the potential for the impact of any individual threshold value to skew the results and reflects an averaged, context-independent decision threshold. To maintain the interpretability and validity of the decision threshold values, any resulting values exceeding 100% (which are not feasible probabilities) were capped at 100%. This capping ensured that all values would remain within the meaningful probability range. Lastly, there were no missing values in the calculation of the DTBT, ensuring the completeness and reliability of the data.

To evaluate whether the DTBT accurately measured the decision threshold, we analyzed its performance using receiver operating characteristic (ROC) curves (Stanislaw & Todorov, 1999). ROC analysis is a standard method in SDT that illustrates the trade-off between sensitivity (the ability to correctly identify true positives) and specificity (the ability to correctly reject false positives) across a range of decision thresholds. By plotting the hit rate (proportion of correct decisions in signal trials) against the false alarm rate (proportion of incorrect decisions in noise trials) for different cut-off values, we gained insight into how effectively the DTBT is able to discriminate between true signals and noise.

To summarize the discriminative accuracy of the DTBT, we calculated the area under the curve (AUC) of the ROC curves. The AUC is a robust and widely accepted measure that captures the overall performance of a decision-making tool. AUC values range from 0.5 (indicating chance-level discrimination, equivalent to random guessing) to 1.0 (indicating

perfect discrimination). This metric provides a single, interpreTable Acore that reflects the DTBT's ability to differentiate between trials with true signals and noise trials, offering a quantitative evaluation of its accuracy (Stanislaw & Todorov, 1999).

Additionally, to assess the plausibility of the children's responses, we examined whether the range of values produced by the DTBT was realistic and consistent with expectations (i.e., between around 50% to 100%). We also calculated the correlation between the objective probability of each item and the percentage of children who were confident enough to make a decision. This correlation provided further validation by testing whether the children's decisions were aligned with the probabilistic structure of the items, thereby supporting the plausibility and validity of the DTBT as a measure of decision thresholds.

For Research Question 2, we wanted to assess the relationships between the decision threshold, draws to decision, and additional variables to assess construct validity. Therefore, we calculated Pearson's correlations between all variables with the R package *Hmisc* (Harrell, 2024). Because there were missing values on most variables, we calculated pairwise complete correlations.

Results

Item Descriptives

Table 2 presents descriptive statistics for all DTBT items. The mean value describes the percentage of children who were certain enough to choose the more probable jar (1) and how many were not able to decide or chose the less probable jar (0). The correlation between the mean value and the objective probability was $r = .86$ (95% CI [.71, .94], $p < .001$). This finding indicates that the children's decisions were plausible and rational, that is, as the objective probability increased, so did their certainty in deciding.

Decision Threshold Distributions

The goal of this analysis was to examine whether the calculation of decision thresholds depended on the specific cut-off values used to separate hits and false alarms from misses and correct rejections. The challenge arises because the placement of this cut-off value, such as using the mean of all items' objective probabilities (79.5%), may influence the resulting decision thresholds, potentially introducing bias or limiting the generalizability of the findings. To address this issue, we calculated decision thresholds using multiple cut-off values to test the robustness of the DTBT. Specifically, we first calculated one decision threshold using the

Table 2

Descriptive Statistics for DTBT Items

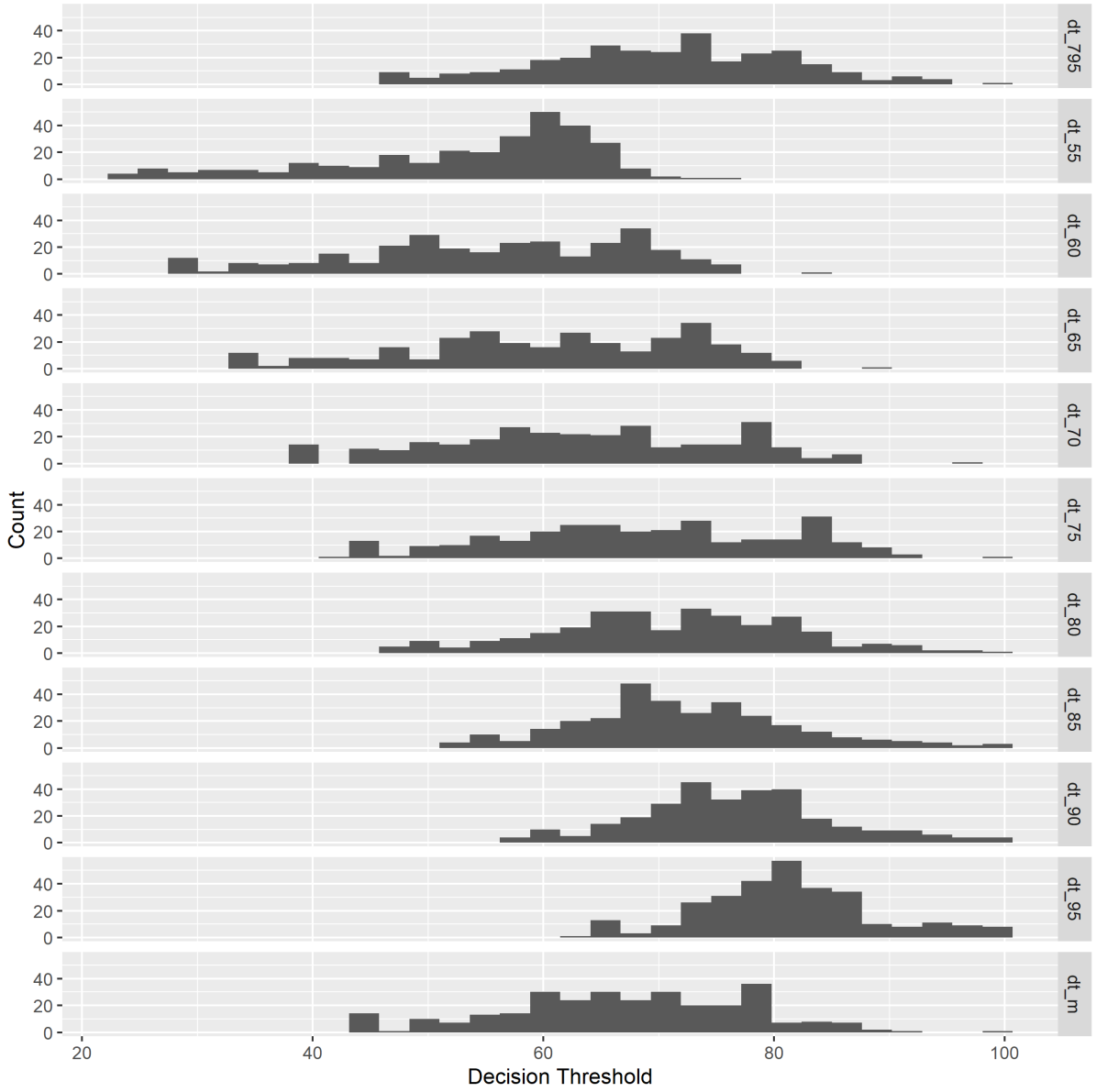
Item	Jars	Number of Beads	Beads Ratio	Objective Probability	*M*	*SD*
1	85:15	1	1:0	.850	0.24	0.43
2	85:15	3	2:1	.850	0.66	0.47
3	85:15	7	6:1	.999	0.93	0.26
4	85:15	7	4:3	.850	0.58	0.49
5	85:15	3	3:0	.995	0.91	0.29
6	85:15	2	1:1	.500	0.21	0.41
7	85:15	11	7:4	.995	0.92	0.27
8	85:15	11	8:3	.999	0.95	0.22
9	70:30	7	4:3	.700	0.56	0.50
10	70:30	7	6:1	.986	0.95	0.21
11	70:30	11	8:3	.986	0.93	0.26
12	70:30	3	2:1	.700	0.45	0.50
13	70:30	11	7:4	.927	0.89	0.31
14	70:30	4	2:2	.500	0.20	0.40
15	70:30	3	3:0	.927	0.90	0.30
16	70:30	1	1:0	.700	0.30	0.46
17	60:40	1	1:0	.600	0.25	0.43
18	60:40	3	3:0	.771	0.78	0.42
19	60:40	11	7:4	.771	0.87	0.34
20	60:40	3	2:1	.600	0.45	0.50
21	60:40	7	4:3	.600	0.49	0.50
22	60:40	7	6:1	.884	0.95	0.21
23	60:40	11	8:3	.884	0.95	0.23
24	60:40	6	3:3	.500	0.16	0.37

Note. All items were answered by all $N = 299$ children.

mean of all objective probabilities (79.5%) as the dividing line. Additionally, we calculated nine more decision thresholds using cut-off values ranging from 55% to 95% in increments of 5%. This approach allowed us to assess how different placements of the signal-noise threshold affected the resulting decision thresholds. Finally, to create a robust overall estimate, we calculated the mean decision threshold across these values by employing the method used by Paulhus and Petrusic (2010) and Goecke et al. (2022). By analyzing these distributions (see Figure 2), we could better understand the results of the decision thresholds across varying cut-off values. This step ensured that our conclusions were not overly dependent on the arbitrary choice of a single cut-off value and thereby provided a comprehensive and nuanced assessment of the decision threshold.

Figure 2

Histograms of Decision Thresholds Based on Cut-Off Values Between Signal and Noise



Note. dt_795 describes the decision threshold with a cut-off value of 79.5%. dt_55 to dt_95 describe the decision thresholds with cut-off values from 55% to 95%. dt_m describes the mean decision threshold for all nine decision thresholds with cut-off values from 55% to 95%.

The decision thresholds exhibited strong correlations with one another, with all correlation coefficients exceeding .85 (see Figure A2 in the Appendix). Thus, the primary differences between these thresholds lie in the range of values they represent. As illustrated in Figure 2, the distributions of the decision thresholds varied considerably. For instance, the distribution of the decision threshold set at the cut-off value of 55% shows a range of values

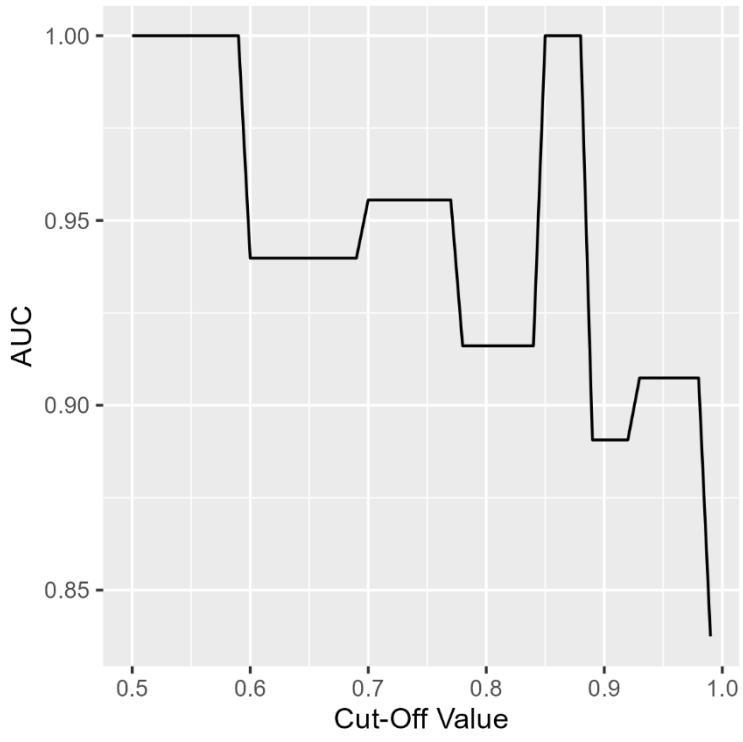
spanning from approximately 20% to 80%. These values are implausible given that the objective probabilities for all items ranged from 50% to 99.9%. By contrast, the mean decision threshold provides more reasonable results, with values ranging from just below 50% to 100%. Consequently, we opted for the mean decision threshold as our final metric, as it can be assumed to yield more robust results (M = 67.02%, SD = 10.47%; similar to Goecke et al., 2022; Paulhus & Petrusic, 2010).

ROC Curve Analysis

To assess the accuracy of the measurement instrument, we analyzed ROC curves (Stanislaw & Todorov, 1999). To do so, the hit rate is plotted against the false alarm rate, and the AUC is calculated to judge the instrument's accuracy. We calculated the AUC for all possible cut-off values set from 50% to 99% in steps of 1% (see Figure 3). It is desirable to find consistent results across all different item thresholds with AUCs close to 1.00. All AUCs ranged from .84 to 1.00. Because these values can vary between .50 (chance level) and 1.00 (perfect accuracy; Stanislaw & Todorov, 1999), the accuracy of the DTBT was consistently high, independent of which cut-off value was chosen. Thus, our instrument measured the decision threshold with high accuracy.

Figure 3

Function of All Possible Cut-Off Values on the AUC



Relationship Between Decision Threshold and Related Constructs

Next, we present bivariate correlations for all assessed variables (see Table 3). As expected, the decision threshold and draws to decision were correlated between .40 and .70 (r = .51, p < .001) and therefore showed partial convergence (Carlson & Herdman, 2012). This finding endorses the assumption that the two instruments measure a similar construct from different perspectives.

As expected, confirmation bias was negatively associated with the decision threshold (r = -.24, p < .001) and draws to decision (r = -.14, p = .013). Therefore, children with a higher confirmation bias have a lower decision threshold and fewer draws to decision.

Interestingly, we found a significant positive correlation between the decision threshold and math grade (r = .19, p = .009) but not between draws to decision and math grade (r = .03, p = .692). These findings indicate that children with better math grades have a higher decision threshold but do not differ in their draws to decision.

Age and grade level were also positively correlated with the decision threshold (r = .20, p < .001; r = .17, p = .003) and draws to decision (r = 0.26, p < .001; r = 0.27, p < .001), respectively. In line with our expectations, these findings suggest that older students and those in higher grades tend to take more draws to reach a decision and have a higher decision threshold.

There were no more significant relationships with the decision threshold or draws to decision, indicating that students with different levels of the alternation bias, epistemic certainty beliefs, self-concept in data-related tasks, general self-efficacy, or gender did not differ in their decision thresholds or draws to decision.

Discussion

With the present study, we aimed to introduce and validate a newly adapted measurement instrument for assessing the decision threshold by using objective probability values and signal detection theory with primary school children. The accuracy of the measurement instrument was demonstrated by the AUCs of the ROC curves, which were high with values between .84 and 1.00. Also, the resulting decision threshold values were plausible in their range and in the fact that we found more decision-making for items with higher certainty. The construct validity of the new task was largely demonstrated. As expected, it was positively correlated with the beads task with an effect size between .40 and .70 and therefore demonstrated partial convergence (Carlson & Herdman, 2012). Moreover, the decision

Table 3

Correlations Between Decision Threshold, Draws to Decision, and Covariates

Variable	1	2	3	4	5	6	7	8	9	10
1. Decision Threshold										
2. Draws to Decision	.51***									
3. Confirmation Bias	-.24***	-.14*								
4. Alternation Bias	-.06	-.02	.01							
5. Epistemic Certainty Beliefs	-.10	-.04	-.06	.00						
6. Math Grade	.19**	.03	-.02	.07	-.09					
7. Self-concept in Data-related Tasks	.06	.05	.20***	.01	-.10	.04				
8. General Self-efficacy	.00	.00	.33***	.02	-.04	.02	.33***			
9. Gender (1 = boys)	.06	.06	-.11	.03	.06	.03	.14*	.01		
10. Age	.20***	.26***	-.05	.04	-.07	-.04	.02	-.03	.03	
11. Grade Level	.17**	.27***	-.07	.04	-.12	-.09	.05	-.01	.02	.88***

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

threshold was significantly positively associated with age, grade level, and math grade and significantly negatively associated with confirmation bias. It was not associated with self-concept in data-related tasks, general self-efficacy, epistemic certainty beliefs, alternation bias, or gender.

Decision Threshold Beads Task Versus Original Beads Task

There are some similarities between the DTBT and the original beads task, but there are also clear differences. Both instruments revealed that the children in this study were rather liberal in their acceptance of hypotheses. The DTBT measured a mean decision threshold of M = 67.02% for all participating children, a level that is far below what healthy adults show (81% – 93%, Moritz et al., 2007; Moritz et al, 2016; Moritz et al., 2020; Veckenstedt et al., 2011). The original beads task measured a mean of M = 5.32 draws to decision in the participating children. This finding is similar to the lower end of healthy adult populations that have shown between 2.60 and 11.19 draws to decision (e.g., Garety et al., 1991; Huq et al., 1988; So & Kwok, 2015; Ward et al., 2018).

The DTBT might be more reliable than the original beads task because the DTBT is assessed with 24 items instead of only two items or, in many cases, only one item (Moritz et al., 2012). In a previous study, Moritz et al. (2013, 2017) found a 4-week test-retest reliability for the classical beads task of only r = .20. However, the test-retest reliability of the DTBT has yet to be determined and should be a topic of future research.

The draws to decision measure from the original beads task can be used to show the JTC bias when participants draw only one or a few beads. More draws to decision are usually considered to reflect a smaller JTC bias. However, there is no clear rule on when participants might be too cautious. The decision threshold from the DTBT can be interpreted more easily, as the certainty needed is depicted as a probability value. From the perspective of science, a decision threshold of 95% might be appropriate, as the common alpha level is 5%. However, the ability to determine which probability values are appropriate might depend on the situation and the person. In some situations, people might have more pressure to make a decision, for instance, when there is time pressure. Or they may already have a strong opinion on a topic and may thereby make their decision even when the current evidence suggests more caution. This might be the case when delusional conviction or conspiracy-related beliefs are involved. Future research should investigate which decision thresholds are appropriate or healthy in different situations.

Construct Validity

We found a correlation between the decision threshold and the draws to decision measure. This association was expected because both measures are indicators of the JTC bias. The measures shared around 25% of their variance. This finding shows that the measures are related but also capture distinct aspects of the construct they were designed to measure.

We expected positive relationships between the decision threshold and other cognitive biases, as cognitive biases can be viewed as misunderstandings of statistical concepts (e.g., Ben-Zvi & Garfield, 2008). As expected, the confirmation bias was significantly correlated with the decision threshold and draws to decision. This association seems theoretically logical because the confirmation bias and the JTC bias both reflect overconfidence in one's own judgment. This finding is also in line with the previous finding that the JTC bias is related to the bias against disconfirming evidence (Balzan et al., 2013; McLean et al., 2016; Moritz et al., 2020; Veckenstedt et al., 2011). Whereas the JTC bias reflects an underestimation of variability, the confirmation bias reflects the cherry-picking of data points to reduce variability in favor of pre-existing beliefs. However, the alternation bias was not significantly related to the decision threshold or draws to decision. This bias reflects the expectation of certain patterns in the variability of random data. Thus, not all cognitive biases in this study can be linked to the understanding of variability in the same way. However, the alternation bias might not be connected to other variables because of the low internal consistency of the measurement instrument we developed.

Next, we expected positive relationships between the decision threshold and cognitive outcome variables. There were no significant relationships with epistemic certainty beliefs, even though more sophisticated epistemic certainty beliefs are described by lower beliefs in the certainty of knowledge (Conley et al., 2004). Therefore, we expected a positive relationship with the decision threshold. Maybe epistemic certainty beliefs were not assessed validly, as their internal consistency was rather low. However, a higher decision threshold was associated with better math grades. A previous study had already found a connection between people’s ability to understand and use odds and draws to decision (McLean et al., 2020), and mathematical competency is positively connected to statistical literacy (Lai et al., 2011). Also, in a previous study (Stark et al., 2025), we found that an extracurricular statistics intervention enhanced talented third- and fourth-graders’ draws to decision. Hence, future studies should investigate how the understanding of certain statistical concepts (e.g., the law of large numbers) is linked to the decision threshold.

We expected the decision threshold to be positively related to self-concept in data-related tasks but not to general self-efficacy. However, both of these relationships were nonsignificant. Therefore, the decision threshold is not related to these motivational beliefs. We expected that self-concept in data-related tasks would be related to the decision threshold, as performance and motivation in a domain are usually correlated (Möller et al., 2020). Thus, the decision threshold might not be a good indicator of performance in the domain of statistics. Alternatively, the way self-concept was measured might not have been accurate, as it showed weak reliability, which could mean that the children did not fully understand what data-related tasks involve. Nonetheless, general self-efficacy was not related to the decision threshold as expected because a previous study showed no link between the JTC bias and self-esteem (Moritz et al., 2020).

Lastly, we expected the decision threshold to be related to age and grade level, and we explored its relationship with gender. As age and grade level were positively correlated with the decision threshold, we can conclude that the decision threshold increases during the years children spend in primary school. However, our study was cross-sectional, and evidence from longitudinal studies is needed to support our conclusion. In addition, no significant relationship was found between the decision threshold and gender. This lack of association supports the idea that there are no gender differences in how the decision threshold was assessed or in the decision threshold itself.

Strengths and Limitations

As a first strength, this study extends the understanding of the decision threshold to children. Previous research has focused primarily on adult populations. But it is essential to study children in order to understand how the JTC bias develops and to determine how to implement early interventions, a need highlighted by Gregersen et al. (2022).

Second, our sample of $N = 299$ primary school children was relatively large and included participants with diverse backgrounds. However, these children were also all from an extracurricular enrichment program for talented primary school children in Germany, thereby limiting the generalizability of the findings to other age groups and children with different abilities or from different educational and cultural settings. Whereas the task was designed for children, future research could explore its applicability to other populations, such as adolescents and adults.

Third, using SDT (Green & Swets, 1966) in this study provided an objective framework for measuring decision-making under uncertainty, and such a framework is essential for

accurately assessing the decision threshold in primary school children. Unlike previous approaches that relied on subjective probability estimates, this SDT approach takes objective probability measures into account, potentially enhancing the reliability and validity of the findings. However, the objective probabilities ranged only from 50% to 99.9%. It might be possible that some people's decision thresholds are so far below 50% that the instrument cannot detect them. Nevertheless, in the context of the beads task and when choosing between two possible options, the range from 50% to 99% was plausible.

Fourth, we used a wide range of variables and assessed their relationships with the decision threshold. This approach offered the opportunity to explore the nomological net around the decision threshold. However, several related constructs (confirmation bias, alternation bias, and epistemic certainty beliefs) exhibited low internal consistencies, thus limiting the reliability of these measures and potentially affecting the validity of the conclusions drawn about their relationships with the decision threshold. Also, it would have been useful to assess additional constructs, such as superstitious, conspiracy-related, and delusional beliefs, which have previously been found to be related to the decision threshold and the JTC bias (e.g., Georgiou et al., 2021; Kuhn et al., 2022; Moritz et al., 2017; Sanchez & Dunning, 2020).

Conclusion

In this study, we introduced and validated a new measurement instrument—the DTBT—for assessing the decision threshold in primary school children. It demonstrated strong accuracy and meaningful correlations with key variables. By employing SDT and using objective probabilities as a basis, the DTBT offers a reliable and innovative approach to understanding decision-making processes and the early development of the JTC bias. Whereas limitations such as limited generalizability and low internal consistencies in some constructs remain, they present valuable opportunities for refinement and future research to enhance educational and psychological applications.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process.

During the preparation of this work the authors used ChatGPT (OpenAI, 2024) in order to improve style and grammar of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Arens, A. K., Trautwein, U., & Hasselhorn, M. (2011). Self-Concept Measurement with Preadolescent Children: Validation of a German Version of the SDQ I. *Zeitschrift für Pädagogische Psychologie*, *25*(2), 131-144. https://doi.org/10.1024/1010-0652/a000030

Balzan, R. P., Delfabbro, P. H., Galletly, C. A., & Woodward, T. S. (2013). Confirmation biases across the psychosis continuum: The contribution of hypersalient evidence-hypothesis matches. *British Journal of Clinical Psychology*, *52*(1), 53-69. https://doi.org/10.1111/bjc.12000

Balzan, R. P., Ephraums, R., Delfabbro, P., & Andreou, C. (2017). Beads task vs. box task: The specificity of the jumping to conclusions bias. *Journal of Behavior Therapy and Experimental Psychiatry*, *56*, 42-50. https://doi.org/10.1016/j.jbtep.2016.07.017

Bandiera, O., Parekh, N., Petrongolo, B., & Rao, M. (2022). Men are from mars, and women too: A bayesian meta-analysis of overconfidence experiments. *Economica*, *89*, S38-S70. https://doi.org/10.1111/ecca.12407

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU) [Short scale for measuring general self-efficacy beliefs (ASKU)]. *Methoden, Daten, Analysen*, *7*(2), 251–278, http://doi.org/10.12758/mda.2013.014.

Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. *School Science and Mathematics*, *108*(8), 355-361. https://doi.org/10.1111/j.1949-8594.2008.tb17850.x

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, *15*(1), 17–32. https://doi.org/10.1177/1094428110392383

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, *29*(2), 186-204. https://doi.org/10.1016/j.cedpsych.2004.01.004

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997a). The effect of self-referent material on the reasoning of people with delusions. *British Journal of Clinical Psychology*, *36*, 575–584. https://doi.org/10.1111/j.2044-8260.1997.tb01262.x

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997b). Normal and abnormal reasoning in people with delusions. *British Journal of Clinical Psychology*, *36*, 243–258. https://doi.org/10.1111/j.2044-8260.1997.tb01410.x

Dudley, R., Taylor, P., Wickham, S., & Hutton, P. (2015). Psychosis, delusions and the "jumping to conclusions" reasoning bias: A systematic review and meta-analysis. *Schizophrenia Bulletin*, *42*(3), 652-665. https://doi.org/10.1093/schbul/sbv150

Fine, C., Gardner, M., Craigie, J., & Gold, I. (2007). Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry*, *12*(1), 46-77. https://doi.org/10.1080/13546800600750597

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction*, *32*,1–9. https://doi. org/10.1016/j.learninstruc.2014.01.001

Freeman, D., Pugh, K., & Garety, P. (2008). Jumping to conclusions and paranoid ideation in the general population. *Schizophrenia Research*, *102*(1-3), 254-260. https://doi.org/10.1016/j.schres.2008.03.020

Gagliardi, L. (2023). The role of cognitive biases in conspiracy beliefs: A literature review. *Journal of Economic Surveys*, 1–34. https://doi.org/10.1111/joes.12604

Garety, P. A., Freeman, D., Jolley, S., Dunn, G., Bebbington, P. E., Fowler, D. G., ... & Dudley, R. (2005). Reasoning, emotions, and delusional conviction in psychosis. *Journal of Abnormal Psychology*, 114(3), 373. https://doi.org/10.1037/0021-843X.114.3.373

Garety, P., Freeman, D., Jolley, S., Ross, K., Waller, H., & Dunn, G. (2011). Jumping to conclusions: The psychology of delusional reasoning. *Advances in Psychiatric Treatment*, *17*(5), 332-339. https://doi.org/10.1192/apt.bp.109.007104

Gaspard, H., Dicke, A. L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, *51*(9), 1226. https://doi.org/10.1037/dev0000028

Georgiou, N., Delfabbro, P., & Balzan, R. (2021). Conspiracy theory beliefs, scientific reasoning and the analytical thinking paradox. *Applied Cognitive Psychology*, *35*(6), 1523-1534. https://doi.org/10.1002/acp.3885

Goecke, B., Schroeders, U., Zettler, I., Schipolowski, S., Golle, J., & Wilhelm, O. (2023). The nomological net of knowledge, self-reported knowledge, and overclaiming in children. *Journal of Personality Assessment*, *105*(5), 702-713. https://doi.org/10.1080/00223891.2022.2144332

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York, NY:Wiley.

Gregersen, M., Rohd, S. B., Jepsen, J. R. M., Brandt, J. M., Søndergaard, A., Hjorthøj, C., ... & Hemager, N. (2022). Jumping to conclusions and its associations with psychotic

experiences in preadolescent children at familial high risk of schizophrenia or bipolar disorder-the Danish high risk and resilience study, VIA 11. *Schizophrenia Bulletin*, *48*(6), 1363-1372. https://doi.org/10.1093/schbul/sbac060

Harrell, Jr. F. (2024). *Hmisc: Harrell Miscellaneous*. R package version 5.1-3, https://CRAN.R-project.org/package=Hmisc.

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology Section A*, *40*(4), 801-812. https://doi.org/10.1080/14640748808402300

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454. https://doi.org/10.1016/0010-0285(72)90016-3

Kuhn, S. A. K., Lieb, R., Freeman, D., Andreou, C., & Zander-Schellenberg, T. (2022). Coronavirus conspiracy beliefs in the German-speaking general population: endorsement rates and links to reasoning biases and paranoia. *Psychological Medicine*, *52*(16), 4162-4176. https://doi.org/10.1017/S0033291721001124

Lai, G., Tanner, J., & Stevens, D. (2011). The importance of mathematics competency in statistical literacy. *Advances in Business Research*, *2*(1), 115-124.

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology*, *103*(2), 152-166. https://doi.org/10.1016/j.jecp.2008.10.002

Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 21–57). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Makowski, D. (2018). The psycho package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software*, *3*(22), 470. Available from https://github.com/neuropsychology/psycho.R

McLean, B. F., Balzan, R. P., & Mattiske, J. K. (2020). Jumping to conclusions in the less-delusion-prone? Further evidence from a more reliable beads task. *Consciousness and Cognition*, *83*, 102956. https://doi.org/10.1016/j.concog.2020.102956

McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2016). Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: A detailed meta-analysis. *Schizophrenia Bulletin*, *43*(2), 344-354. https://doi.org/10.1093/schbul/sbw056

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research*, *90*(3), 376-419. https://doi.org/10.3102/0034654320919354

Moritz, S., Scheu, F., Andreou, C., Pfueller, U., Weisbrod, M., & Roesch-Ely, D. (2016). Reasoning in psychosis: risky but not necessarily hasty. *Cognitive Neuropsychiatry*, *21*(2), 91-106. https://doi.org/10.1080/13546805.2015.1136611

Moritz, S., Scheunemann, J., Lüdtke, T., Westermann, S., Pfuhl, G., Balzan, R. P., & Andreou, C. (2020). Prolonged rather than hasty decision-making in schizophrenia using the box task. Must we rethink the jumping to conclusions account of paranoia? *Schizophrenia Research*, *222*, 202-208. https://doi.org/10.1016/j.schres.2020.05.056

Moritz, S., Woodward, T. S., & Hausmann, D. (2006). Incautious reasoning as a pathogenetic factor for the development of psychotic symptoms in schizophrenia. *Schizophrenia Bulletin*, *32*(2), 327-331. https://doi.org/10.1093/schbul/sbj034

Moritz, S., Woodward, T. S., & Lambert, M. (2007). Under what circumstances do patients with schizophrenia jump to conclusions? A liberal acceptance account. *British Journal of Clinical Psychology*, *46*(2), 127-137. https://doi.org/10.1348/014466506X129862

Moritz, S., Van Quaquebeke, N., & Lincoln, T. M. (2012). Jumping to conclusions is associated with paranoia but not general suspiciousness: A comparison of two versions of the probabilistic reasoning paradigm. *Schizophrenia Research and Treatment*, 2012. https://doi.org/10.1155/2012/384039

Moritz, S., Veckenstedt, R., Bohn, F., Hottenrott, B., Scheu, F., Randjbar, S., ... & Roesch-Ely, D. (2013). Complementary group Metacognitive Training (MCT) reduces delusional ideation in schizophrenia. *Schizophrenia Research*, *151*(1-3), 61-69. https://doi.org/10.1016/j.schres.2013.10.007

OpenAI. (2024). *ChatGPT* (Dec 20 version) [Large language model]. https://chat.openai.com/chat

Paulhus, D. L., Petrusic, W. M. (2010). *Measuring individual differences with signal detection analysis: A guide to indices based on knowledge ratings.* Unpublished manuscript

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rassin, E. (2008). Individual differences in the susceptibility to confirmation bias. *Netherlands Journal of Psychology*, *64*(2), 87-93. https://doi.org/10.1007/BF03076410

Ross, R. M., McKay, R., Coltheart, M., & Langdon, R. (2015). Jumping to conclusions about the beads task? A meta-analysis of delusional ideation and data-gathering. *Schizophrenia Bulletin*, *41*(5), 1183-1191. https://doi.org/10.1093/schbul/sbu187

Ross, R. M., Pennycook, G., McKay, R., Gervais, W. M., Langdon, R., & Coltheart, M. (2016). Analytic cognitive style, not delusional ideation, predicts data gathering in a large beads task study. *Cognitive Neuropsychiatry*, *21*(4), 300-314. https://doi.org/10.1080/13546805.2016.1192025

Sanchez, C., & Dunning, D. (2021). Jumping to conclusions: Implications for reasoning errors, false belief, knowledge corruption, and impeded learning. *Journal of Personality and Social Psychology*, *120*(3), 789. https://doi.org/10.1037/pspp0000375

So, S. H. W., & Kwok, N. T. K. (2015). Jumping to conclusions style along the continuum of delusions: Delusion-prone individuals are not hastier in decision making than healthy individuals. *PloS one*, 10(3). https://doi.org/10.1371/journal.pone.0121347

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instrum. Comput. 31*, 137–149. https://doi.org/10.3758/BF03207704

Stark, L., Krummenauer, J., Jaggy, A.-K., Kremer, F., Kuntze, S., Nagengast, B., Trautwein, U., Golle, J. (2025). *Evaluating the efficacy of a statistical literacy intervention.* Manuscript in preparation.

Trautwein, U., Golle, J., Jaggy, A. K., 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. https://doi.org/10.1111/nyas.15074

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105. https://doi.org/10.1037/h0031322

Urhahne, D. & Hopf, M. (2004). Epistemologische Überzeugungen in den Naturwissenschaften und ihre Zusammenhänge mit Motivation, Selbstkonzept und Lernstrategien [Epistemological beliefs in science and their relationships with motivation, self-concept, and learning strategies]. *Zeitschrift für die Didaktik der Naturwissenschaften*, *10*, 70-86. ftp://ftp.rz.uni-kiel.de/pub/ipn/zfdn/2004/4.Urhahne_Hopf_071-088.pdf

van der Leer, L., Hartig, B., Goldmanis, M., & McKay, R. (2015). Delusion proneness and 'jumping to conclusions': Relative and absolute effects. *Psychological Medicine*, *45*(6), 1253-1262. https://doi.org/10.1017/S0033291714002359

Veckenstedt, R., Randjbar, S., Vitzthum, F., Hottenrott, B., Woodward, T. S., & Moritz, S. (2011). Incorrigibility, jumping to conclusions, and decision threshold in schizophrenia. *Cognitive Neuropsychiatry*, *16*(2), 174-192. https://doi.org/10.1080/13546805.2010.536084

Ward, T., Peters, E., Jackson, M., Day, F., & Garety, P. A. (2018). Data-gathering, belief flexibility, and reasoning across the psychosis continuum. *Schizophrenia Bulletin*, 44(1), 126-136. https://doi.org/10.1093/schbul/sbx029

Appendix

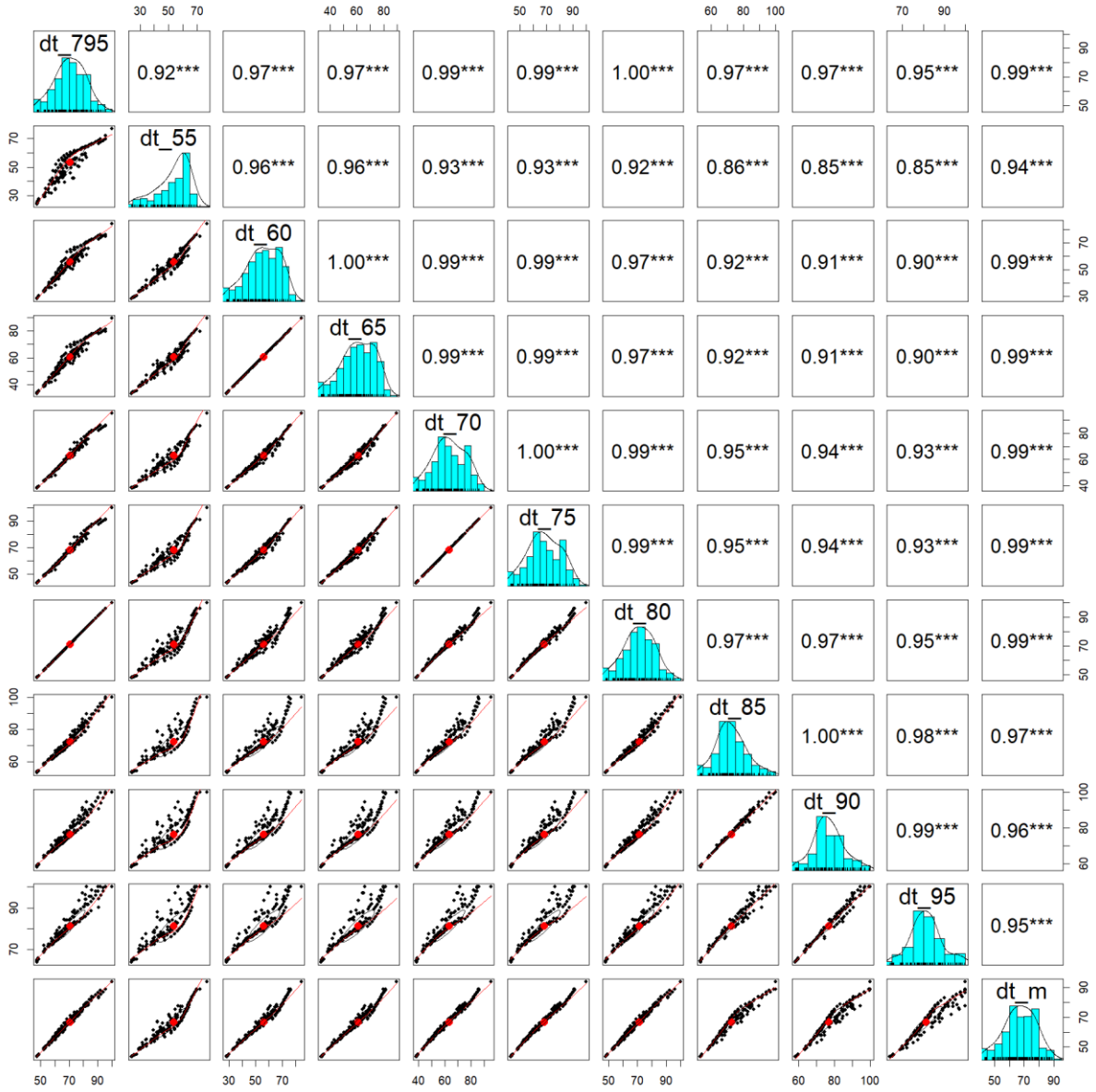
Table A1

Descriptive Statistics of All Variables

Variable	*N* Items	*N*	*M*	*SD*	*Min*	*Max*	α
Decision Threshold	24	299	67.04	10.53	43.43	99.15	.83
Draws to Decision	2	299	5.32	3.30	1.00	20.00	.72
Confirmation Bias	6	299	3.80	0.56	2.00	5.00	.50
Alternation Bias	3	257	11.11	2.43	0.67	18.33	.52
Epistemic Certainty Beliefs	9	255	2.39	0.56	1.00	4.12	.52
Math Grade	1	182	5.85	0.36	5.00	6.00	
Self-concept in Data-related Tasks	6	294	4.05	0.82	1.00	5.00	.90
General Self-efficacy	3	297	4.10	0.71	1.00	5.00	.65
Gender (1 = boys)	1	298	0.55	0.50	0.00	1.00	
Age	1	297	8.69	1.04	6.30	12.13	
Grade Level	1	292	2.79	0.88	1.00	4.00	

Figure A2

Correlation Plot of Decision Thresholds for All Ten Cut-Off Values and Overall Mean



Note. dt_795 describes the decision threshold with a cut-off value of 79.5%. dt_55 to dt_95 describe the decision thresholds with cut-off values from 55% to 95%. dt_m describes the mean decision threshold of all cut-off values from 55% to 95%. Distributions are depicted on the diagonal.

*** = $p < .001$.

4

Study 3: Promoting Primary School Children's Statistical Literacy: Results of a Randomized Controlled Field Trial

Stark, L., Krummenauer, J., Jaggy, A.-K., Kremer, F., Kuntze, S., Golle, J., Nagengast, B., Trautwein, U. (2025). *Promoting primary school children's statistical literacy: Results of a randomized controlled field trial*. Manuscript in preparation.

The Hector Foundation II supported this work. Lucas Stark and Fabienne Kremer are doctoral students at the LEAD Graduate School & Research Network [GSC 1028], funded by the Baden-Württemberg Ministry of Science, Research and the Arts within the framework of sustainability funding for the projects of the Excellence Initiative II.

Abstract

Interpreting statistical data is an important basis for data-based argumentation and informed decision-making in today's societies. At the same time, interpreting data is challenging, and statistical misconceptions and biased interpretations are frequently described in the literature. These issues highlight the need for age-appropriate learning opportunities to foster the development of statistical literacy. In this preregistered study, we evaluated the effectiveness of a statistical literacy intervention for elementary school students when implemented in practice in an extracurricular STEM enrichment program. In a previous efficacy study, the intervention yielded positive effects on third- and fourth-graders' statistical literacy and motivation under highly controlled conditions (i.e., when taught by university staff). The current effectiveness study was conducted to test whether the intervention would also demonstrate effectiveness when conducted by trained course instructors from the field. Participants were 87 third- and fourth-grade children from an extracurricular enrichment program. We tested the effectiveness of the intervention using a multisite randomized controlled field trial with pre- and posttests. Participants were randomly assigned to the intervention or the waitlist control group. Multilevel multiple regression analyses showed significant intervention effects on children's data-based argumentation, decision threshold, and alternation bias as well as on their self-concept and attainment value in data-related tasks. No effects were found on views on variability, draws to decision, intrinsic value, or general self-efficacy. Overall, the results indicated the effectiveness of the intervention in a field setting and the malleability of aspects of statistical literacy in primary-school-aged children.

*Keywo*rds: statistical literacy, data-based argumentation, randomized controlled field trial, primary school children

Promoting Primary School Children's Statistical Literacy: Results of a Randomized Controlled Field Trial

In today's societies, citizens find themselves submerged in an ever-growing flood of information, as an abundance of data continue to pour in. On almost any topic, citizens can find a plethora of diverse interpretations of data, suggesting different and possibly contradictory actions regarding, for instance, their relationships, careers, and nutrition. For example, during the COVID-19 pandemic, conflicting interpretations about the efficacy and safety of various vaccines made it difficult for some individuals to decide whether to get vaccinated. With misinformation on the rise, statistical literacy has increasingly been emphasized as necessary for citizens to be able to effectively navigate their everyday lives and for democracies to function (e.g., United Nations Economic Commission for Europe [UNECE] 2012; Wallman, 1993). Citizens need to be able to understand and interpret data on their own to be able to participate and engage in data-based argumentation and make well-informed decisions and should thereby be trained to do so from early on (Watson & Callingham, 2020; Weiland, 2017).

So far, research on statistical literacy has focused primarily on statistics education in tertiary education, especially introductory statistics courses (e.g., Gopal et al. 2018; Schutz et al., 1998). However, there is evidence that children already begin to develop ideas about statistical concepts in primary school (e.g., Piaget & Inhelder, 1976; Watson & Kelly, 2005; Watson & Moritz, 2000), and early statistical literacy interventions have been called for (e.g., Engel, 2017; Gal, 2002). However, only a few educational interventions (e.g., Ben-Zvi & Sharett-Amir, 2005) have focused on promoting statistical literacy in the primary school years in general, and even fewer interventions (e.g., English & Watson, 2013; Stark et al., 2025b) have been implemented and tested in regular curricular or extracurricular educational settings. However, studies in educational field settings (or similar settings) are important for gauging the potential contributions of interventions conducted under conditions that are as close as possible to their intended applications (e.g., Barnett, 2011; Lazowski & Hulleman, 2016).

To expand research in this area, we developed an intervention for third- and fourth-graders from an extracurricular enrichment program for talented children to promote their statistical literacy. We strategically selected talented students from an extracurricular STEM enrichment program called the Hector Children's Academy Program (HCAP, for more information, see Trautwein et al., 2023) to address their special needs (Özdemir & Işıksal Bostan, 2021) and to explore upper performance limits. The intervention presented in the present study is based on the predict-observe-explain approach (Gunstone & White, 1981) and

cooperative learning methods (e.g., Capar & Tarim, 2015). A previous efficacy study (Stark et al., 2025b), in which university staff served as instructors in the intervention, already showed promising results that spoke in favor of the efficacy of the intervention under standardized conditions. In the present study, the intervention's effectiveness was tested under field conditions, including the condition that course instructors from the field were trained by us to conduct the intervention themselves.

Statistical Literacy

Despite the widespread use of the term *statistical literacy*, several approaches for describing statistical literacy exist (e.g., Ben-Zvi & Garfield, 2004; Gal, 2002; Schield, 1999; Wallman, 1993; Watson & Callingham, 2003), with no consensus that any one of them should predominate (Sharma, 2017). For example, in his description of statistical literacy, Gal (2002) concentrated on *mathematical, statistical,* or *context knowledge elements* and added also what he called *dispositional elements*, such as *scepticism* and *motivational beliefs* as an important base of statistical literacy. By contrast, Watson and Callingham (2003) defined hierarchical ability tiers crucial to statistical literacy, such as recognizing the *need for data* or evaluating claims on statistical data. Moreover, it has been argued (e.g., Jones et al. 2000; Watson & Callingham, 2003) that the mastery of diverse tasks is associated with statistical literacy, such as *graph reading* or *reducing data*. Thus, there is a somewhat broad spectrum of approaches that have been used to define, study, and promote statistical literacy, which may cause some disagreement about what should be measured in the respective studies.

However, there are also commonalities in the field. First, a statistically literate person is able to *interpret* (e.g., Gal, 2002; Jones et al., 2000; Schield, 1999; Wild & Pfannkuch, 1999) and *critically evaluate* statistical information (e.g., Gal, 2002; Wallman, 1993; Wild & Pfannkuch, 1999). These abilities are central to statistical literacy because it enables people to participate in social discourse and to become active citizens (Weiland, 2017), and it can be summarized under the term data-based argumentation (Krummenauer & Kuntze, 2018). Consequently, in the present study, we understand statistical literacy to be the ability to understand and interpret statistical information in order to be able to participate in data-based argumentation. Second, many researchers (e.g., McKenzie, 2004; Watson & Callingham, 2003) have identified the concept of variability as central to statistical literacy. Variability is omnipresent (Moore, 1990), and it is the reason why statistical methods are being used (Wild & Pfannkuch, 1999). There are many misconceptions about variability (e.g., Chan & Ismail, 2013), and these misconceptions hinder the development of appropriate data-based arguments.

Therefore, it is essential for a statistically literate person to be able to deal with statistical variability, and variability is also a central pillar in our study. Third, a statistically literate person must be motivated to think statistically (e.g., Gal, 2019; Wild & Pfannkuch, 1999) or else they will not be able to make use of their statistical literacy. Also, higher motivation in statistics has been found to be related to higher performance (Hood et al., 2012). Therefore, it is crucial to positively impact learners' motivational beliefs about data-related tasks when promoting statistical literacy, and for this reason, we also included motivational outcomes in our study.

Thus, in this study, we focused on aspects of statistical literacy that are crucial for young learners to become active citizens. More specifically, we focused on data-based argumentation, dealing with variability, and motivation in data-related tasks.

Data-Based Argumentation

Statistical literacy includes the ability to interpret data and communicate the interpretation in a clear and concise manner, but it also entails the ability to critically evaluate interpretations made by others (e.g., Gal, 2002; Wild & Pfannkuch, 1999). Such requirements can be summarized under the term data-based argumentation (Krummenauer & Kuntze, 2018) and are necessary for citizens to be able to participate in social discourse. Consequently, statistical literacy interventions should focus in particular on enhancing participants' data-based argumentation. This emphasis will help to ensure that citizens are able to participate effectively in social discourse and make informed decisions about the issues that affect their lives. But at what age can data-based argumentation already be observed?

There is evidence that even young children are able to develop data-based arguments. A study by Krummenauer and Kuntze (2018) found that about one third of $N = 385$ fourth-graders were able to use data in their arguments. There is also initial evidence that children's data-based argumentation skills can be enhanced from an early age on. A study by Ben-Zvi (2006) found that fifth-graders were able to enhance their argumentation skills by making inferences based on samples that became successively larger (Bakker & Gravemeijer, 2004). Similarly, Paparistodemou and Meletiou-Mayrotheris (2008) found that using real data sets and letting third-graders explore them with a data visualization tool could enhance the children's data-based argumentation. In our previous efficacy study (Stark et al., 2025b), we found that children's data-based argumentation could be enhanced by an extracurricular intervention that was based on the predict-observe-explain approach (Gunstone & White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015). Taken together, there is initial

evidence that data-based argumentation is malleable even at a fairly young age, at least to a certain degree.

Dealing With Variability

Many researchers and statisticians see phenomena of *variability* as located at the heart of statistics (e.g., McKenzie, 2004; Watson & Callingham, 2003); thus, to be statistically literate and produce data-based arguments, it is also necessary to know about the statistical concept of variability. Variability describes the phenomenon that the value of an observed variable may differ between observations or investigation units (e.g., Reading & Shaughnessy, 2004). It is crucial for students to learn to deal with variability in order to become statistically literate, especially young learners who do not yet know much about statistics. Variability in statistics is omnipresent (Cobb & Moore, 1997) and needs to be anticipated when developing data-based arguments. When generating data-based arguments, for instance, variability in data can require learners to anticipate that data may have to be viewed as ambiguous evidence and that, therefore, they must check for whether alternative interpretations or arguments are possible. But even though variability can induce ambiguity, the main tendencies can still be identified and justified by data in many cases.

Weaknesses in dealing with statistical variability have often been found to coincide with so-called *cognitive biases* (e.g., Ben-Zvi & Garfield, 2008; Tversky & Kahnemann, 1971). For example, the *jumping to conclusions bias* (Garety et al., 2005) describes a form of disregarding or underestimating the existence of variability. People with this bias draw conclusions on the basis of too little evidence. The *alternation bias* (Kahneman & Tversky, 1972; Tversky & Kahneman, 1971) describes the idea that people tend to assume that random events alternate more often than they would on average. People with this bias may expect less variability than statistical methods would suggest. The *confirmation bias* (Kahneman & Tversky, 1972; Tversky & Kahneman, 1971) describes people's tendency to neglect evidence that goes against their current worldview and to accept evidence that confirms it. It can be viewed as the tendency to cherry-pick data to reduce variability in favor of their current worldview. Taken together, early interventions to reduce such cognitive biases have been called for (e.g., Gregersen et al., 2022).

Children usually tend to neglect variability, as their perspective on knowledge is rather absolutist (e.g., Walker et al., 2020): that is, something is either right or wrong. Variability is just seen as differences between data (Lehrer & Kim, 2009) rather than as noise around a signal (Konold & Pollatsek, 2002). According to Piaget (1976), children before the age of 11 begin

to develop an intuition about chance in variability, but they are not able to reason abstractly about hypothetical situations. If this idea is correct, it would clearly limit the prospects of any interventions for children below the age of 10. However, there is some empirical evidence that primary-school-aged children's abilities to deal with statistical variability may already be malleable. For example, Ben-Zvi and Sharett-Amir (2005) found that when three second-graders were actively involved in exploratory data analysis with real data and exchanged arguments in classroom discourse, they began to reason about distributions. English and Watson (2013) let children from five fourth-grade classes explore variability by measuring arm spans and experiencing measurement variability and found that they learned to acknowledge the uncertainty that comes with reasoning about variable data. This finding indicates that primary school children already have the potential to deal with variability in a statistical context and that actively involving children in the whole process of data gathering, analysis, and interpretation could be effective. However, stronger empirical evidence is needed to show how it can be developed systematically.

Motivational Beliefs

Developing the ability to construct arguments on the basis of data is a valuable skill for active citizenship. However, this ability requires not only statistical knowledge but also the motivation and confidence to use it. Motivational beliefs are seen as crucial for statistical literacy (e.g., Gal, 2002; Watson & Callingham, 2003). They have been positively linked to statistics engagement (e.g., Gopal et al., 2018; Schutz et al., 1998) and achievement (e.g., Hood, Creed, & Neumann, 2012) in tertiary education. According to Eccles' expectancy-value theory (e.g., Eccles & Wigfield, 2020), a person is more likely to execute a certain behavior when they feel competent enough to execute it and they see enough value in it or in its potential outcomes. Therefore, it is essential to foster motivational beliefs in young learners, such as *self-concept*, *attainment value*, and *intrinsic value*. Building these beliefs from an early age can empower them to utilize their statistical literacy skills effectively.

To our knowledge, there is no research on promoting motivational beliefs in statistics education in primary school children. Intervention effects on statistics motivation have been studied only in tertiary education. From this research, it is known that active involvement in data-related problems can enhance statistics motivation (Gopal et al., 2018; Khoshnoodifar, Ashouri, & Taheri, 2023) and that cooperative learning methods and feedback may enhance students' self-concept in statistics (e.g., Krause, Stark, & Mandl, 2009).

Stepwise Evaluation of Educational Interventions

It has been argued (see Gottfredson et al., 2015; Greene, 2015; Herbein et al., 2018) that educational interventions should be tested in a stepwise fashion, moving from standardized conditions toward their real-world context. According to this argumentation, so-called efficacy studies assess the success of an intervention under ideal, controlled conditions, whereas effectiveness studies evaluate its impact in realistic field settings (Gottfredson et al., 2015; Greene, 2015; Herbein et al., 2018). After a promising efficacy study (Stark et al., 2025), our current aim was to take the next step in the process of evaluating our intervention, which took place as an extracurricular enrichment program for talented primary school children. The context of an enrichment program opens up the possibility of testing the effectiveness of the intervention under field conditions that also offer rich ground for learning effects. The statistical literacy intervention used in this study already showed significant positive effects on cognitive and motivational constructs in our previous efficacy study (Stark et al., 2025b). At that time, university staff were the course instructors, but in the current study, we aimed to determine whether the intervention would also be effective under field conditions. For this reason, a diverse sample of course instructors were trained to offer the statistical literacy intervention. Whereas during the efficacy study, the university staff was taught in weekly sessions in close contact with the course developer, in the less standardized conditions of the effectiveness study, course instructors were all trained with asynchronous online material with a joint synchronous kick-off and closing session before the pretest.

To label an intervention as effective, there needs to be evidence that the intervention provides the intended effects in a target group (Fixsen et al., 2013). Additionally, there needs to be evidence that the intervention was implemented as intended by the course instructors. To gather such evidence, intervention fidelity should be assessed (Hulleman & Cordray, 2009; Humphrey et al., 2016). To ensure quality implementation fidelity in the present study, course instructors received a course manual and participated in an online workshop. An assessment of intervention fidelity can be applied to check for whether the theoretical assumptions about core components and the change model can be confirmed (Carroll et al., 2007). It can be useful to assess multiple perspectives on the intervention's fidelity, for example, from the instructors' and the children's perspectives (Schultes et al., 2015).

The Present Study

There is initial evidence that statistical literacy can be promoted from an early age on, but most existing studies have been of an exploratory or quasi-experimental nature. The goal

of the present study was the practical implementation of a statistical literacy intervention that was recently developed for primary school children and shown to be effective in a first efficacy study with 53 children under standardized conditions (Stark et al., 2025b).

We formulated and preregistered one confirmatory and several exploratory research questions. First, we asked whether the intervention would have a significant effect on cognitive aspects of primary school children's statistical literacy compared with a waitlist control group (confirmatory research question). We predicted positive effects of the intervention on performance measures and negative effects on cognitive biases. Second, we wanted to check for whether the intervention would have a significant effect on motivational aspects of primary school children's statistical literacy compared with the waitlist control group (exploratory research question). We assumed an exploratory approach, as positive intervention effects might be counteracted by reference group effects. When talented children come from their mixed classrooms into groups with exclusively talented children, their motivation tends to decrease as their reference group becomes more capable (Preckel et al., 2010; Zeidner & Schleyer, 1999).

Moreover, we preregistered and analyzed additional exploratory research questions. With these questions, we wanted to explore the possibility of differential intervention effects and associations between course fidelity measures and intervention outcomes.

Method

Participants and Procedure

A total of 101 third- and fourth-graders agreed to participate in the study. The final sample consisted of 87 of them (39.1% girls, M_{age} = 9.10 years, SD_{age} = 0.48 years, see Table 1), as 14 did not attend the pretest and posttest sessions. Figure 1 presents the attrition data. The intervention was implemented as part of the HCAP, an extracurricular enrichment program in the German federal state of Baden-Württemberg, in the second term of the 2022/2023 school year. The HCAP is hosted at 69 different local sites and is tailored to talented, interested, and motivated primary school children (for more information about the HCAP, see Trautwein et al., 2023). Children are nominated for the program by their teachers. The nominated children can choose from a variety of courses with a focus on Science, Technology, Engineering, and Mathematics (STEM) subjects.

On the basis of results from our previous study (Stark et al., 2025b), we determined the minimum required sample size with a power analysis. As preregistered, we aimed to recruit at least between 8 and 22 local sites with approximately 13 children at each site. With 87 children

Table 1

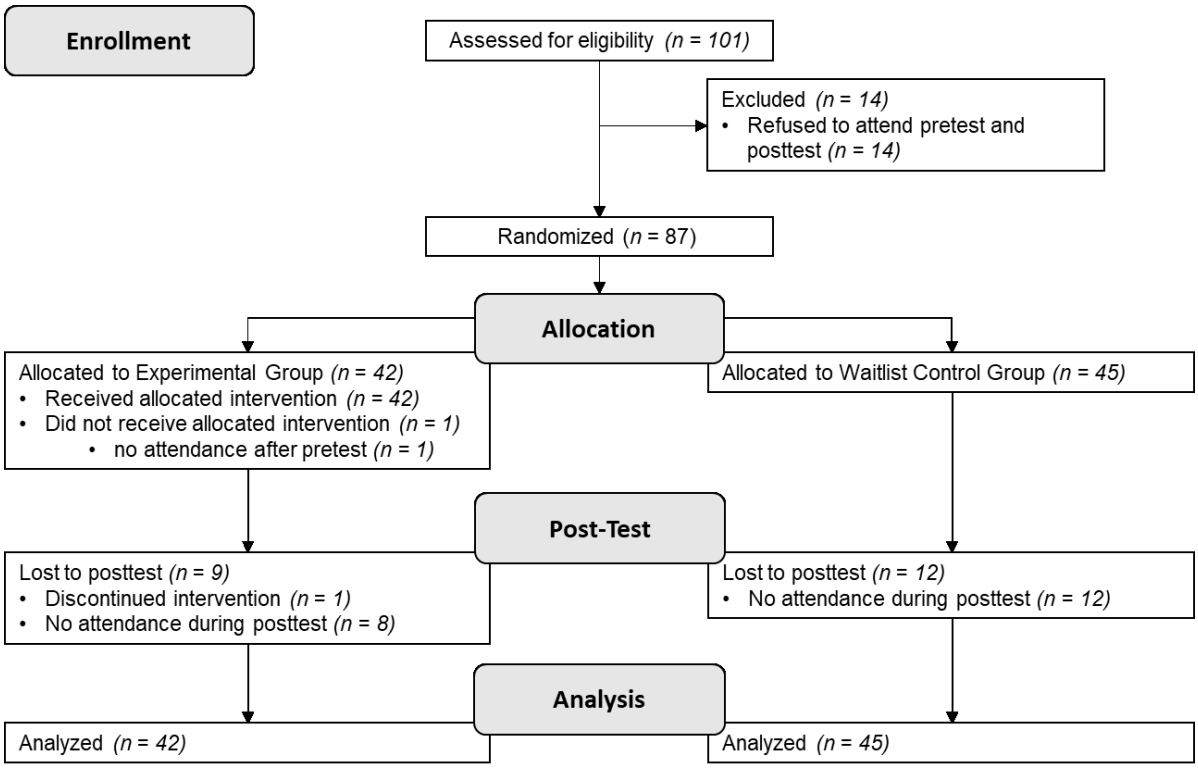
Description of the Sample

Group	*N*	Boys	Girls	Age (in years)	Grade 3	Grade 4
Intervention group	42	57.1%	42.9%	*M* = 9.07 (*SD* = 0.43)	23	14
Waitlist control group	45	62.2%	35.6%	*M* = 9.12 (*SD* = 0.52)	24	18

Note. *N* = number of participating children; *M* = mean; *SD* = standard deviation.

Figure 1

Attrition Flow Chart



from nine local sites of the HCAP, we ended up at the lower end of the minimum required sample size.

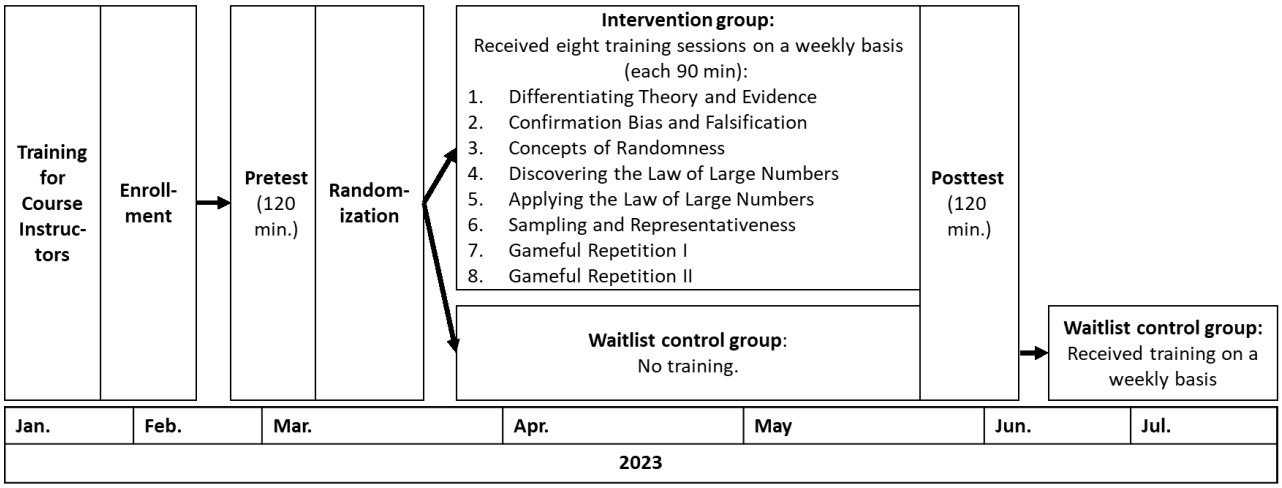
The intervention was offered as a course called “Luck or genius? Understanding data and making predictions” in the course programs of nine local sites of the HCAP. An announcement provided information about the intervention’s contents and goals as well as some initial information about the study. The intervention was open to all third- and fourth-graders who were nominated for the program and had written parental consent to participate in the study. For each site, up to 16 children could be admitted to the course.

Study Design

To evaluate the efficacy of the intervention, we conducted a randomized controlled field trial with a pretest, a posttest (Friedman et al., 2010), and a waitlist control group. The pretest measurement took place 1 day to 1 week before the intervention began (see Figure 2).

Figure 2

Study Procedure and Contents of the Intervention



After the pretest, the children were randomly assigned to one of the two conditions. At three of the sites, there were not enough children to form groups in both conditions within the site, so they were all randomly assigned to either the intervention or the control condition (resulting in one experimental group and two waitlist control groups across the three sites). Children in the intervention condition (n = 42) received weekly training for a total of eight sessions, each 90 min long. Children in the waitlist control group (n = 45) received the training after the posttest. During that time, it was possible for the children from both the intervention condition and the waitlist control condition to also participate in other courses offered by the HCAP. After the eight sessions were taught in the intervention group, all children participated in the posttest. Additionally, the parents were asked to fill out a questionnaire on family background and to provide further information about the children.

Description of the Intervention

We based the intervention on two core components, which can be defined as intervention properties that are hypothesized to make the intervention effective (Abry et al., 2015). These core components can help in the designing of the intervention but also in assessing its fidelity. We have based the conceptualization of this statistical literacy intervention on the *predict-observe-explain approach* (Gunstone & White, 1981) and *cooperative learning methods* (e.g., Capar & Tarim, 2015).

The predict-observe-explain approach involves three steps, in which learners make predictions, observe a phenomenon, and then explain the correctness of their prediction with the observation in mind. In a meta-analysis, Gustina et al. (2023) found a moderate effect (d = 0.60, 95% CI [0.49, 0.70]) on students' learning and critical thinking in the domains of mathematics and natural sciences. By stating a prediction first, learners have the opportunity to observe contradictory evidence, which then leads to cognitive dissonance and makes the learner more motivated to close the gap in their knowledge (e.g., Yahya & Sukmayadi, 2020). This motivation can be used in the explanation phase to help learners construct their knowledge.

Cooperative learning methods (e.g., Capar & Tarim, 2015) encompass methods that involve group-based learning strategies where children work together to achieve common learning goals. These methods incorporate activities that promote exchanges between children so they can share their understanding of statistical concepts. From research on tertiary statistics education, there is evidence that cooperative learning can be used in statistics courses to help students overcome misconceptions and enhance their learning and motivation (e.g., Giraud, 1997; Garfield, 1993).

As an example, in Session 3 (see Figure 2), the goal is to explore concepts of randomness. In the prediction phase of the session, the children are asked to write down a sequence of 20 coin toss results that look random to them. Then, in the observation phase, the children toss 20 coins and write down the actual results. In the explanation phase, the children work in pairs to discuss what they think the difference is between the imagined and thrown sequences. In a game, the pairs then try to guess what the imagined and thrown sequences of the other pairs of students look like. Finally, in a teacher-led discussion, the children explore the concept of randomness and the alternation bias by aggregating the observed data across all the children.

As another example, in Session 5 (see Figure 2), the goal is to deepen knowledge about the law of large numbers. In the prediction phase of the session, the children are asked to predict whether a paper airplane with a sharp tip or one with a flat tip will fly further in general. Then, in the observation phase, each child throws the paper airplane multiple times and records their flying distances. In the explanation phase, the children start explaining which paper airplane flies further and how their views changed the more they threw the planes. Then, in a teacher-led discussion, the children explain how more observations were connected to a more secure decision about which paper airplane flies further in general, and they integrate the law of large numbers.

All intervention materials in German and the materials from an example session in English are available online on the Open Science Framework (OSF; *https://osf.io/j2vyw*).

Measures

Several instruments were selected to cover the central constructs that were supposed to be affected by the intervention. We assessed the children's data-based argumentation, five indicators of their concept of variability, motivational beliefs, fluid intelligence, and intervention fidelity. All instruments and the corresponding descriptive statistics are presented in Table 2.

Data-Based Argumentation

As a general indicator of statistical literacy, we assessed data-based argumentation with an adapted version of the measurement instrument from the previous study (Stark et al., 2025b). In 12 items with increasing difficulty, the children had to evaluate whether a statement was true and justify their decision on the basis of the available data. For the easiest items, the children had to name only one data point to justify their evaluation. For more difficult items, test takers had to identify multiple data points, variability, and differences in samples, if any, to solve the task correctly. The answers were rated by three raters (0 = *false*, 1 = *correct*). The rating with the highest agreement level among the raters was selected as the final rating for each item. Items 11 and 12 were excluded because the solution frequencies were under 6% across both measurement occasions. To check for interrater reliability, we calculated Fleiss' Kappa (Fleiss & Cohen, 1973), which confirmed good reliability (K_{T1} = .821, K_{T2} = .697). The internal consistency of the scale was assessed by applying the Kuder-Richardson coefficient for binary variables with varying difficulties (KR-20; Kuder & Richardson, 1937; KR-20_{T1} = .50, KR-20_{T2} = .60). A sum score for the remaining 10 items was calculated for each child.

Dealing With Variability

To take a comprehensive look at the broad construct of dealing with variability, we assessed five different variables that each showed a different facet of variability.

We assessed the children's *views on variability* with five items by asking them to generate distributions on the basis of prompts, such as "On one street, a squirrel lives in every third tree on average." The children were asked to tick boxes below tree icons to create a distribution that looked typical to them (see Figure 3). The children's expectations of variability in random distributions were then sorted into three levels: (a) no variability (i.e., one squirrel in every third tree), (b) little variability (i.e., one squirrel in every group of three trees), and (c)

Table 2

Descriptive Statistics for Dependent Variables and Fluid Intelligence: Means, Standard Deviations, Reliabilities, and Number of Items

			T1				T2			
Construct	*N* items	Group	*N*	*M*	*SD*	α	*N*	*M*	*SD*	α
Data-based Argumentation*	14	All	82	2.63	1.59	.50	67	4.09	1.94	.60
		IG	40	2.70	1.73		34	4.53	2.25	
		CG	42	2.57	1.47		33	3.64	1.48	
Views on Variability	5	All	75	1.87	0.57	.69	64	1.95	0.57	.66
		IG	38	1.85	0.51		33	1.99	0.50	
		CG	37	1.90	0.62		31	1.92	0.65	
Draws to Decision	2	All	81	8.78	5.53	.77	67	9.37	4.60	.70
		IG	40	9.07	5.72		34	10.44	4.91	
		CG	41	8.50	5.40		33	8.27	4.03	
Decision Threshold*	24	All	81	77.41%	13.76%	.90	67	83.26%	11.52%	.91
		IG	40	78.44%	14.10%		34	88.42%	9.63%	
		CG	41	76.20%	13.42%		33	77.92%	11.01%	
Alternation Bias	3	All	73	10.61	2.69	.58	63	10.01	2.71	.63
		IG	34	10.64	2.64		33	9.33	2.99	
		CG	39	10.59	2.77		30	10.74	2.18	
Confirmation Bias	8	All	77	3.71	0.60	.75	64	3.67	0.61	.69
		IG	38	3.80	0.59		33	3.66	0.60	
		CG	39	3.62	0.61		31	3.67	0.64	
Self-concept	6	All	81	3.79	0.90	.93	66	4.02	0.80	.88
		IG	39	3.80	0.89		33	4.32	0.62	
		CG	42	3.79	0.92		33	3.71	0.86	
Intrinsic Value	6	All	80	3.82	0.86	.91	66	3.74	1.07	.95
		IG	39	3.85	0.84		33	3.99	1.08	
		CG	41	3.79	0.88		33	3.49	1.01	
Attainment Value	3	All	77	4.05	0.86	.77	66	4.00	0.84	.76
		IG	37	4.08	0.83		33	4.24	0.80	
		CG	40	4.03	0.89		33	3.76	0.82	
General Self-efficacy	3	All	78	4.14	0.63	.65	65	4.16	0.68	.62
		IG	38	4.20	0.64		33	4.20	0.61	
		CG	40	4.09	0.63		32	4.11	0.76	
Fluid Intelligence	16	All	82	9.00	3.06	.71 *				
		IG	40	9.20	2.69					
		CG	42	8.81	3.39					

Note. *N* = number of participating children; IG = intervention group; CG = control group.

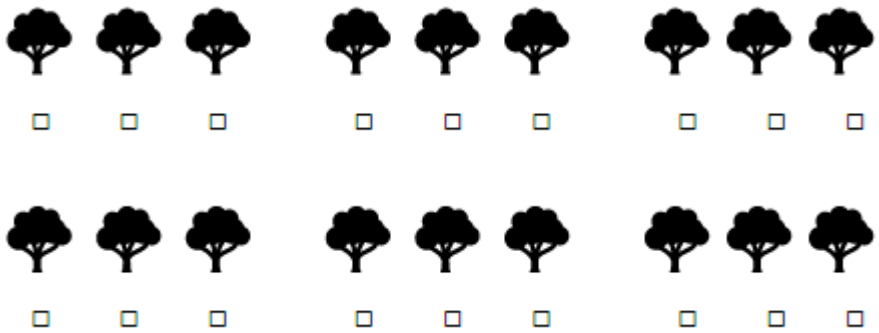
***** For binary variables, the Kuder-Richardson coefficient was calculated instead of Cronbach's alpha.

Measurement points: T1 = February to March 2023, T2 = May to June 2023.

full variability (i.e., approximately one squirrel per three trees on average, but the squirrels showed no regularity). A mean value was created for each student (α_{T1} = .62, α_{T2} = .61).

Figure 3

Sample Item for Views on Variability



The *jumping to conclusions bias* was assessed with the 85:15 and 60:40 versions of the beads task (Dudley, John, Young, & Over, 1997a, 1997b; Garety et al., 2005). In the first of two items, children were shown two different jars, one with 85 yellow beads and 15 black beads and one with 15 yellow beads and 85 black beads. In the second item, the ratios were 60 to 40 and 40 to 60, respectively. They were then told that one of the two jars was chosen at random, and beads would be drawn from this jar in an order that had been determined beforehand. After each draw, the children could decide whether they wanted to see another bead or if they already knew which jar the beads were drawn from. All the drawn beads were visible at any time to aid memory. *Draws to decision* were then used as an outcome measure with fewer draws suggesting a greater jumping to conclusions bias. We created a mean value for each participant. Because there were only two items, we additionally calculated the Spearman-Brown coefficient between the items to judge whether the scale's reliability was sufficient (Eisinga, Te Grotenhuis, & Pelzer, 2013; α_{T1} = .74, α_{T2} = .69; r_{T1} = .61, r_{T2} = .49).

To assess the children's *decision threshold* as a complement to the beads task, we developed a new measurement instrument (Stark et al., 2025a). The decision threshold provides information about how certain a child needs to be to decide they are sure about something. In three scenarios with eight items each, children see jars that are similar to the beads task (60:40, 70:30, 85:15), and then they have to decide whether they know from which jar a set of beads (e.g., three orange and one blue bead) was drawn. By using signal detection theory (e.g., Wickens, 2001), we calculated the children's criterion locations. To apply this procedure, we separated signal trials from signal and noise trials. Signal trials were items for which the certainty was over 80% in knowing from which jar the beads were drawn. Signal and noise trials included items for which the certainty was under 80%. We then transformed each child's

criterion value into a decision threshold in the form of a percentage. For example, a decision threshold of 95% is equivalent to an alpha level of 5%.

The *alternation bias* was assessed with another self-developed measurement instrument. Children were asked to generate three sequences each containing 20 coin toss results with the following prompt: "Imagine you flip a coin 20 times. It can show heads or tails after each flip. What sequences of heads and tails would you expect? Think of three sequences of 20 coin flips each. They should look as random as possible to you." Then the alternations between heads and tails were counted for each sequence (e.g., 14 alternations in HTHTTTHTHTTHTHHTHTHH). A mean value was generated for each child (α_{T1} = .59, α_{T2} = .63). Given that a total of 0 to 19 alternations are possible, mean values further away from 9.5 alternations are considered more biased.

Children's *confirmation bias* was assessed with an adapted version of the self-assessment questionnaire by Rassin (2008). The children were instructed to rate themselves on eight items (e.g., "Sometimes I know things before there is actual proof of them") on a 5-point rating scale ranging from 1 (*not true at all*) to 5 (*exactly right*). Then a mean was computed for each participant (α_{T1} = .75, α_{T2} = .69).

Motivational Beliefs

We assessed three domain-specific and one general motivational variable. Self-concept was measured with six adapted items (α_{T1} = .93, α_{T2} = .88; based on Arens et al., 2011; Gaspard et al., 2015, e.g., "Everything that has to do with data comes easy to me"). Intrinsic value was assessed with six items (α_{T1} = .86, α_{T2} = .95; based on Stalder, 2013, e.g., "I like to do everything that has to do with data"). Attainment value was measured with three adapted items (α_{T1} = .77, α_{T2} = .76; Ramm et al., 2006, e.g., "Everything that has to do with data is important to me"). Additionally, we measured general self-efficacy (α_{T1} = .65, α_{T2} = .62; based on Beierlein et al., 2012, e.g., "In difficult situations, I can rely on my abilities").

All items were rated on a scale ranging from 1 (*not true at all*) to 5 (*exactly right*). Mean values were calculated for each child.

Fluid Intelligence

Fluid intelligence, which was used as a covariate in our analyses, was measured at pretest with 16 figural items from an age-adapted version of the Berlin Test of Fluid and Crystallized Intelligence (BEFKI; see, e.g., Schroeders et al., 2020), appropriate for primary school children (Schroeders et al., 2016). For each item, the children were asked to select one

of three possible figures for every fourth and fifth unit within a sequence that needed to be completed. The time was limited to 15 min. A binary coding scheme was applied to score the answers (0 = *false*, 1 = *correct*), and a sum score was calculated. The internal consistency was high ($KR\text{-}20_{T1} = .71$).

Implementation Fidelity

We assessed implementation fidelity in three different ways to obtain multiple perspectives on the effectiveness of the intervention design on the children's outcomes (Humphrey et al., 2016).

First, we assessed the course instructors' view of the quantity and quality of the implemented exercises (Greene, 2015; Odom et al., 2010). The compliance with the exercises in the course manual is labeled *adherence* ("Was the exercise conducted?"; dummy-coded: 0 = *no*, 1 = *yes*), whereas the perceived quality of those exercises was labeled *quality of delivery* ("How well was the exercise implemented?"; 3-point rating scale: 1 = *not well*, 2 = *okay*, 3 = *well*; see Humphrey et al., 2016; Nelson et al., 2012; Schoenwald et al., 2011). To this end, course instructors filled out a questionnaire after each course unit.

Second, the course instructors rated how well aspects of the core components were implemented in each session. Three items that were rated from 1 (*not at all*) to 5 (*very much*) were used for the predict-observe-explain approach (e.g., "How surprised were the children by the results?"). For cooperative learning methods, we used three items per session (e.g., "In today's session, the children cooperated a lot") rated from 1 (*not true at all)* to 5 (*exactly right*). We first calculated the mean for each item across all sessions. Then we calculated internal consistency across these mean item values for both the predict-observe-explain approach (Cronbach's $\alpha = .62$) and the cooperative learning methods (Cronbach's $\alpha = .84$).

Third, we also assessed the children's views on the implementation fidelity of the core components. At posttest, children from the intervention group were asked whether certain activities helped them learn. The items were answered on a rating scale ranging from 1 (*not at all*) to 5 (*very much*). For cooperative learning methods, we used three items (e.g., "sharing my views with other children"; Cronbach's $\alpha = .78$). For the predict-observe-explain approach, we used four items (e.g., "checking assumptions with data"; Cronbach's $\alpha = .85$).

Analyses

In a first step, we checked for baseline differences between the intervention and control groups as a prerequisite for isolating the effect of the intervention from any pre-existing group

differences. We preregistered that we would check for nonsignificant baseline differences to assess the baseline equivalence of all pretest values of the dependent variables: gender, age, grade level, and fluid intelligence according to the guidelines of the What Works Clearinghouse (2022). To be in accordance with these guidelines, if there were any mean value differences between the two groups on the pretest values of any variable with an effect size $|d| > 0.05$ despite randomization, these variables had to be included as additional predictors in further analyses, even when nonsignificant. Because grade level and age are closely intertwined, we exclusively employed age as a predictor because of its greater granularity. In line with recommendations for measures of baseline equivalence, we used Hedges' *g* (Hedges, 1981) for continuous variables and the Cox index (Cox, 1970) for binary variables. The intervention variable was dummy-coded (1 = *intervention*, 0 = *waitlist control group*).

In a second step, we examined implementation fidelity (see Hulleman & Cordray, 2009; Humphrey et al., 2016; O'Donnell, 2008). To do so, we looked at the individual course instructors' ratings of the adherence and quality of the delivery values of the course sessions. We also looked at their assessments of the extent to which the two core components of the intervention were successfully implemented. To complement the course leaders' view, we also looked at the extent to which the children from the intervention condition rated various aspects of the core components as useful for their learning success. High scores on all of these measurement instruments should indicate that the intervention was implemented as intended and that any intervention effects can be traced back to the conceptualization of our intervention.

Third, we analyzed the effectiveness of the intervention with hierarchical multiple regression analysis with the lavaan (Rosseel, 2012) package. Intention-To-Treat (ITT) effects, which are considered a "strict test" (Fisher et al., 1990) of effects, were estimated for each outcome variable separately to identify intervention effects. By including all participants, regardless of whether they complied with the intervention or dropped out, the ITT analysis can account for potential biases that can arise when analyzing only those who completed and complied with the intervention. The pretest score was included in each model to increase power (e.g., Aiken et al., 2003). Due to the standardization of the dependent and predictor variables, the regression coefficients can be interpreted as Cohen's *d* (Cohen, 1988). To assess differential effects (e.g., for the children's pretest scores), the cross-product of the corresponding variables with the intervention condition was included as an additional predictor variable in additional models.

To account for the nesting of children in local sites, multiple linear regressions were computed with a design-based correction of standard errors (Asparouhov & Muthén, 2006) by

adding the cluster argument in the lavaan (Rosseel, 2012) package to correct the standard errors for the nested structure of the data. In contrast to the preregistration, we did not use the lavaan.survey (Oberski, 2014) package because the standard errors were already adjusted due to the cluster argument in lavaan.

Missing values in the primary variables ranged from 3.37% to 28.09%. To deal with missing data due to the children's absence at pretest or posttest or due to nonresponse for individual scales, we used the full information maximum likelihood approach (Enders, 2001).

All analyses were conducted with a significance level of $p < .05$. For our confirmatory directed research question, one-tailed tests were applied (Hales, 2023). For all exploratory research questions, we applied two-tailed tests. Finally, we used the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to control the false discovery rate due to multiple testing within each research question.

Transparency and Openness

We adhere to the Journal Article Reporting Standards (JARS; Kazak, 2018). All data and analysis codes for the main analyses are available on the OSF (*https://osf.io/f6bxq*). Additional materials are available upon request. Data were analyzed with R (version 3.6.1; R Core Team, 2019). This study's design, research questions, and analysis plan were preregistered prospectively on the Registry of Efficacy and Effectiveness Studies (REES; see https://sreereg.icpsr.umich.edu/sreereg/subEntry/23780/pdf?action=view) before the posttest data were collected. Cases in which we had to deviate from the preregistration are described in the section on statistical analysis. The study protocol was approved by the Ethics Committee of the Faculty of Economics and Social Sciences at the University of Tübingen (A2.5.4-260_bi). ChatGPT (OpenAI, 2024) was used for editing grammar and style, but not for research purposes.

Results

Preliminary Analysis

Descriptive statistics for all study variables can be found in Table 2. Bivariate correlations between study variables are shown in Table 3.

Table 3

Correlations Between the Dependent Variables and the Covariates at Pretest (Below the Diagonal), at Posttest (Above the Diagonal), and Between Points in Time (On the Diagonal)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Data-based Argumentation	.54***	.22	-.06	-.11	-.15	-.34**	-.05	.03	-.20	-.32**	.37**	-.07	.02	.16
2. Views on Variability	-.02	.35**	.21	.18	.00	-.04	.00	.10	-.08	-.24	.04	-.14	-.15	.01
3. Draws to Decision	.13	.04	.38**	.44***	-.17	.02	.14	.21	.07	.10	-.08	-.14	-.29*	-.26*
4. Decision Threshold	.15	.09	.59***	.22	-.30*	.14	.27*	.18	.01	.04	-.05	-.17	-.09	-.23
5. Alternation Bias	.08	-.27*	.03	.16	.19	-.04	-.27*	-.18	-.18	-.01	-.01	.30*	.10	-.10
6. Confirmation Bias	-.10	-.22	-.18	-.27*	.11	.54***	.38**	.43***	.40***	.60***	-.25	-.21	-.13	.01
7. Self-concept	.12	.07	-.22	-.15	.19	.19	.51***	.67***	.48***	.52***	.07	-.15	-.02	.11
8. Intrinsic Value	.18	-.04	-.08	-.04	-.02	.37***	.56***	.38**	.48***	.42***	-.10	-.29*	-.17	.01
9. Attainment Value	.13	.04	-.24*	-.19	-.06	.25*	.42***	.72***	.40**	.50***	-.06	-.19	-.05	.08
10. General Self-efficacy	-.09	-.13	-.09	-.20	.17	.63***	.42***	.29*	.11	.63***	-.09	-.16	-.07	.07
11. Fluid Intelligence	.26*	.00	.01	.17	.06	-.03	.12	-.05	-.02	.13				
12. Gender (1 = boys)	-.09	.23*	-.14	-.04	-.06	-.29*	.14	-.07	-.07	-.18	-.08			
13. Age	.08	.04	-.03	.09	-.02	-.14	.08	.11	-.01	-.04	.32**	.04		
14. Grade Level (1= fourth)	.26*	.12	-.09	.11	.09	-.05	.27*	.24*	.12	-.02	.42***	.04	.71***	

Note. The diagonal contains the correlations between pretest and posttest. Fluid intelligence was assessed only at pretest. Correlations between fluid intelligence, gender, grade level, and age are reported only once because the correlations did not vary between pretest and posttest.

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

As expected, draws to decision and the decision threshold were significantly correlated, as they both measure the jumping to conclusions bias. Also, as expected, all motivational variables were positively correlated with each other. Interestingly, the confirmation bias was significantly positively correlated with all motivational variables, possibly because of a similar response format (self-assessment questionnaire.) Data-based argumentation at pretest and posttest was significantly positively correlated with fluid intelligence.

Baseline equivalence analyses revealed that 12 out of 14 variables had baseline differences greater than an effect size of $|d| = 0.05$ (see Table 4). So, we included all of these variables as predictors in the regression analyses to control for whether any of the posttest values could be better explained by these baseline differences in contrast to any effects of the intervention.

Table 4

Baseline Differences in the Intervention Group Compared With the Waitlist Control Group

	Baseline differences			
Variable	*g*	*t*	*df*	*p*
Data-based Argumentation	0.08	0.36	76.60	.718
Views on Variability	-0.20	-0.40	74.55	.390
Draws to Decision	0.10	0.47	78.47	.643
Decision Threshold	0.16	0.73	78.60	.467
Alternation Bias	0.02	0.08	70.40	.935
Confirmation Bias	0.30	1.33	75.00	.188
Self-concept	0.01	0.03	78.88	.973
Intrinsic Value	0.07	0.30	78.00	.762
Attainment Value	0.07	0.29	75.00	.776
Self-efficacy	0.17	0.73	75.66	.466
Gender (1 = *boys*)	0.19			
Age	-0.10	-0.42	70.72	.676
Grade Level (1 = *Grade 4*)	0.05			
Fluid Intelligence	0.13	0.58	77.54	.564

Note. Instead of Hedges' g, the Cox index was used for binary variables.

Implementation Fidelity

The analysis of the implementation fidelity of our intervention revealed that most of the course instructors adhered to the course manual at a high-quality level (see Hulleman & Cordray, 2009; Humphrey et al., 2016; O'Donnell, 2008). Table 4 shows the fidelity scores for each of the course instructors. Adherence to the course manual was high (78% to 100%). Also,

the quality of delivery was high, ranging from 2.52 to a maximum value of 3.00. When looking at the core components (see Table 5), the instructor-rated implementation of the predict-observe-explain approach was realized well with a mean value of M = 3.29 (SD = 0.33). Cooperative learning methods were implemented to a high level with a mean value of M = 4.49 (SD = 0.41). When the children in the intervention group were asked how well the core components helped them learn (see Table 6), both core components performed well, but in this case, the predict-observe-explain approach performed better (M = 4.34, SD = 0.66) than the cooperative learning methods (M = 4.00, SD = 0.80). Altogether, this finding implies that the intended intervention was well-implemented.

Effects of the Intervention on Cognitive Aspects of Children's Statistical Literacy

With our confirmatory research question, we asked whether there were positive effects of the intervention on aspects of children's statistical literacy. Figure 4 shows the time by condition plot of all targeted variables. Table 8 presents the results of the regression analyses for the target variables. The full tables with all predictors and additional analyses are included in the Appendix. As expected, statistically significant intervention effects on the posttest values were found for data-based argumentation (B = 0.58, 95% CI [0.25, 0.91], p < .001), the decision threshold (B = 0.81, 95% CI [0.41, 1.21], p < .001), and alternation bias (B = -0.52, 95% CI [-0.89, -0.14], p = .007). However, we did not find significant effects of the intervention on views on variability (B = -0.03, 95% CI [-0.43, 0.38], p = .445), draws to decision (B = 0.41, 95% CI [-0.05, 0.87], p = .060), or confirmation bias (B = -0.26, 95% CI [-0.79, 0.27], p = .198).

Effects of the Intervention on Children's Motivational Beliefs

With our first exploratory research question, we asked whether there were any effects of the intervention on the children's motivational beliefs. The regression analysis results (see Table 9) showed a positive and significant effect of the course on self-concept in data-related tasks (B = 0.58, 95% CI [0.27, 0.89], p < .001) and attainment value (B = 0.46, 95% CI [0.15, 0.78], p = .008), whereas we did not find a statistically significant positive effect on intrinsic value (B = 0.37, 95% CI [-0.15, 0.89], p = .219). Similarly, there was no significant effect on general self-efficacy (B = -0.10, 95% CI [-0.54, 0.33], p = .643). Overall, these findings revealed that some of the primary school children's motivational beliefs were fostered by the intervention in this study.

Table 5

Implementation Fidelity: Adherence and Quality of Delivery

	Course instructors														Total	
	A		B		C		D		E		F		G			
Course unit	Adh. %	Quality *M* (*SD*)	Adh. %	Quality *M* (*SD*)	Adh. %	Quality *M* (*SD*)	Adh. %	Quality *M* (*SD*)	Adh. %	Quality *M* (*SD*)	Adh. %	Quality *M* (*SD*)	Adh. %	Quality *M* (*SD*)	Adh. %	Quality *M* (*SD*)
1	1.00	2.50 (0.58)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	2.25 (0.96)	1.00	2.75 (0.50)	1.00	2.79 (0.29)
2	1.00	2.60 (0.55)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	2.60 (0.55)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	2.60 (0.55)	1.00	2.83 (0.23)
3	1.00	2.60 (0.55)	1.00	3.00 (0.00)	1.00	2.80 (0.45)	0.86	3.00 (0.00)	0.86	3.00 (0.00)	1.00	2.75 (0.50)	1.00	2.00 (0.82)	0.96	2.74 (0.33)
4	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	0.83	3.00 (0.00)	0.71	3.00 (0.00)	0.86	3.00 (0.00)	0.86	2.33 (0.58)	0.89	2.90 (0.08)
5	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	0.80	3.00 (0.00)	0.67	3.00 (0.00)	1.00	2.33 (0.58)	0.83	2.67 (0.58)	0.90	2.86 (0.16)
6	1.00	2.50 (0.71)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	0.80	3.00 (0.00)	1.00	2.33 (0.58)	0.67	3.00 (0.00)	1.00	2.50 (0.71)	0.92	2.76 (0.28)
7	0.83	2.00 (NA)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	0.80	3.00 (0.00)	0.00	NA (NA)	0.83	3.00 (0.00)	1.00	2.33 (0.58)	0.78	2.72 (0.12)
8	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	0.67	3.00 (0.00)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	1.00	3.00 (0.00)	0.95	3.00 (0.00)
Total	0.98	2.65 (0.34)	1.00	3.00 (0.00)	1.00	2.98 (0.06)	0.84	2.95 (0.07)	0.78	2.90 (0.08)	0.92	2.79 (0.25)	0.96	2.52 (0.54)	0.93	2.82 (0.19)

Note. Adh.= adherence (i.e., the percentage of exercises conducted by each course instructor and in each course unit). Quality = quality of delivery. n.a.= no data were available because the corresponding session was not conducted.

Table 6

Descriptive Statistics for Adherence to Core Components Rated by Course Instructors

Construct	*N* items	*N*	*M*	*SD*	α
Predict-observe-explain	3	7	3.29	0.33	.62
Cooperative Learning	3	7	4.49	0.41	.84

Note. *N* = number of course instructors.

Table 7

Descriptive Statistics for Value of Core Components Rated by the Children in the Intervention Group

Construct	*N* items	*N*	*M*	*SD*	α
Predict-observe-explain	3	33	4.34	0.66	.85
Cooperative Learning	4	34	4.00	0.80	.78

Note. *N* = number of children.

Differential Intervention Effects

With another exploratory research question, we wanted to explore whether there were any differential intervention effects by gender, age, grade level, pretest values, or fluid intelligence. None of the 55 differential effects were significant ($.229 < p < .990$; see Appendix). Thus, the intervention was equally effective for all children.

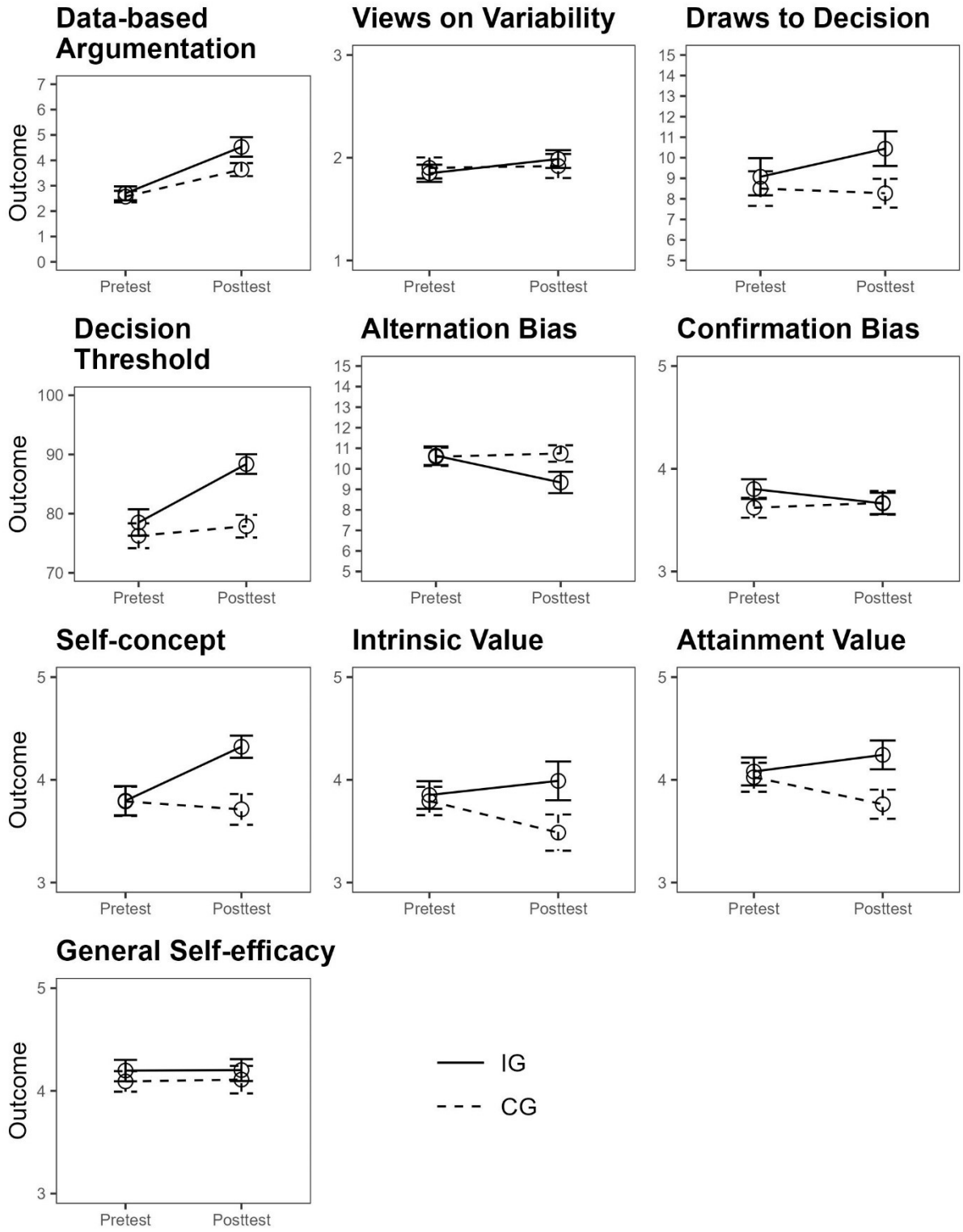
Associations Between Fidelity Measures and Posttest Values

With another exploratory research question, we explored whether the children's posttest values were related to assessments of intervention fidelity within the intervention condition. Results from all 40 models are presented in the Appendix. The findings showed that fidelity did not seem to make a huge difference, as most of the associations were nonsignificant. Higher posttest values were not related to adherence, except for one outcome (draws to decision: $B = 0.49$, 95% CI [0.19, 0.79], $p = .020$), and posttest values were not related to quality of delivery ($.800 \leq p \leq .981$). However, the predict-observe-explain approach was shown to be important for the motivational outcomes.

When looking at ratings of the fidelity of the core components, there were some statistically significant associations with motivational beliefs. The instructor-rated fidelity of the predict-observe-explain approach was positively related to intrinsic value ($B = 0.32$, 95% CI [0.16, 0.48], $p < .001$). The child-rated values of the core component showed a somewhat

Figure 4

Time by Condition Plots of Targeted Variables



Note. IG = Intervention Group, CG = Control Group. This graphic was produced by using the ggplot2 package (Wickham, 2016).

Table 8

Intervention Effects on Statistical Literacy Outcomes Including Baseline Differences

Variable	Outcome				
	B	*SE*	95% CI	*p*	p_{BH}
Data-based Argumentation					
Intervention	0.58***	0.17	[0.25, 0.91]	<.001	<.001
Pretest Value	0.47***	0.13	[0.23, 0.72]	<.001	
R^2	.50				
Views on Variability					
Intervention	-0.03	0.21	[-0.43, 0.38]	.445	.445
Pretest Value	0.40***	0.10	[0.21, 0.59]	<.001	
R^2	.32				
Draws to Decision					
Intervention	0.41	0.24	[-0.05, 0.87]	.040	.060
Pretest Value	0.48*	0.24	[0.01, 0.96]	.023	
R^2	.30				
Decision Threshold					
Intervention	0.81***	0.20	[0.41, 1.21]	<.001	<.001
Pretest Value	0.26*	0.13	[0.02, 0.51]	.018	
R^2	.39				
Alternation Bias					
Intervention	-0.52**	0.19	[-0.89, -0.14]	.003	.007
Pretest Value	0.22	0.19	[-0.16, 0.59]	.126	
R^2	.35				
Confirmation Bias					
Intervention	-0.26	0.27	[-0.79, 0.27]	.165	.198
Pretest Value	0.45**	0.19	[0.07, 0.83]	.010	
R^2	.43				

Note. For the intervention condition, one-tailed significance levels are reported because directional hypotheses were tested. The *p*-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure. The complete table with effects of baseline differences is reported in the Appendix.

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

similar pattern. The predict-observe-explain approach was positively and significantly related to self-concept (B = 0.33, 95% CI [0.11, 0.55], p = .030) and intrinsic value (B = 0.49, 95% CI [0.14, 0.84], p = .047). There were no significant correspondences between ratings of cooperative learning methods and intervention outcomes ($.072 \leq p \leq .998$).

Altogether, the findings suggest that some of the fidelity measures were related to some of the outcome variables. The fidelity measures of the predict-observe-explain approach had the most coherent associations with intrinsic value.

Table 9

Intervention Effects on Motivational Beliefs Including Baseline Differences

Variable	Outcome				
	B	*SE*	95% CI	*p*	p_{BH}
Self-concept					
Intervention	0.58***	0.16	[0.27, 0.89]	<.001	<.001
Pretest Value	0.59***	0.14	[0.31, 0.87]	<.001	
R^2	.51				
Intrinsic Value					
Intervention	0.37	0.26	[-0.15, 0.89]	.164	.219
Pretest Value	0.43**	0.15	[0.14, 0.72]	.002	
R^2	.36				
Attainment Value					
Intervention	0.46**	0.16	[0.15, 0.78]	.004	.008
Pretest Value	0.44**	0.16	[0.12, 0.77]	.004	
R^2	.34				
General Self-efficacy					
Intervention	-0.10	0.22	[-0.54, 0.33]	.643	.643
Pretest Value	0.52***	0.10	[0.33, 0.71]	<.001	
R^2	.46				

Note. Two-tailed significance levels are reported. The *p*-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure. The complete table with effects of baseline differences is reported in the Appendix.

*p < .05. **p < .01. ***p < .001.

Discussion

In this preregistered randomized controlled field trial, we studied the effectiveness of an 8-week course called “Luck or genius? Understanding data and making predictions” on primary school children's statistical literacy. The results showed that the statistical literacy

intervention successfully promoted the majority of the investigated cognitive and motivational aspects of third- and fourth-graders' statistical literacy. The intervention was successfully implemented into practice; it enhanced children's data-based argumentation, decision threshold, data-related self-concept, and attainment value; and it reduced students' alternation bias. No intervention effects were found for views on variability, draws to decision, confirmation bias, intrinsic value, or general self-efficacy. Overall, this study offers further evidence of the malleability of children's statistical literacy in primary school (e.g., Ben-Zvi & Sharett-Amir, 2005; English & Watson, 2013) and the effectiveness of the intervention when put into practice by course instructors from the field. In the following, the intervention effects are discussed in terms of their relevance for theory and practice.

Effects on Cognitive Aspects of Statistical Literacy

With our first research question, we wanted to test whether the intervention promoted cognitive aspects of children's statistical literacy. Three of our six measures (i.e., data-based argumentation, decision threshold, alternation bias) were significantly affected as intended. As preregistered, the effect sizes were interpreted as Cohen's *d* (small: $|d| = 0.20$, medium: $|d| = 0.50$, large: $|d| = 0.80$; Cohen, 1988). However, educational interventions usually achieve much smaller effects and could also be interpreted in accordance with Kraft (2020; small: $|d| < 0.05$, medium: $0.05 < |d| < 0.20$, large: $|d| > 0.20$).

With our intervention, we wanted to promote children's data-based argumentation skills to increase the odds that they will be able to effectively participate in social discourse and make informed decisions in their lives. In line with our expectations, data-based argumentation was positively affected by our intervention with a medium effect size. As in our previous study (Stark et al., 2025b), the children in the intervention group were able to develop more data-based arguments than the children in the control group.

Contrary to our expectations, the draws to decision variable was not significantly promoted by our intervention. In our previous study (Stark et al., 2025b), we found a large positive effect of the intervention on draws to decision (see Table 11). In this study, there was a nonsignificant medium-sized intervention effect. As the intervention was implemented under less standardized conditions, a smaller effect size was expected. With a larger sample size, it seems plausible to assume that the effect would have become significant. This view is supported by the significant intervention effect on the decision threshold. In line with our expectations, the decision threshold was positively affected by our intervention with a large effect size. Both measures are indicators of the jumping to conclusions bias. Therefore, we can

assume that the children in the intervention condition tended to jump to conclusions less often than the children in the control condition. We can also see that the children did not become too cautious, as their decision threshold grew from M = 78.44% (SD = 14.10%) to M = 88.22% (SD = 9.63%), which is comparable to a healthy adult's decision threshold (e.g., Moritz et al., 2020) and to an alpha level of 11.78%, which is not too conservative in science. This finding shows that children's statistical literacy was strengthened, as they usually tend to be overconfident in their decisions (e.g., Lipko et al., 2009). This overconfidence has been previously found to be related to superstitious and conspiracy-related beliefs (e.g., Kuhn et al., 2022), and early interventions have been called for (Gregersen et al., 2022).

As expected, the alternation bias was reduced by our intervention. There was a statistically significant medium-sized intervention effect on the alternation bias. Children in the intervention group had a mean of M = 10.64 (SD = 2.64) alternations between heads and tails at pretest and M = 9.33 (SD = 2.99) at posttest, whereas children in the waitlist control condition had a mean of M = 10.59 (SD = 2.77) alternations at pretest and M = 10.74 (SD = 2.18) at posttest. A completely unbiased sample would expect 9.50 alternations in 20 coin throws. Thus, the intervention group went from expecting more than one alternation too much to expecting just 0.17 less than an unbiased sample would, whereas the control condition's value did not change. So, the intervention may actually help to un-bias participants instead of giving them an opposing bias (i.e., not expecting enough alternations).

There was no effect on children's views on variability. Therefore, the present intervention did not seem to affect how children expect variability when asked to generate random distributions. In the previous study (Stark et al., 2025b), there was a significant large positive effect on children's views on variability. As this measurement instrument captures something similar to the alternation bias, we would have expected both scales to be affected if one was. However, perhaps the views on variability were not affected as much because the context deviated further from the intervention's content.

There was no effect of the intervention on the confirmation bias. However, a reduction in the confirmation bias is possible in the context of STEM education. Hane and Brister (2022) found that STEM university students showed a lower confirmation bias when they were helped to recognize this bias in their own thinking. It is possible that it is harder to counteract the confirmation bias in primary school children than in university students, as children show a strong tendency to be overconfident in their views (e.g., Lipko et al., 2009). It is also possible that there was no lasting effect on the confirmation bias, as it was targeted in only one of the earlier sessions of the intervention. Perhaps the intervention should focus more often and more

deeply on the confirmation bias. Further research is needed to determine how and how much the confirmation bias can be reduced in primary school children.

Analysis of differential effects for children of different genders, ages, grade levels, prior knowledge, and fluid intelligence did not yield any statistically significant effects, thus supporting the notion that, in general, the intervention works similarly for all children. However, given the sample size and power considerations, this conclusion should be somewhat tentative, especially in light of findings from other studies with similar interventions (e.g., Rebholz et al., 2022; Schiefer et al., 2021) that worked better for some subpopulations rather than others.

So, what can our study contribute to the question of whether statistical literary can be promoted in primary school? Altogether, the results of this study strengthen the existing evidence of the malleability of primary school children's data-based argumentation (Bakker & Gravemeiher, 2004; Ben-Zvi, 2006; Paparistodemou & Meletiou-Mayrotheris, 2008) and their ability to deal with variability (e.g., English & Watson, 2013; Lehrer & Kim, 2009; Piaget & Inhelder, 1976; Watson & Kelly, 2005; Watson & Moritz, 2000) and show the effectiveness of our intervention.

Effects on Motivational Aspects of Statistical Literacy

Because motivational beliefs are seen as crucial for statistical literacy (e.g., Gal, 2002; Watson & Callingham, 2003), we explored whether our intervention affected primary school children's motivational beliefs. We did not expect positive effects of the intervention on motivational beliefs because reference group effects might lead to lower motivation in the intervention group (Preckel et al., 2010; Zeidner & Schleyer, 1999). Nevertheless, two out of four motivational variables were positively affected by our intervention.

We found a medium-sized positive intervention effect on children's self-concept in data-related tasks. This finding means that children from the intervention group gained higher feelings of confidence in their abilities in data-related tasks, even though self-concept is a relatively stable construct (Bong & Skaalvik, 2003). As self-concept is based on previous experiences (Bong & Skaalvik, 2003), children in the intervention condition might have had positive experiences regarding their own perceived competence, which led to the change in their self-concept. This change, in turn, might also impact their achievement in the domain, as there are reciprocal effects between self-concept and achievement (Marsh et al., 2005).

Attainment value was also positively affected by our intervention with a medium effect size. This finding means that children in the intervention condition changed to consider data to

be more important after the intervention than the students in the control group did. Therefore, according to Eccles' expectancy-value theory (e.g., Eccles & Wigfield, 2020), the children in the intervention condition should be more likely to execute data-related behavior because they consider it to be a meaningful activity.

By contrast, there was no intervention effect on children's intrinsic value in data-related tasks. This finding suggests that, relative to the control group, the enjoyment of data-related tasks remained unchanged in the intervention group. However, child-rated and instructor-rated fidelity of the predict-observe-explain approach was associated with higher outcomes on intrinsic value in the intervention group. This finding implies that the better the predict-observe-explain approach was implemented, the more intrinsically motivated the children became.

There was no intervention effect on general self-efficacy. As statistical knowledge can be applied to many different aspects of a person's own life, an effect on general self-efficacy could have been conceivable. However, it is possible that it will take longer to see lasting effects on general self-efficacy. Moreover, the finding that general self-efficacy was not affected also suggests that the children from the intervention condition did not simply rate all the items higher due to social desirability (Dodou & de Winter, 2014).

Altogether, our findings provide evidence that some dimensions of statistics motivation can be enhanced even before tertiary education (e.g., Gopal et al., 2018; Khoshnoodifar, Ashouri, & Taheri, 2023; Krause, Stark, & Mandl, 2009).

Putting the Intervention Into Practice

Amidst the growing emphasis on investigating interventions and their implementation (Hulleman & Cordray, 2009; Humphrey et al., 2016; Lendrum & Humphrey, 2012), our study serves as a demonstration that a theoretically derived intervention can effectively be implemented in the real world. In the previous efficacy study (Stark et al., 2025b), scientific staff conducted the intervention and strictly adhered to the intervention manual. In the current effectiveness study, a diverse sample of nonscientific course instructors from the field conducted the intervention. Because of the diversity in the instructors, we expected greater variability in how the intervention was conducted. Therefore, we assessed several fidelity measures to explore whether the intervention was implemented as intended (Hulleman & Cordray, 2009; Humphrey et al., 2016) and to potentially explain differences in the intervention's effects.

Adherence to the course manual and implementation quality generally received high ratings from all course instructors. Also, the implementation of the core components was highly rated by the course instructors and children. Therefore, we can assume that the intervention was implemented as intended and that the above described intervention effects were due to the intervention.

For the cognitive aspects of statistical literacy, we could replicate only one of the three significant effects of the intervention. Whereas the effect on data-based argumentation was maintained, the effects on views on variability and draws to decision were much smaller and were nonsignificant in the effectiveness study. However, two out of three effects on the new cognitive measures were significant. The intervention significantly affected children's decision threshold and alternation bias but not their confirmation bias.

For motivational beliefs, we were able to replicate the effect on self-concept in data-related tasks. Also, there was no effect on intrinsic value in either study. However, there was no effect on attainment value in the efficacy study, but this effect turned out to be positive in the effectiveness study. Therefore, two out of three effects on domain-specific motivational beliefs were significant, but there was no effect on general self-efficacy.

Altogether, we were partially successful in replicating the effects when moving from the efficacy study to the effectiveness study. Because the intervention was implemented under less standardized conditions, some of the effects were not replicated.

Strength and Limitations

Our study revealed beneficial effects of an extracurricular intervention on children's cognitive and motivational aspects of statistical literacy. One noTable Atrength of this study was its design: a randomized controlled field trial coupled with a waitlist control group. Randomized controlled trials are considered the gold standard in educational research (Torgerson & Torgerson, 2008). This methodology not only enhances the rigor of a study but also facilitates a robust evaluation of the intervention's effects.

However, some limitations should be considered. First, we used a sample of children from an extracurricular enrichment program. The use of a unique sample limits the generalizability of our results to that specific group of children. It is possible that mathematically talented children benefit more from such interventions on the basis of their higher order reasoning skills (Greenes, 1981). Therefore, this sample may have provided somewhat optimal conditions for testing the effectiveness of our intervention. At the same time, intervention studies often find more pronounced effects for students whose initial scores are

comparably low (e.g., Clemens et al., 2019; Herbein et al., 2018). Moreover, we did not find any statistically significant interaction effects between initial scores on our core constructs and learning gains, which might signal that our intervention would be effective in less selective samples; however, the power to detect any such interaction effects was arguably fairly low. Consequently, to get a better understanding of generalizability, future research could focus on testing this intervention in a broader target group.

Second, our sample of 87 third- and fourth-graders from nine local sites was relatively small. On the basis of our preregistered power analysis, we wanted to recruit between eight and 22 local sites of the HCAP to replicate the findings from our previous study (Stark et al., 2025b). So, even though we reached the lower end of this goal, some effects might not have been detectable due to low statistical power.

Third, even though our intervention was based on the predict-observe-explain approach (Gunstone & White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015), our study design did not allow us to conclude that the observed effects were due to these two core components. However, we applied several different fidelity measures so that we could gain insights from different perspectives, as often suggested (Schultes et al., 2015). Adherence and quality of delivery generally received high ratings. Therefore, we can assume that our intervention was implemented successfully. Additionally, for the predict-observe-explain approach, we found associations between two intervention outcomes and the fidelity measures in the intervention condition. No such associations were found for the implementation of cooperative learning methods. These findings imply that higher implementation of the predict-observe-explain approach was related to higher self-concept and intrinsic value at posttest. However, future studies should use study designs that enable researchers to identify the intervention's causal mechanisms.

Finally, our findings do not include any long-term effects on children's statistical literacy. Posttests took place 1 day to 1 week after the last intervention session. However, other outcomes, such as children's achievement in school, interest in science, and vocational choices could be of interest (Robertson et al., 2010). Therefore, future studies should include follow-up tests at later points in time to identify any lasting effects of the intervention.

Conclusion

This study provides evidence that cognitive and motivational aspects of statistical literacy can be fostered at the primary school level and that our newly developed statistical literacy intervention is still effective when implemented in the field. There were positive

intervention effects on children's data-based argumentation, decision threshold, self-concept, and attainment value. Also, their alternation bias was reduced. Future studies could focus on applying the intervention on a larger scale or for a broader target group.

References

Abry, T., Hulleman, C. S., & Rimm-Kaufman, S. E. (2015). Using indices of fidelity to intervention core components to identify program active ingredients. *American Journal of Evaluation*, *36*(3), 320-338. http://dx.doi.org/10.1177'/1098214014557009

Aiken, L. S., West, S. G., & Pitts, S. C. (2003). Multiple linear regression. In Handbook of Psychology (pp. 481–507). John Wiley & Sons, Inc. https://doi.org/10.1051/eas/1466005

Arens, A. K., Trautwein, U., & Hasselhorn, M. (2011). Erfassung des Selbstkonzepts im mittleren Kindesalter: Validierung einer deutschen Version des SDQ I [Assessment of self-concept in middle childhood: Validation of a German version of the SDQ I]. *Zeitschrift für Pädagogische Psychologie.* https://doi.org/10.1024/1010-0652/a000030

Asparouhov, T., & Muthén, B. O. (2006). Multilevel modeling of complex survey data. Proceedings of the Joint Statistical Meeting, 2718–2726. https://doi.org/10.1111/j.1467-985X.2006.00426.x

Bakker, A., & Gravemeijer, K.P.E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp.147-168). Dordrecht, the Netherlands: Kluwer Academic Publishers. Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science, 333*(6045), 975-978. https://doi.org/10.1126/science.1204534

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU)[An instrument for measuring subjective competency expectations: The Short Scale for Measuring General Self-efficacy Beliefs (ASKU)]. *Köln, Germany: GESIS–Leibniz Intitut für Sozialwissenschaften.* https://doi.org/10.23668/psycharchives.418

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman, & B. Chance (Eds.), Proceedings of the Seventh International Conference on Teaching Statistics [CD-ROM]. Voorburg, The Netherlands: International Association for Statistics Education. https://www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/2D1_BENZ.pdf

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004). The challenge of developing statistical literacy, reasoning and thinking (pp. 3-16). Dordrecht: Kluwer academic publishers.

Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. School Science and Mathematics, 108(8), 355-361.

Ben-Zvi, D., & Sharett-Amir, Y. (2005). How do primary school students begin to reason about distributions? In K. Makar (Ed.), Proceedings of SRTL-4. University of Queensland.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really?. *Educational Psychology Review*, *15*, 1-40. https://www.jstor.org/stable/23361533

Capar, G., & Tarim, K. (2015). Efficacy of the cooperative learning method on mathematics achievement and attitude: A meta-analysis research. *Educational Sciences: Theory and Practice*, *15*(2), 553-559. https://doi.org/10.12738/estp.2015.2.2098

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*, 1-9. http://dx.doi.org/10.1186/1748-5908-2-40

Chan, S. W., & Ismail, Z. (2013). Assessing misconceptions in reasoning about variability among high school students. *Procedia - Social and Behavioral Sciences*, *93*, 1478-1483.

Clemens, N. H., Oslund, E., Kwok, O. M., Fogarty, M., Simmons, D., & Davis, J. L. (2019). Skill moderators of the effects of a reading comprehension intervention. *Exceptional Children*, *85*(2), 197-211. https://doi.org/10.1177/0014402918787339

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801-823. https://doi.org/10.2307/2975286

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Edition). Hillsdale, NJ: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Cox, D. R. (1970). Analysis of binary data. New York: Chapman & Hall/CRC. https://doi.org/10.1201/9781315137391

Dodou, D., & de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. Computers in Human Behavior, 36, 487-495. https://doi.org/10.1016/j.chb.2014.04.005

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997a). The effect of self-referent material on the reasoning of people with delusions. British Journal of Clinical Psychology, 36, 575–584. https://doi.org/10.1111/j.2044-8260.1997.tb01262.x

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997b). Normal and abnormal reasoning in people with delusions. British Journal of Clinical Psychology, 36, 243–258. https://doi.org/10.1111/j.2044-8260.1997.tb01410.x

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?. *International Journal of Public Health*, *58*(4), 637-642. https://doi.org/10.1007/s00038-012-0416-3

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*(5), 713–740. https://doi.org/10.1177/00131640121971482

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, *16*(1), 44-49. https://doi.org/10.52041/serj.v16i1.213

English, L., & Watson, J. (2013). Beginning inference in fourth grade: Exploring variation in measurement. Mathematics Education Research Group of Australasia, 274-281. https://doi.org/10.1186/s40594-015-0016-x

Fisher, L. D., Dixon, D. O., Herson, J., Frankowski, R. F., & Pearce, K. E. (1990). Analysis of randomized clinical trials: Intention to treat. Statistical Issues in Drug Research and Development, 331, 331-345.

Fixsen, D., Blase, K., Metz, A., & Van Dyke, M. (2013). Statewide implementation of evidence-based programs. *Exceptional Children*, *79*(2), 213-230. https://doi.org/10.1177/001440291307900206

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33(3), 613-619. https://doi.org/10.1177/001316447303300309

Friedman, L. M., Furberg, C., & DeMets, D. L. (2010). Fundamentals of clinical trials (4th ed.). New York, NY: Springer.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. International statistical review, 70(1), 1-25. https://doi.org/10.2307/1403713

Gal, I. (2019). Understanding statistical literacy: about knowledge of contexts and models. In J. M. Contreras, M. M. Gea, M. M. López-Martín & E. Molina-Portillo (Eds.), *Actas*

del Tercer Congreso Internacional Virtual de Educación Estadística. www.ugr.es/local/fqm126/civeest.html

Garety, P. A., Freeman, D., Jolley, S., Dunn, G., Bebbington, P. E., Fowler, D. G., ... & Dudley, R. (2005). Reasoning, emotions, and delusional conviction in psychosis. *Journal of Abnormal Psychology*, *114*(3), 373. https://doi.org/10.1037/0021-843X.114.3.373

Garfield, J. (1993). Teaching statistics using small-group cooperative learning. Journal of Statistics Education, *1*(1). https://doi.org/10.1080/10691898.1993.11910455

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, *4*(1), 92-99.

Gaspard, H., Dicke, A. L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, *51*(9), 1226. https://doi.org/10.1037/dev0000028

Giraud, G. (1997). Cooperative learning and statistics instruction. *Journal of Statistics Education*, *5*(3). https://doi.org/10.1080/10691898.1997.11910598

Gopal, K., Salim, N. R., & Ayub, A. F. M. (2018, October). RStudio as a tool to motivate students to learn statistics: A study in a Malaysian public university. In *AIP Conference Proceedings* (Vol. 2013, No. 1). AIP Publishing. https://doi.org/10.1063/1.5054226

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science, 16*(7), 893–926. http://dx.doi.org/10.1007/s11121-015-0555-x.

Greene, J. A. (2015). Serious challenges require serious scholarship: Integrating implementation science into the scholarly discourse. *Contemporary Educational Psychology, 40*, 112–120. https://doi.org/10.1016/j.cedpsych.2014.10.007.

Greenes, C. (1981). Identifying the gifted student in mathematics. *The Arithmetic Teacher*, *28*(6), 14-17.

Gregersen, M., Rohd, S. B., Jepsen, J. R. M., Brandt, J. M., Søndergaard, A., Hjorthøj, C., ... & Hemager, N. (2022). Jumping to conclusions and its associations with psychotic experiences in preadolescent children at familial high risk of schizophrenia or bipolar disorder-The Danish high risk and resilience study, VIA 11. *Schizophrenia Bulletin*, *48*(6), 1363-1372. https://doi.org/10.1093/schbul/sbac060

Gunstone, R. F., & White, R. T. (1981). Understanding of gravity. *Science Education*, *65*(3), 291-299. https://doi.org/10.1002/sce.3730650308

Gustina, R., Hastuti, I. D., Nizaar, M., & Syaharuddin, S. (2023). Predict Observe Explain Learning Model: Implementation and Its Influence on Students' Critical Thinking Ability and Learning Outcomes (A Meta-Analysis Study). *Jurnal Kependidikan: Jurnal Hasil Penelitian dan Kajian Kepustakaan di Bidang Pendidikan, Pengajaran dan Pembelajaran*, *9*(2). https://doi.org/10.33394/jk.v9i2.7388

Hales, A. H. (2023). One-Tailed Tests: Let's Do This (Responsibly). *Psychological Methods*. https://doi.org/10.1037/met0000610

Hane, E. N., & Brister, E. (2022). A classroom intervention to reduce confirmation bias. CourseSource. https://doi.org/10.24918/cs.2022.7

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107-128. https://doi.org/10.2307/1164588

Herbein, E., Golle, J., Tibus, M., Zettler, I., & Trautwein, U. (2018). 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. https://doi.org/10.1016/j.cedpsych.2018.09.003

Hood, M., Creed, P. A., & Neumann, D. L. (2012). Using the expectancy value model of motivation to understand the relationship between student attitudes and achievement in statistics. *Statistics Education Research Journal*, *11*(2), 72-85. https://doi.org/10.52041/serj.v11i2.330

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88-110. http://dx.doi.org/10.1080/19345740802539325

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook. *Education Endowment Foundation*, *1-32*. http://dx.doi.org/10.1017/CBO9781107415324.004

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*(4), 269-307. https://doi.org/10.1207/S15327833MTL0204_3

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430-454.

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Khoshnoodifar, M., Ashouri, A., & Taheri, M. (2023). Effectiveness of gamification in enhancing learning and attitudes: a study of statistics education for health school students. *Journal of Advances in Medical Education & Professionalism*, *11*(4), 230. https://doi.org/10.30476/jamp.2023.98953.1817

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Krause, U. M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and Instruction*, *19*(2), 158-170. http://dx.doi.org/10.1016/j.learninstruc.2008.03.003

Krummenauer, J., & Kuntze, S. (2018). Primary student's data-based argumentation – an empirical reanalysis. In Bergqvist, E., Österholm, M., Granberg, C., & Sumpter, L. (Eds.), *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 251–258). Umeå, Sweden: PME.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160.

Kuhn, S. A. K., Lieb, R., Freeman, D., Andreou, C., & Zander-Schellenberg, T. (2022). Coronavirus conspiracy beliefs in the German-speaking general population: endorsement rates and links to reasoning biases and paranoia. *Psychological Medicine*, *52*(16), 4162-4176. https://doi.org/10.1017/S0033291721001124

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research, 86*(2), 602-640. https://doi.org/10.3102/0034654315617832

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal*, *21*(2), 116-133. https://doi.org/10.1007/BF03217548

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology*, *103*(2), 152-166. https://doi.org/10.1016/j.jecp.2008.10.002

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of

causal ordering. Child development, 76(2), 397-416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

McKenzie, J. D., Jr. (2004). Conveying the Core Concepts. In ASA Section on Statistical Education. (pp. 2755 - 2757). http://www.statlit.org/pdf/2004mckenzieasa.pdf

Moore, D. (1990). Uncertainty. In L. A. Steen (Ed.), *On the Shoulders of Giants: New Approaches in Numeracy* (pp. 95-137). Washington: National Academy Press.

Moritz, S., Scheunemann, J., Lüdtke, T., Westermann, S., Pfuhl, G., Balzan, R. P., & Andreou, C. (2020). Prolonged rather than hasty decision-making in schizophrenia using the box task. Must we rethink the jumping to conclusions account of paranoia*?*. *Schizophrenia Research*, *222*, 202-208.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research*, *39*, 374-396. https://doi.org/10.1007/s11414-012-9295-x

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

Oberski, D. (2014). lavaan.survey: An R package for complex survey analysis of structural equation models. *Journal of Statistical Software*, *57*(1), 1-27. https://doi.org/10.18637/jss.v057.i01

Odom, S. L., Fleming, K., Diamond, K., Lieber, J., Hanson, M., Butera, G., ... & Marquis, J. (2010). Examining different forms of implementation and in early childhood curriculum research. *Early Childhood Research Quarterly, 25*(3), 314-328.

OpenAI. (2024). *ChatGPT* (Dec 10 version) [Large language model]. https://chat.openai.com/chat

Özdemir, D. A., & Işiksal Bostan, M. (2021). Mathematically gifted students' differentiated needs: what kind of support do they need?. *International Journal of Mathematical Education in Science and Technology*, *52*(1), 65-83.

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, 7(2), 83-106. https://doi.org/10.52041/serj.v7i2.471

Preckel, F., Götz, T., & Frenzel, A. (2010). Ability grouping of gifted students: Effects on academic self-concept and boredom. *British Journal of Educational Psychology*, *80*(3), 451-472. https://doi.org/10.1348/000709909X480716

Piaget, J., & Inhelder, B. (1976). The origin of the idea of chance in children (Leake, P. B., Fishbein, H. D., & Leake, L., Trans.). W. W. Norton and Company, Inc.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., . . . Schiefele, U. (Eds.). (2006). *PISA 2003. Dokumentation der Erhebungsinstrumente [PISA 2003. Documentation of assessment instruments]*. Münster, Germany: Waxmann.

Rassin, E. (2008). Individual differences in the susceptibility to confirmation bias. *Netherlands Journal of Psychology, 64*, 87-93. https://doi .org/10.1007/BF03076410

Reading, C., & Shaughnessy, M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), The Challenge of Developing Statistical Literacy, Reasoning and Thinking (pp. 201-226). Dordrecht: Kluwer Academic Publishers.

Rebholz, F., Golle, J., Tibus, M., Ruth-Herbein, E., Moeller, K., & Trautwein, U. (2022). Getting fit for the Mathematical Olympiad: positive effects on achievement and motivation?. *Zeitschrift für Erziehungswissenschaft*, *25*(5), 1175-1198. https://doi.org/10.1007/s11618-022-01106-y

Robertson, K. F., Smeets, S., Lubinski, D., & Benbow, C. P. (2010). Beyond the threshold hypothesis even among the gifted and top math/science graduate students, cognitive abilities, vocational interests, and lifestyle preferences matter for career choice, performance, and persistence. *Current Directions in Psychological Science*, *19*, 346–351. https://doi.org/10.1177/0963721410391442

Rosseel Y (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1–36. doi: 10.18637/jss.v048.i02.

Scherrer, V., & Preckel, F. (2019). Development of motivational variables and self-esteem during the school career: A meta-analysis of longitudinal studies. *Review of Educational Research, 89*(2), 211-258. https://doi.org/10.3102/0034654318819127

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. https://doi.org/10.1037/edu0000630

Schield, M. (1999). Statistical literacy: Thinking critically about statistics. *Of Significance, 1*(1), 15-20.

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of

implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38*, 32-43. http://dx.doi.org/10.1007/s10488-010-0321-0

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2020). Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 5. Bis 7. Klasse (BEFKI 5-7). Hogrefe.

Schroeders, U., Schipolowski, S., Zettler, I., Golle, J., & Wilhelm, O. (2016). Do the smart get smarter? Development of fluid and crystallized intelligence in 3rd grade. Intelligence, 59, 84–95. https://doi.org/10.1016/j.intell.2016.08.00

Schultes, M. T., Jöstl, G., Finsterwald, M., Schober, B., & Spiel, C. (2015). Measuring intervention fidelity from different perspectives with multiple methods: The Reflect program as an example. *Studies in Educational Evaluation*, *47*, 102-112. https://doi.org/10.1016/j.stueduc.2015.10.001

Schutz, P. A., Drogosz, L. M., White, V. E., & Distefano, C. (1998). Prior knowledge, attitude, and strategy use in an introduction to statistics course. *Learning and Individual Differences*, *10*(4), 291-308. https://doi.org/10.1016/S1041-6080(99)80124-1

Sharma, S. (2017). Definitions and models of statistical literacy: a literature review. *Open Review of Educational Research*, *4*(1), 118-133. https://doi.org/10.1080/23265507.2017.1354313

Stalder, U. M. (2013). *Leselust in Risikogruppen: Gruppenspezifische Wirkungszusammenhänge [Reading pleasure in risk groups: Group-specific interdependencies]*. Springer Fachmedien. https://doi.org/10.1007/978-3-658-01701-9

Stark, L., Goecke, B., Jaggy, A.-K., Krummenauer, J., Kuntze, S., Golle, J., Nagengast, B. (2025a). *Assessing decision thresholds in primary school students using signal detection theory: Validating an adapted version of the beads task*. Manuscript in preparation.

Stark, L., Krummenauer, J., Jaggy, A.-K., Kremer, F., Kuntze, S., Nagengast, B., Trautwein, U., Golle, J. (2025b). *Evaluating the efficacy of a statistical literacy intervention.* Manuscript in preparation.

Torgerson, C. J., and D. J. Torgerson. (2008). *Designing Randomised Controlled Trials in Health, Education, and the Social Sciences: An Intr.* New York: Palgrave Macmillan

Trautwein, U., Golle, J., Jaggy, A. K., 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. https://doi.org/10.1111/nyas.15074

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105. https://doi.org/10.1037/h0031322

United Nations Economic Commission for Europe (2012). Making Data Meaningful Part 4: How to improve statistical literacy: A guide for statistical organizations, Geneva. Retrieved from: https://www.unece.org/fileadmin/DAM/stats/documents/writing/Making_Data_Meaningf ul_Part_4_for_Web.pdf

Van Dijke-Droogers, M., Drijvers, P., & Tolboom, J. (2017). Enhancing statistical literacy. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10, February 1 –5,2017)* (pp. 860–867). Dublin, Ireland: DCU Institute of Education and ERME. https://hal.science/hal-01927707

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, *88*(421), 1-8. https://doi.org/10.1080/01621459.1993.10594283

Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46.

Watson, J., & Callingham, R. (2020). COVID-19 and the need for statistical literacy. *Australian Mathematics Education Journal*, *2*(2), 16-21.

Watson, J. M., & Kelly, B. A. (2005). The winds are variable: Student intuitions about variation. *School Science and Mathematics*, *105*(5), 252-269.

Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *The Journal of Mathematical Behavior, 19*(1), 109-136.

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics*, *96*(1), 33-47.

What Works Clearinghouse. (2022). Procedures and standards handbook version 5.0. What Works Clearinghouse. Retrieved from https://ies.ed.gov/ncee/wwc/Handbooks

Wickens, T. (2002). *Elementary signal detection theory.* New York, NY: Oxford University Press.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

Yahya, A. H., & Sukmayadi, V. (2020). A review of cognitive dissonance theory and its relevance to current social issues. MIMBAR: *Jurnal Sosial Dan Pembangunan*, *36*(2), 480-488. https://doi.org/10.29313/mimbar.v36i2.6652

Zeidner, M., & Schleyer, E. J. (1999). 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. https://doi.org/10.1006/ceps.1998.0985

Appendix

Table A1

Intervention Effects on Statistical Literacy Outcomes

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Intervention	0.62	0.31	[0.00; 1.23]	.048	.076	0.05	0.20	[-0.34;0.44]	.788	.788	0.52	0.25	[0.03;1.01]	.036	.076
R^2	0.09					0.00					0.07				
Model 2															
Intervention	0.59*	0.24	[0.11;1.06]	.016	.048	0.06	0.17	[-0.26;0.39]	.711	.711	0.45	0.23	[-0.01;0.90]	.054	.081
Pretest Value	0.57***	0.11	[0.37;0.78]	<.001		0.35*	0.15	[0.06;0.65]	.019		0.39*	0.19	[0.02;0.76]	.041	
R^2	0.39					0.12					0.22				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A1

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Intervention	0.86*	0.30	[0.28; 1.45]	.004	.024	-0.49	0.25	[-0.98; 0.00]	.051	.076	-0.18	0.35	[-0.87; 0.51]	.604	.725
R^2	0.19					0.06					0.01				
Model 2															
Intervention	0.77*	0.26	[0.26;1.29]	.003	.018	-0.51	0.23	[-0.97;-0.05]	.028	.056	-0.32	0.30	[-0.90;0.26]	.278	.334
Pretest Value	0.22*	0.10	[0.02;0.42]	.033		0.18	0.17	[-0.16; 0.52]	.296		0.53***	0.15	[0.24;0.82]	<.001	
R^2	0.22					0.09					0.28				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A2

Intervention Effects on Motivational Beliefs

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Intervention	0.56*	0.20	[0.18; 0.94]	.004	.016	0.45	0.29	[-0.12; 1.02]	.120	.160	0.48	0.24	[0.00; 0.96]	.049	.098
R^2	0.08					0.05					0.06				
Model 2															
Intervention	0.61**	0.20	[0.22; 1.00]	.002	.008	0.46	0.31	[-0.15;1.08]	.140	.187	0.53*	0.21	[0.13;0.94]	.010	.020
Pretest Value	0.55***	0.11	[0.34; 0.76]	<.001		0.42**	0.13	[0.17;0.67]	.001		0.40*	0.16	[0.09;0.71]	.011	
R^2	0.40					0.23					0.23				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A2

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Intervention	-0.03	0.21	[-0.45;0.38]	.875	.875
R^2	0.00				
Model 2					
Intervention	-0.09	0.23	[-0.54;0.36]	.696	.696
Pretest Value	0.59***	0.05	[0.48;0.70]	<.001	
R^2	0.36				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A3

Intervention Effects on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.58***	0.17	[0.25; 0.91]	<.001	<.001	-0.03	0.21	[-0.43; 0.38]	.445	.445	0.41	0.24	[-0.05;0.87]	.040	.060
Data-based Argumentation	0.47	0.13	[0.23; 0.72]	<.001		0.25	0.08	[0.08; 0.42]	.003		-0.03	0.09	[-0.20;0.14]	.736	
Views on Variability	-0.01	0.13	[-0.27; 0.24]	.907		0.40	0.10	[0.21; 0.59]	<.001		-0.05	0.06	[-0.17;0.07]	.419	
Draws to Decision	0.01	0.14	[-0.27; 0.30]	.918		-0.11	0.14	[-0.37; 0.16]	.435		0.48	0.24	[0.01;0.96]	.023	
Decision Threshold	-0.06	0.19	[-0.42; 0.30]	.752		0.14	0.19	[-0.23; 0.50]	.463		-0.11	0.15	[-0.40;0.18]	.467	
Confirmation Bias	-0.15	0.06	[-0.27;-0.02]	.022		0.22	0.09	[0.05; 0.40]	.012		0.16	0.11	[-0.06;0.37]	.153	
Intrinsic Value	0.07	0.19	[-0.30; 0.43]	.716		-0.01	0.08	[-0.17; 0.15]	.900		0.19	0.16	[-0.12;0.50]	.234	
Attainment Value	-0.13	0.12	[-0.36; 0.10]	.268		0.06	0.15	[-0.23; 0.34]	.707		-0.02	0.15	[-0.31;0.27]	.885	
General Self-efficacy	-0.21	0.11	[-0.42; 0.01]	.057		-0.20	0.11	[-0.41; 0.01]	.062		-0.05	0.11	[-0.27;0.17]	.644	
Gender (1 = male)	-0.24	0.19	[-0.61; 0.13]	.207		-0.46	0.21	[-0.86;-0.05]	.028		0.03	0.24	[-0.43;0.50]	.891	
Age	0.07	0.13	[-0.19; 0.33]	.609		-0.17	0.19	[-0.54; 0.21]	.388		-0.16	0.10	[-0.35;0.03]	.096	
Fluid Intelligence	0.19	0.09	[0.01; 0.36]	.034		-0.03	0.24	[-0.50; 0.45]	.915		0.07	0.14	[-0.22;0.35]	.648	
R^2	0.50					0.32					0.30				

Note. For the intervention condition one-tailed significance levels are reported because directional hypotheses were tested. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A3

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.81***	0.20	[0.41;1.21]	<.001	<.001	-0.52**	0.19	[-0.89;-0.14]	.003	.006	-0.26	0.27	[-0.79;0.27]	.165	.198
Data-based Argumentation	-0.02	0.14	[-0.29;0.25]	.886		-0.09	0.12	[-0.33; 0.15]	.455		-0.10	0.17	[-0.43;0.24]	.563	
Views on Variability	0.10	0.07	[-0.03;0.23]	.126		0.13	0.11	[-0.08; 0.35]	.214		0.22	0.13	[-0.04;0.47]	.092	
Draws to Decision	0.15	0.11	[-0.07;0.36]	.176		-0.21	0.13	[-0.47; 0.04]	.101		-0.07	0.13	[-0.32;0.18]	.579	
Decision Threshold	0.26	0.13	[0.02;0.51]	.018		-0.01	0.07	[-0.14; 0.12]	.877		0.00	0.11	[-0.21;0.21]	.990	
Confirmation Bias	0.08	0.18	[-0.27;0.42]	.656		0.11	0.11	[-0.11; 0.33]	.319		0.45	0.19	[0.07;0.83]	.010	
Intrinsic Value	-0.25	0.21	[-0.66;0.17]	.241		-0.09	0.20	[-0.49; 0.31]	.646		-0.07	0.20	[-0.46;0.32]	.717	
Attainment Value	0.33	0.13	[0.07;0.58]	.011		-0.25	0.12	[-0.47;-0.02]	.033		-0.01	0.16	[-0.33;0.31]	.951	
General Self-efficacy	0.19	0.17	[-0.14;0.51]	.261		0.07	0.10	[-0.13; 0.26]	.515		0.16	0.08	[0.00;0.32]	.050	
Gender (1 = male)	-0.17	0.26	[-0.67;0.34]	.523		0.49	0.18	[0.15; 0.84]	.005		-0.22	0.36	[-0.92;0.48]	.532	
Age	0.08	0.11	[-0.13;0.29]	.456		-0.04	0.13	[-0.29; 0.21]	.741		-0.08	0.14	[-0.35;0.20]	.585	
Fluid Intelligence	-0.18	0.14	[-0.44;0.09]	.199		0.04	0.14	[-0.23; 0.32]	.765		-0.20	0.12	[-0.44;0.04]	.096	
Alternation Bias						0.22	0.19	[-0.16; 0.59]	.126						
R^2	0.39					0.35					0.43				

Note. For the intervention condition one-tailed significance levels are reported because directional hypotheses were tested. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A4

Intervention Effects on Motivational Beliefs Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.58***	0.16	[0.27; 0.89]	<.001	<.001	0.37	0.26	[-0.15;0.89]	.164	.219	0.46**	0.16	[0.15; 0.78]	.004	.008
Data-based Argumentation	-0.10	0.10	[-0.29; 0.09]	.307		-0.04	0.14	[-0.31;0.23]	.767		-0.33	0.13	[-0.59;-0.07]	.012	
Views on Variability	0.10	0.13	[-0.15; 0.34]	.448		0.14	0.09	[-0.03;0.32]	.112		-0.06	0.15	[-0.35; 0.23]	.698	
Draws to Decision	0.07	0.14	[-0.20; 0.34]	.622		0.04	0.12	[-0.20;0.27]	.766		0.06	0.18	[-0.28; 0.41]	.718	
Decision Threshold	-0.11	0.11	[-0.32; 0.10]	.318		0.02	0.14	[-0.25;0.29]	.893		-0.08	0.12	[-0.31; 0.15]	.501	
Confirmation Bias	0.00	0.12	[-0.23; 0.24]	.973		0.01	0.17	[-0.33;0.36]	.934		0.12	0.10	[-0.08; 0.32]	.231	
Intrinsic Value	-0.11	0.15	[-0.41; 0.18]	.446		0.43	0.15	[0.14;0.72]	.002		-0.06	0.15	[-0.35; 0.22]	.661	
Attainment Value	-0.06	0.20	[-0.45; 0.34]	.784		-0.09	0.18	[-0.45;0.27]	.619		0.44	0.16	[0.12; 0.77]	.004	
General Self-efficacy	0.17	0.14	[-0.10; 0.44]	.227		0.18	0.09	[0.01;0.35]	.040		-0.01	0.11	[-0.22; 0.21]	.960	
Gender (1 = male)	-0.42	0.17	[-0.75;-0.10]	.011		-0.45	0.33	[-1.11;0.20]	.173		-0.12	0.32	[-0.76; 0.52]	.714	
Age	-0.02	0.11	[-0.23; 0.19]	.857		-0.15	0.11	[-0.37;0.07]	.171		-0.02	0.13	[-0.28; 0.24]	.881	
Fluid Intelligence	-0.02	0.15	[-0.32; 0.28]	.886		-0.03	0.19	[-0.40;0.33]	.861		0.04	0.15	[-0.26; 0.33]	.804	
Self-concept	0.59	0.14	[0.31; 0.87]	<.001											
R^2	0.51					0.36					0.36				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A4

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Intervention condition	-0.10	0.22	[-0.54;0.33]	.643	.643
Data-based argumentation	-0.26	0.16	[-0.57;0.04]	.093	
Views on Variability	-0.05	0.10	[-0.25;0.14]	.591	
Draws to Decision	-0.05	0.13	[-0.30;0.19]	.675	
Decision Threshold	0.02	0.10	[-0.18;0.21]	.877	
Confirmation Bias	0.06	0.12	[-0.19;0.30]	.640	
Intrinsic Value	0.10	0.18	[-0.27;0.46]	.607	
Attainment Value	-0.07	0.16	[-0.39;0.25]	.660	
General Self-efficacy	0.52	0.10	[0.33;0.71]	<.001	
Gender (1 = male)	0.00	0.23	[-0.44;0.44]	.996	
Age	-0.11	0.11	[-0.32;0.11]	.334	
Fluid Intelligence	0.02	0.12	[-0.22;0.27]	.849	
R^2	0.46				

Note. Two-tailed significance levels are reported. The p-values of the intervention condition predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A5

Differential Intervention Effects of Previous Knowledge on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.51*	0.18	[0.15; 0.86]	.005		-0.07	0.21	[-0.48; 0.33]	.725		0.41	0.24	[-0.06;0.89]	.088	
Pretest Value × Intervention condition	0.46	0.29	[-0.12; 1.04]	.119	.238	-0.19	0.16	[-0.50; 0.12]	.223	.257	-0.05	0.27	[-0.58;0.48]	.858	.858
Data-based Argumentation	0.20	0.19	[-0.17; 0.58]	.294		0.25**	0.08	[0.09; 0.41]	.002		-0.03	0.09	[-0.20;0.14]	.748	
Views on Variability	0.02	0.13	[-0.23; 0.27]	.879		0.49***	0.12	[0.26; 0.72]	<.001		-0.04	0.07	[-0.18;0.09]	.521	
Draws to Decision	0.01	0.14	[-0.27; 0.28]	.966		-0.11	0.13	[-0.37; 0.14]	.378		0.51	0.27	[-0.02;1.04]	.058	
Decision Threshold	-0.04	0.18	[-0.38; 0.31]	.833		0.15	0.17	[-0.18; 0.48]	.370		-0.11	0.15	[-0.41;0.19]	.490	
Confirmation Bias	-0.15*	0.07	[-0.29;-0.01]	.030		0.26***	0.07	[0.12; 0.41]	<.001		0.16	0.10	[-0.02;0.35]	.089	
Intrinsic Value	0.04	0.18	[-0.31; 0.39]	.822		-0.03	0.09	[-0.21; 0.14]	.689		0.18	0.15	[-0.10;0.47]	.213	
Attainment Value	-0.09	0.13	[-0.35; 0.16]	.474		0.05	0.14	[-0.22; 0.33]	.698		-0.01	0.14	[-0.29;0.26]	.921	
General Self-efficacy	-0.12	0.12	[-0.35; 0.12]	.343		-0.21	0.11	[-0.43; 0.02]	.068		-0.06	0.15	[-0.35;0.23]	.705	
Gender (1 = male)	-0.31	0.21	[-0.72; 0.09]	.129		-0.48*	0.20	[-0.86;-0.10]	.014		0.03	0.24	[-0.44;0.50]	.902	
Age	0.08	0.13	[-0.18; 0.34]	.562		-0.18	0.19	[-0.55; 0.19]	.343		-0.16	0.10	[-0.35;0.04]	.113	
Fluid Intelligence	0.18*	0.08	[0.01; 0.34]	.036		0.00	0.25	[-0.49; 0.49]	.988		0.06	0.14	[-0.22;0.34]	.675	
R^2	0.51					0.34					0.30				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A5

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.81***	0.21	[0.39; 1.23]	<.001		-0.51*	0.18	[-0.87;-0.16]	.005		-0.25	0.22	[-0.68;0.18]	.248	
Pretest Value × Intervention condition	-0.32	0.14	[-0.61;-0.04]	.026	.238	-0.33	0.28	[-0.88; 0.21]	.231	.257	-0.49	0.25	[-0.98;0.01]	.054	.238
Data-based Argumentation	-0.02	0.13	[-0.27; 0.24]	.904		-0.08	0.11	[-0.29; 0.13]	.470		-0.17	0.18	[-0.53;0.18]	.335	
Views on Variability	0.11	0.07	[-0.03; 0.25]	.122		0.15	0.10	[-0.04; 0.34]	.111		0.24	0.14	[-0.03;0.52]	.081	
Draws to Decision	0.17	0.11	[-0.05; 0.39]	.136		-0.21	0.14	[-0.48; 0.06]	.133		-0.07	0.12	[-0.31;0.18]	.598	
Decision Threshold	0.45**	0.16	[0.15; 0.76]	.004		0.03	0.09	[-0.14; 0.20]	.740		0.00	0.10	[-0.19;0.20]	.959	
Confirmation Bias	0.14	0.17	[-0.18; 0.46]	.397		0.16	0.11	[-0.06; 0.39]	.155		0.67**	0.21	[0.25;1.09]	.002	
Intrinsic Value	-0.26	0.22	[-0.69; 0.17]	.241		-0.11	0.19	[-0.49; 0.26]	.545		-0.03	0.21	[-0.45;0.38]	.875	
Attainment Value	0.34*	0.14	[0.06; 0.61]	.016		-0.25*	0.13	[-0.50;-0.01]	.042		-0.04	0.17	[-0.37;0.28]	.789	
General Self-efficacy	0.12	0.14	[-0.15; 0.40]	.374		0.11	0.10	[-0.08; 0.30]	.257		0.20	0.11	[-0.02;0.41]	.069	
Gender (1 = male)	-0.19	0.26	[-0.70; 0.33]	.476		0.48**	0.17	[0.15; 0.81]	.004		-0.26	0.32	[-0.89;0.37]	.418	
Age	0.11	0.11	[-0.12; 0.33]	.354		-0.06	0.11	[-0.29; 0.16]	.584		-0.05	0.13	[-0.30;0.20]	.706	
Fluid Intelligence	-0.18	0.13	[-0.44; 0.07]	.164		0.06	0.15	[-0.23; 0.35]	.696		-0.23	0.13	[-0.49;0.02]	.068	
Alternation Bias						0.39*	0.15	[0.09; 0.69]	.011						
R^2	0.43					0.37					0.48				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A6

Differential Intervention Effects of Previous Knowledge on Motivational Beliefs Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.56**	0.16	[0.24; 0.88]	.001		0.40	0.28	[-0.15;0.95]	.152		0.48*	0.19	[0.11; 0.84]	.011	
Pretest Value × Intervention condition	-0.24	0.14	[-0.51; 0.04]	.095	.238	0.44	0.27	[-0.09;0.98]	.105	.238	0.44	0.33	[-0.21; 1.09]	.184	.257
Data-based Argumentation	-0.10	0.10	[-0.29; 0.08]	.286		-0.03	0.13	[-0.28;0.22]	.788		-0.33*	0.13	[-0.59;-0.07]	.013	
Views on Variability	0.05	0.12	[-0.18; 0.28]	.656		0.18	0.09	[0.00;0.35]	.050		-0.05	0.14	[-0.33; 0.23]	.729	
Draws to Decision	0.05	0.13	[-0.20; 0.30]	.710		0.07	0.11	[-0.15;0.28]	.549		0.08	0.18	[-0.27; 0.43]	.651	
Decision Threshold	-0.10	0.11	[-0.32; 0.11]	.343		0.00	0.13	[-0.26;0.26]	.981		-0.09	0.13	[-0.34; 0.16]	.490	
Confirmation Bias	-0.01	0.10	[-0.21; 0.19]	.931		0.02	0.18	[-0.33;0.37]	.904		0.17*	0.09	[0.00; 0.34]	.047	
Intrinsic Value	-0.09	0.14	[-0.37; 0.20]	.554		0.18	0.20	[-0.21;0.58]	.360		-0.05	0.14	[-0.32; 0.21]	.693	
Attainment Value	-0.08	0.21	[-0.49; 0.32]	.689		-0.04	0.18	[-0.38;0.31]	.833		0.28	0.25	[-0.21; 0.76]	.263	
General Self-efficacy	0.19	0.13	[-0.06; 0.44]	.140		0.17*	0.07	[0.02;0.31]	.023		-0.06	0.12	[-0.30; 0.18]	.625	
Gender (1 = male)	-0.40*	0.18	[-0.76;-0.05]	.026		-0.44	0.30	[-1.02;0.15]	.142		-0.05	0.29	[-0.61; 0.51]	.862	
Age	-0.04	0.10	[-0.23; 0.16]	.725		-0.17	0.11	[-0.39;0.04]	.108		-0.03	0.13	[-0.29; 0.23]	.813	
Fluid Intelligence	-0.03	0.15	[-0.33; 0.27]	.832		-0.02	0.19	[-0.39;0.34]	.902		0.07	0.12	[-0.17; 0.31]	.582	
Self-concept	0.68***	0.15	[0.38; 0.97]	<.001											
R^2	0.51					0.39					0.42				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A6

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Intervention Condition	-0.09	0.22	[-0.53;0.35]	.676	
Pretest Value × Intervention Condition	-0.20	0.16	[-0.52;0.11]	.203	.257
Data-based Argumentation	-0.30	0.16	[-0.61;0.01]	.062	
Views on Variability	-0.04	0.10	[-0.24;0.15]	.670	
Draws to Decision	-0.10	0.15	[-0.39;0.19]	.503	
Decision Threshold	0.02	0.10	[-0.17;0.21]	.833	
Confirmation Bias	0.04	0.11	[-0.17;0.25]	.734	
Intrinsic Value	0.10	0.18	[-0.26;0.46]	.580	
Attainment Value	-0.07	0.16	[-0.38;0.25]	.671	
General Self-efficacy	0.63***	0.10	[0.44;0.81]	<.001	
Gender (1 = male)	-0.04	0.22	[-0.47;0.39]	.858	
Age	-0.10	0.11	[-0.32;0.11]	.355	
Fluid Intelligence	-0.01	0.12	[-0.24;0.22]	.957	
R^2	0.47				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A7

Differential Intervention Effects of Age on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.58**	0.19	[0.21; 0.95]	.002		-0.02	0.22	[-0.46; 0.42]	.919		0.42	0.24	[-0.05;0.90]	.082	
Age × Intervention Condition	0.01	0.38	[-0.73; 0.76]	.969	.990	0.07	0.33	[-0.58; 0.71]	.835	.990	0.15	0.25	[-0.34;0.64]	.558	.946
Data-based Argumentation	0.47***	0.12	[0.25; 0.70]	<.001		0.25**	0.09	[0.08; 0.42]	.004		-0.02	0.09	[-0.20;0.16]	.821	
Views on Variability	-0.01	0.13	[-0.26; 0.24]	.923		0.40***	0.10	[0.21; 0.60]	<.001		-0.05	0.06	[-0.17;0.07]	.411	
Draws to Decision	0.00	0.13	[-0.24; 0.25]	.971		-0.10	0.14	[-0.39; 0.18]	.480		0.50*	0.23	[0.04;0.96]	.032	
Decision Threshold	-0.05	0.17	[-0.39; 0.28]	.747		0.13	0.19	[-0.25; 0.51]	.506		-0.13	0.14	[-0.42;0.15]	.360	
Confirmation Bias	-0.16*	0.07	[-0.29; -0.03]	.016		0.21*	0.10	[0.01; 0.40]	.042		0.15	0.11	[-0.06;0.35]	.167	
Intrinsic Value	0.07	0.18	[-0.29; 0.43]	.711		-0.01	0.08	[-0.16; 0.14]	.909		0.19	0.16	[-0.13;0.50]	.244	
Attainment Value	-0.13	0.12	[-0.36; 0.10]	.261		0.05	0.15	[-0.23; 0.34]	.711		-0.02	0.15	[-0.31;0.27]	.892	
General Self-efficacy	-0.20	0.12	[-0.43; 0.04]	.106		-0.19	0.11	[-0.40; 0.02]	.081		-0.06	0.11	[-0.27;0.16]	.607	
Gender (1 = male)	-0.25	0.17	[-0.59; 0.10]	.158		-0.45*	0.20	[-0.84;-0.06]	.025		0.06	0.24	[-0.41;0.52]	.811	
Age	0.05	0.22	[-0.38; 0.49]	.806		-0.20	0.25	[-0.68; 0.28]	.421		-0.22	0.12	[-0.44;0.01]	.061	
Fluid Intelligence	0.19	0.10	[0.00; 0.38]	.051		-0.03	0.24	[-0.50; 0.45]	.912		0.06	0.14	[-0.22;0.34]	.668	
R^2	0.50					0.32					0.30				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A7

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.79***	0.21	[0.37;1.20]	<.001		-0.51**	0.19	[-0.88; -0.14]	.007		-0.27	0.26	[-0.78;0.24]	.299	
Age × Intervention condition	-0.36	0.27	[-0.90;0.17]	.182	.566	0.08	0.22	[-0.36; 0.52]	.716	.946	-0.05	0.25	[-0.55;0.44]	.842	.990
Data-based Argumentation	-0.05	0.13	[-0.31;0.20]	.690		-0.08	0.12	[-0.31; 0.15]	.489		-0.10	0.17	[-0.44;0.23]	.545	
Views on Variability	0.11*	0.05	[0.01;0.22]	.036		0.13	0.11	[-0.08; 0.35]	.223		0.22	0.13	[-0.03;0.47]	.090	
Draws to Decision	0.10	0.13	[-0.15;0.34]	.451		-0.20	0.14	[-0.48; 0.08]	.156		-0.08	0.12	[-0.30;0.15]	.503	
Decision Threshold	0.33*	0.13	[0.07;0.59]	.011		-0.03	0.09	[-0.21; 0.15]	.768		0.01	0.11	[-0.22;0.23]	.938	
Confirmation Bias	0.09	0.19	[-0.28;0.45]	.638		0.10	0.11	[-0.11; 0.32]	.330		0.45*	0.20	[0.07;0.84]	.021	
Intrinsic Value	-0.24	0.20	[-0.64;0.16]	.232		-0.09	0.20	[-0.49; 0.31]	.644		-0.07	0.20	[-0.46;0.32]	.730	
Attainment Value	0.33*	0.13	[0.08;0.58]	.011		-0.25*	0.12	[-0.48;-0.02]	.030		-0.01	0.17	[-0.34;0.32]	.953	
General Self-efficacy	0.21	0.16	[-0.11;0.52]	.194		0.06	0.10	[-0.13; 0.26]	.511		0.16	0.08	[0.00;0.33]	.056	
Gender (1 = male)	-0.25	0.23	[-0.70;0.19]	.264		0.51**	0.16	[0.20; 0.82]	.001		-0.23	0.34	[-0.89;0.43]	.488	
Age	0.20	0.13	[-0.06;0.46]	.125		-0.07	0.14	[-0.35; 0.21]	.627		-0.06	0.19	[-0.44;0.31]	.746	
Fluid Intelligence	-0.16	0.12	[-0.40;0.07]	.176		0.04	0.14	[-0.23; 0.31]	.782		-0.20	0.12	[-0.42;0.03]	.094	
Alternation Bias						0.23	0.18	[-0.12; 0.58]	.195						
R^2	0.43					0.35					0.43				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A8

Differential Intervention Effects of Age on Motivational Beliefs Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.60***	0.16	[0.29; 0.91]	<.001		0.39	0.26	[-0.11;0.89]	.128		0.47**	0.15	[0.17; 0.77]	.002	
Age × Intervention Condition	0.44	0.17	[0.11; 0.77]	.010	.229	0.30	0.26	[-0.22;0.81]	.260	.693	0.07	0.18	[-0.29; 0.43]	.693	.946
Data-based Argumentation	-0.07	0.10	[-0.27; 0.12]	.475		-0.02	0.16	[-0.33;0.28]	.879		-0.33*	0.13	[-0.59; -0.06]	.015	
Views on Variability	0.08	0.12	[-0.16; 0.32]	.509		0.14	0.09	[-0.03;0.31]	.110		-0.06	0.15	[-0.35; 0.23]	.680	
Draws to Decision	0.14	0.13	[-0.11; 0.39]	.268		0.06	0.10	[-0.14;0.26]	.546		0.08	0.17	[-0.26; 0.42]	.651	
Decision Threshold	-0.18	0.09	[-0.36; 0.00]	.053		-0.03	0.12	[-0.27;0.22]	.821		-0.09	0.11	[-0.32; 0.13]	.410	
Confirmation Bias	0.00	0.14	[-0.27; 0.28]	.981		0.00	0.19	[-0.37;0.37]	.987		0.12	0.10	[-0.07; 0.32]	.224	
Intrinsic Value	-0.19	0.16	[-0.51; 0.13]	.247		0.43**	0.15	[0.14;0.71]	.004		-0.06	0.15	[-0.35; 0.22]	.660	
Attainment Value	-0.06	0.20	[-0.44; 0.32]	.754		-0.09	0.18	[-0.44;0.25]	.597		0.45**	0.17	[0.12; 0.77]	.007	
General Self-efficacy	0.12	0.14	[-0.15; 0.39]	.393		0.16*	0.08	[0.00;0.32]	.045		-0.01	0.11	[-0.23; 0.20]	.899	
Gender (1 = male)	-0.38*	0.17	[-0.72;-0.04]	.029		-0.39	0.31	[-1.01;0.22]	.212		-0.10	0.32	[-0.72; 0.52]	.751	
Age	-0.15	0.11	[-0.37; 0.06]	.167		-0.25	0.14	[-0.53;0.03]	.083		-0.04	0.13	[-0.30; 0.22]	.755	
Fluid Intelligence	-0.07	0.16	[-0.39; 0.25]	.674		-0.05	0.20	[-0.44;0.34]	.804		0.03	0.15	[-0.26; 0.33]	.816	
Self-concept	0.69***	0.12	[0.45; 0.92]	<.001											
R^2	0.56					0.36					0.36				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A8

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Intervention Condition	-0.09	0.22	[-0.51;0.33]	.674	
Age × Intervention Condition	0.18	0.14	[-0.09;0.46]	.193	.566
Data-based argumentation	-0.25	0.16	[-0.57;0.07]	.131	
Views on Variability	-0.05	0.10	[-0.25;0.14]	.592	
Draws to Decision	-0.03	0.12	[-0.27;0.20]	.773	
Decision Threshold	-0.01	0.10	[-0.21;0.18]	.886	
Confirmation Bias	0.06	0.12	[-0.18;0.29]	.647	
Intrinsic Value	0.10	0.19	[-0.27;0.47]	.608	
Attainment Value	-0.08	0.16	[-0.40;0.24]	.631	
General Self-efficacy	0.51***	0.10	[0.32;0.70]	<.001	
Gender (1 = male)	0.04	0.21	[-0.37;0.45]	.854	
Age	-0.17	0.13	[-0.43;0.09]	.194	
Fluid Intelligence	0.01	0.12	[-0.22;0.25]	.908	
R^2	0.46				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A9

Differential Intervention Effects of Grade Level on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.33	0.71	[-1.06;1.72]	.641		-0.51	0.57	[-1.63; 0.61]	.371		-0.80	0.62	[-2.01;0.41]	.195	
Grade Level × Intervention Condition	0.17	0.55	[-0.91;1.25]	.757	.946	0.35	0.53	[-0.68; 1.38]	.509	.946	0.82	0.40	[0.04;1.60]	.040	.229
Data-based Argumentation	0.48***	0.11	[0.26;0.70]	<.001		0.25*	0.10	[0.05; 0.45]	.015		0.03	0.12	[-0.20;0.26]	.795	
Views on Variability	-0.02	0.11	[-0.24;0.20]	.849		0.39**	0.12	[0.15; 0.63]	.001		-0.02	0.06	[-0.13;0.09]	.722	
Draws to Decision	0.02	0.15	[-0.26;0.31]	.881		-0.11	0.13	[-0.38; 0.15]	.410		0.45	0.25	[-0.03;0.93]	.066	
Decision Threshold	-0.06	0.17	[-0.40;0.28]	.711		0.13	0.19	[-0.25; 0.51]	.499		-0.11	0.14	[-0.39;0.16]	.409	
Confirmation Bias	-0.15	0.08	[-0.31;0.02]	.078		0.21*	0.10	[0.01; 0.41]	.039		0.21*	0.09	[0.03;0.39]	.021	
Intrinsic Value	0.04	0.22	[-0.38;0.47]	.841		-0.03	0.06	[-0.14; 0.08]	.574		0.16	0.18	[-0.20;0.52]	.383	
Attainment Value	-0.10	0.13	[-0.36;0.15]	.434		0.08	0.15	[-0.22; 0.37]	.610		0.02	0.14	[-0.26;0.30]	.888	
General Self-efficacy	-0.20	0.11	[-0.42;0.02]	.072		-0.19	0.11	[-0.39; 0.02]	.072		-0.04	0.10	[-0.24;0.16]	.705	
Grade Level (1 = fourth)	-0.14	0.45	[-1.03;0.74]	.748		-0.13	0.52	[-1.15; 0.89]	.802		-0.69	0.45	[-1.58;0.20]	.130	
Gender (1 = male)	-0.22	0.20	[-0.60;0.16]	.254		-0.43*	0.17	[-0.77;-0.10]	.011		0.10	0.19	[-0.27;0.48]	.588	
Age	0.12	0.15	[-0.19;0.42]	.450		-0.19	0.23	[-0.65; 0.28]	.430		-0.06	0.13	[-0.32;0.20]	.664	
Fluid Intelligence	0.17	0.09	[-0.01;0.35]	.060		-0.04	0.25	[-0.53; 0.45]	.870		0.08	0.13	[-0.17;0.33]	.521	
R^2	0.51					0.33					0.33				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A9

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.96	0.50	[-0.03;1.94]	.057		0.69	0.59	[-0.46; 1.84]	.241		-1.67*	0.74	[-3.11;-0.22]	.024	
Grade Level × Intervention Condition	-0.18	0.37	[-0.90;0.55]	.634	.946	-0.88	0.35	[-1.57;-0.19]	.012	.229	1.00	0.46	[0.11; 1.90]	.027	.229
Data-based Argumentation	0.02	0.16	[-0.28;0.33]	.886		-0.07	0.12	[-0.31; 0.16]	.537		-0.09	0.19	[-0.45; 0.28]	.639	
Views on Variability	0.20***	0.04	[0.11;0.29]	<.001		0.20	0.12	[-0.04; 0.44]	.106		0.16	0.11	[-0.06; 0.37]	.154	
Draws to Decision	0.10	0.11	[-0.11;0.30]	.363		-0.25*	0.13	[-0.50; 0.00]	.049		-0.03	0.13	[-0.28; 0.21]	.798	
Decision Threshold	0.30*	0.12	[0.07;0.53]	.011		0.05	0.04	[-0.02; 0.12]	.187		-0.05	0.10	[-0.24; 0.14]	.592	
Confirmation Bias	0.10	0.17	[-0.23;0.43]	.548		0.09	0.10	[-0.10; 0.29]	.351		0.50**	0.16	[0.19; 0.82]	.002	
Intrinsic Value	-0.15	0.23	[-0.59;0.30]	.511		0.06	0.22	[-0.37; 0.49]	.786		-0.21	0.19	[-0.59; 0.16]	.260	
Attainment Value	0.25	0.15	[-0.05;0.55]	.103		-0.34**	0.13	[-0.59;-0.09]	.008		0.09	0.14	[-0.19; 0.37]	.516	
General Self-efficacy	0.17	0.12	[-0.06;0.41]	.148		0.04	0.12	[-0.21; 0.28]	.776		0.20*	0.08	[0.04; 0.36]	.016	
Grade Level (1 = fourth)	-0.56	0.39	[-1.33;0.20]	.149		-0.20	0.25	[-0.69; 0.30]	.436		-0.16	0.44	[-1.03; 0.71]	.720	
Gender (1 = male)	-0.25	0.24	[-0.73;0.23]	.308		0.38*	0.18	[0.02; 0.73]	.038		-0.08	0.32	[-0.70; 0.54]	.800	
Age	0.20	0.14	[-0.09;0.48]	.173		0.07	0.13	[-0.19; 0.33]	.585		-0.09	0.18	[-0.44; 0.26]	.617	
Fluid Intelligence	-0.11	0.11	[-0.32;0.11]	.319		0.13	0.14	[-0.14; 0.40]	.352		-0.25*	0.12	[-0.49;-0.01]	.045	
Alternation Bias						0.21	0.16	[-0.10; 0.52]	.181						
R^2	0.44					0.41					0.49				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A10

Differential Intervention Effects of Grade Level on Motivational Outcomes Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.32	0.93	[-1.50; 2.15]	.728		-0.68	0.79	[-2.22;0.87]	.392		0.67	0.68	[-0.67; 2.01]	.327	
Grade Level × Intervention Condition	0.19	0.61	[-1.00; 1.39]	.751	.946	0.75	0.48	[-0.20;1.70]	.121	.484	-0.07	0.52	[-1.08; 0.95]	.898	.990
Data-based Argumentation	-0.10	0.12	[-0.33; 0.12]	.375		-0.03	0.18	[-0.37;0.32]	.884		-0.39*	0.15	[-0.69;-0.09]	.011	
Views on Variability	0.08	0.13	[-0.18; 0.34]	.544		0.10	0.11	[-0.12;0.33]	.355		-0.16	0.14	[-0.44; 0.11]	.247	
Draws to Decision	0.07	0.15	[-0.22; 0.36]	.633		0.05	0.15	[-0.26;0.35]	.761		0.12	0.19	[-0.25; 0.50]	.520	
Decision Threshold	-0.12	0.12	[-0.35; 0.11]	.293		-0.01	0.16	[-0.32;0.30]	.939		-0.11	0.11	[-0.33; 0.12]	.344	
Confirmation Bias	-0.01	0.11	[-0.23; 0.22]	.965		0.04	0.18	[-0.32;0.39]	.843		0.09	0.11	[-0.13; 0.30]	.444	
Intrinsic Value	-0.13	0.15	[-0.42; 0.17]	.392		0.34*	0.16	[0.02;0.66]	.038		-0.16	0.14	[-0.43; 0.11]	.242	
Attainment Value	-0.04	0.20	[-0.44; 0.36]	.843		-0.03	0.16	[-0.35;0.29]	.852		0.51**	0.16	[0.20; 0.82]	.001	
General Self-efficacy	0.18	0.15	[-0.12; 0.48]	.246		0.20*	0.10	[0.01;0.40]	.042		0.01	0.09	[-0.16; 0.18]	.912	
Grade Level (1 = fourth)	0.04	0.52	[-0.99; 1.07]	.939		-0.12	0.54	[-1.17;0.93]	.816		0.81*	0.31	[0.20; 1.43]	.010	
Gender (1 = male)	-0.40**	0.13	[-0.66;-0.13]	.003		-0.35	0.30	[-0.95;0.24]	.239		-0.05	0.31	[-0.66; 0.56]	.871	
Age	-0.05	0.16	[-0.37; 0.27]	.756		-0.16	0.12	[-0.40;0.08]	.188		-0.18	0.17	[-0.52; 0.16]	.296	
Fluid Intelligence	-0.03	0.15	[-0.33; 0.26]	.822		-0.08	0.19	[-0.46;0.30]	.681		-0.03	0.16	[-0.34; 0.27]	.831	
Self-concept	0.57***	0.16	[0.26; 0.88]	<.001											
R^2	0.51					0.37					0.43				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A10

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Intervention Condition	-0.62	0.41	[-1.42; 0.17]	.126	
Grade Level × Intervention Condition	0.45	0.22	[0.02; 0.88]	.039	.229
Data-based argumentation	-0.32	0.17	[-0.65; 0.01]	.061	
Views on Variability	-0.18	0.11	[-0.38; 0.03]	.097	
Draws to Decision	0.01	0.13	[-0.25; 0.27]	.943	
Decision Threshold	-0.04	0.08	[-0.19; 0.11]	.598	
Confirmation Bias	0.05	0.14	[-0.22; 0.32]	.723	
Intrinsic Value	-0.08	0.21	[-0.49; 0.32]	.682	
Attainment Value	0.05	0.15	[-0.24; 0.34]	.746	
General Self-efficacy	0.54***	0.08	[0.37; 0.70]	<.001	
Grade Level (1 = fourth)	0.73*	0.29	[0.16; 1.31]	.012	
Gender (1 = male)	0.12	0.17	[-0.21; 0.46]	.473	
Age	-0.31*	0.12	[-0.55;-0.07]	.011	
Fluid Intelligence	-0.08	0.09	[-0.25; 0.09]	.369	
R^2	0.55				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A11

Differential Intervention Effects of Gender on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.15	0.22	[-0.29; 0.58]	.511		-0.46	0.34	[-1.13; 0.22]	.185		0.94**	0.36	[0.23; 1.64]	.009	
Gender × Intervention Condition	0.73	0.34	[0.07; 1.39]	.030	.229	0.74	0.33	[0.09; 1.39]	.025	.229	-0.91	0.46	[-1.82;-0.01]	.047	.235
Data-based Argumentation	0.43***	0.11	[0.22; 0.65]	<.001		0.21**	0.08	[0.05; 0.37]	.009		0.02	0.10	[-0.17; 0.22]	.825	
Views on Variability	0.04	0.13	[-0.22; 0.30]	.776		0.44***	0.11	[0.22; 0.66]	<.001		-0.10	0.07	[-0.25; 0.04]	.168	
Draws to Decision	0.03	0.14	[-0.25; 0.31]	.833		-0.09	0.16	[-0.41; 0.22]	.558		0.46*	0.23	[0.01; 0.92]	.045	
Decision Threshold	-0.04	0.17	[-0.39; 0.30]	.807		0.15	0.19	[-0.22; 0.52]	.418		-0.12	0.13	[-0.38; 0.13]	.344	
Confirmation Bias	-0.14*	0.06	[-0.26;-0.01]	.035		0.23**	0.08	[0.06; 0.40]	.007		0.16	0.13	[-0.10; 0.42]	.235	
Intrinsic Value	0.07	0.20	[-0.31; 0.45]	.720		0.00	0.08	[-0.16; 0.15]	.966		0.18	0.16	[-0.13; 0.49]	.259	
Attainment Value	-0.12	0.14	[-0.39; 0.15]	.384		0.05	0.16	[-0.27; 0.37]	.779		-0.04	0.14	[-0.31; 0.24]	.799	
General Self-efficacy	-0.20	0.12	[-0.43; 0.04]	.098		-0.19*	0.09	[-0.37;-0.01]	.043		-0.07	0.11	[-0.29; 0.16]	.568	
Gender (1 = male)	-0.65***	0.17	[-0.99;-0.31]	<.001		-0.88***	0.24	[-1.35;-0.41]	<.001		0.55	0.39	[-0.21; 1.30]	.156	
Age	0.12	0.12	[-0.12; 0.36]	.341		-0.13	0.19	[-0.50; 0.23]	.474		-0.20**	0.06	[-0.32;-0.08]	.001	
Fluid Intelligence	0.18*	0.09	[0.01; 0.36]	.042		-0.03	0.23	[-0.49; 0.43]	.896		0.07	0.13	[-0.19; 0.33]	.594	
R^2	0.54					0.36					0.33				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A11

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.99*	0.47	[0.07;1.91]	.034		-0.52	0.33	[-1.16; 0.13]	.117		0.03	0.47	[-0.88;0.95]	.944	
Gender × Intervention Condition	-0.30	0.57	[-1.42;0.81]	.596	.946	-0.01	0.40	[-0.80; 0.79]	.990	.990	-0.52	0.50	[-1.49;0.45]	.298	.745
Data-based Argumentation	0.00	0.15	[-0.29;0.29]	.977		-0.09	0.12	[-0.33; 0.14]	.444		-0.07	0.17	[-0.39;0.26]	.687	
Views on Variability	0.09	0.08	[-0.07;0.26]	.268		0.13	0.12	[-0.09; 0.36]	.244		0.19	0.15	[-0.10;0.49]	.196	
Draws to Decision	0.14	0.11	[-0.07;0.35]	.181		-0.21	0.13	[-0.47; 0.04]	.100		-0.08	0.14	[-0.35;0.19]	.554	
Decision Threshold	0.26*	0.12	[0.02;0.49]	.033		-0.01	0.07	[-0.15; 0.13]	.911		-0.01	0.11	[-0.22;0.20]	.941	
Confirmation Bias	0.08	0.17	[-0.25;0.40]	.648		0.12	0.11	[-0.11; 0.34]	.311		0.45*	0.18	[0.08;0.81]	.016	
Intrinsic Value	-0.24	0.21	[-0.66;0.17]	.247		-0.09	0.20	[-0.49; 0.30]	.634		-0.08	0.19	[-0.46;0.29]	.654	
Attainment Value	0.32*	0.12	[0.08;0.56]	.010		-0.25*	0.11	[-0.47;-0.03]	.029		-0.01	0.17	[-0.34;0.31]	.928	
General Self-efficacy	0.18	0.17	[-0.16;0.53]	.294		0.06	0.10	[-0.13; 0.26]	.529		0.15	0.09	[-0.01;0.32]	.071	
Gender (1 = male)	0.00	0.50	[-0.99;0.99]	1.000		0.50	0.36	[-0.20; 1.20]	.165		0.06	0.53	[-0.98;1.10]	.912	
Age	0.06	0.08	[-0.11;0.22]	.494		-0.04	0.13	[-0.29; 0.20]	.725		-0.09	0.13	[-0.34;0.15]	.456	
Fluid Intelligence	-0.17	0.14	[-0.46;0.11]	.225		0.04	0.14	[-0.24; 0.31]	.781		-0.20	0.12	[-0.43;0.02]	.081	
Alternation Bias						0.22	0.20	[-0.17; 0.60]	.273						
R^2	0.39					0.35					0.44				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A12

Differential Intervention Effects of Gender on Motivational Outcomes Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.65**	0.20	[0.25; 1.05]	.001		0.50	0.34	[-0.17;1.17]	.142		0.68*	0.26	[0.16; 1.20]	.010	
Gender × Intervention Condition	-0.11	0.31	[-0.72; 0.49]	.714	.946	-0.23	0.42	[-1.06;0.60]	.592	.946	-0.36	0.41	[-1.16; 0.44]	.380	.894
Data-based Argumentation	-0.10	0.09	[-0.28; 0.09]	.290		-0.03	0.14	[-0.30;0.24]	.807		-0.32**	0.12	[-0.55;-0.09]	.006	
Views on Variability	0.08	0.13	[-0.18; 0.34]	.550		0.12	0.09	[-0.05;0.30]	.174		-0.10	0.16	[-0.42; 0.22]	.543	
Draws to Decision	0.07	0.14	[-0.20; 0.34]	.619		0.03	0.12	[-0.20;0.26]	.783		0.06	0.17	[-0.28; 0.41]	.710	
Decision Threshold	-0.11	0.11	[-0.32; 0.10]	.296		0.02	0.13	[-0.24;0.28]	.910		-0.08	0.11	[-0.31; 0.14]	.454	
Confirmation Bias	0.00	0.12	[-0.24; 0.24]	.997		0.01	0.18	[-0.34;0.36]	.953		0.12	0.11	[-0.10; 0.33]	.294	
Intrinsic Value	-0.12	0.16	[-0.42; 0.19]	.461		0.43**	0.15	[0.14;0.71]	.003		-0.08	0.14	[-0.35; 0.20]	.582	
Attainment Value	-0.05	0.21	[-0.46; 0.35]	.795		-0.09	0.19	[-0.46;0.28]	.619		0.46**	0.17	[0.12; 0.79]	.007	
General Self-efficacy	0.17	0.14	[-0.10; 0.44]	.225		0.18*	0.09	[0.00;0.35]	.048		-0.01	0.11	[-0.22; 0.20]	.917	
Gender (1 = male)	-0.36*	0.14	[-0.63;-0.08]	.011		-0.33	0.26	[-0.85;0.18]	.204		0.09	0.34	[-0.58; 0.76]	.785	
Age	-0.03	0.12	[-0.26; 0.21]	.822		-0.16	0.12	[-0.39;0.07]	.166		-0.03	0.13	[-0.29; 0.24]	.841	
Fluid Intelligence	-0.02	0.15	[-0.31; 0.27]	.895		-0.03	0.18	[-0.38;0.32]	.852		0.04	0.14	[-0.25; 0.32]	.794	
Self-concept	0.58***	0.15	[0.29; 0.88]	<.001											
R^2	0.51					0.35					0.38				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A12

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Intervention Condition	-0.10	0.33	[-0.74;0.55]	.772	
Gender × Intervention Condition	-0.01	0.37	[-0.74;0.72]	.977	.990
Data-based argumentation	-0.26	0.15	[-0.55;0.02]	.071	
Views on Variability	-0.05	0.11	[-0.27;0.17]	.628	
Draws to Decision	-0.05	0.12	[-0.30;0.19]	.665	
Decision Threshold	0.02	0.10	[-0.18;0.22]	.867	
Confirmation Bias	0.06	0.12	[-0.19;0.30]	.634	
Intrinsic Value	0.09	0.18	[-0.26;0.45]	.610	
Attainment Value	-0.07	0.16	[-0.39;0.25]	.680	
General Self-efficacy	0.52***	0.10	[0.33;0.71]	<.001	
Gender (1 = male)	0.00	0.32	[-0.63;0.64]	.990	
Age	-0.11	0.11	[-0.33;0.11]	.340	
Fluid Intelligence	0.02	0.13	[-0.22;0.27]	.845	
R^2	0.46				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A13

Differential Intervention Effects of Fluid Intelligence on Statistical Literacy Outcomes Including Baseline Differences

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.51***	0.13	[0.25; 0.76]	<.001		-0.01	0.17	[-0.35; 0.33]	.951		0.41	0.23	[-0.04;0.85]	.072	
Fluid Intelligence × Intervention Condition	0.44	0.26	[-0.07; 0.96]	.093	.413	-0.10	0.21	[-0.50; 0.30]	.625	.946	0.01	0.27	[-0.52;0.55]	.964	.990
Data-based Argumentation	0.45***	0.12	[0.22; 0.68]	<.001		0.26**	0.09	[0.09; 0.43]	.003		-0.03	0.08	[-0.19;0.14]	.759	
Views on Variability	-0.04	0.12	[-0.29; 0.20]	.728		0.41***	0.11	[0.20; 0.62]	<.001		-0.05	0.07	[-0.18;0.08]	.412	
Draws to Decision	0.09	0.13	[-0.17; 0.34]	.497		-0.12	0.13	[-0.37; 0.13]	.346		0.48*	0.24	[0.01;0.96]	.045	
Decision Threshold	-0.07	0.16	[-0.40; 0.25]	.651		0.15	0.18	[-0.22; 0.51]	.431		-0.11	0.15	[-0.39;0.18]	.467	
Confirmation Bias	-0.12	0.08	[-0.28; 0.03]	.109		0.24**	0.09	[0.07; 0.41]	.006		0.16	0.10	[-0.04;0.37]	.111	
Intrinsic Value	0.03	0.14	[-0.25; 0.31]	.830		-0.02	0.09	[-0.19; 0.15]	.822		0.18	0.15	[-0.11;0.48]	.219	
Attainment Value	-0.08	0.11	[-0.29; 0.12]	.438		0.06	0.15	[-0.23; 0.35]	.689		-0.02	0.15	[-0.31;0.27]	.887	
General Self-efficacy	-0.16*	0.07	[-0.31;-0.01]	.032		-0.20	0.10	[-0.41; 0.00]	.053		-0.05	0.11	[-0.26;0.16]	.650	
Gender (1 = male)	-0.16	0.20	[-0.54; 0.22]	.409		-0.45*	0.21	[-0.86;-0.04]	.030		0.04	0.24	[-0.43;0.51]	.875	
Age	0.05	0.14	[-0.23; 0.33]	.723		-0.16	0.19	[-0.53; 0.21]	.391		-0.16	0.10	[-0.35;0.04]	.109	
Fluid Intelligence	0.02	0.13	[-0.24; 0.27]	.893		0.02	0.29	[-0.54; 0.58]	.940		0.06	0.15	[-0.24;0.36]	.695	
R^2	0.52					0.33					0.30				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.

* p < .05. ** p < .01. *** p < .001.

Table A13

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.87***	0.24	[0.39;1.35]	<.001		-0.49*	0.20	[-0.88;-0.10]	.015		-0.25	0.29	[-0.82;0.32]	.391	
Fluid Intelligence × Intervention Condition	-0.37	0.26	[-0.88;0.14]	.153	.556	-0.17	0.24	[-0.64; 0.31]	.486	.946	-0.09	0.20	[-0.48;0.31]	.665	.946
Data-based Argumentation	0.00	0.13	[-0.26;0.25]	.977		-0.08	0.12	[-0.31; 0.16]	.520		-0.09	0.17	[-0.43;0.25]	.607	
Views on Variability	0.13*	0.06	[0.01;0.24]	.035		0.16	0.11	[-0.05; 0.37]	.144		0.22	0.13	[-0.02;0.47]	.075	
Draws to Decision	0.09	0.12	[-0.15;0.33]	.459		-0.24	0.15	[-0.53; 0.05]	.103		-0.09	0.13	[-0.34;0.17]	.499	
Decision Threshold	0.28*	0.12	[0.04;0.51]	.021		0.00	0.07	[-0.14; 0.14]	.964		0.01	0.11	[-0.21;0.22]	.964	
Confirmation Bias	0.06	0.18	[-0.29;0.41]	.721		0.12	0.11	[-0.11; 0.34]	.312		0.45*	0.20	[0.06;0.84]	.023	
Intrinsic Value	-0.22	0.20	[-0.62;0.18]	.274		-0.09	0.21	[-0.51; 0.33]	.686		-0.07	0.20	[-0.46;0.31]	.717	
Attainment Value	0.29*	0.12	[0.05;0.54]	.018		-0.26*	0.12	[-0.50;-0.02]	.035		-0.01	0.15	[-0.30;0.27]	.922	
General Self-efficacy	0.15	0.16	[-0.16;0.45]	.340		0.05	0.11	[-0.16; 0.26]	.626		0.15	0.08	[-0.01;0.32]	.068	
Gender (1 = male)	-0.22	0.25	[-0.71;0.27]	.375		0.47**	0.18	[0.12; 0.82]	.008		-0.23	0.36	[-0.93;0.47]	.515	
Age	0.09	0.12	[-0.15;0.33]	.444		-0.04	0.13	[-0.29; 0.21]	.753		-0.07	0.15	[-0.35;0.22]	.638	
Fluid Intelligence	-0.03	0.17	[-0.36;0.29]	.848		0.11	0.16	[-0.20; 0.42]	.501		-0.16*	0.08	[-0.32;0.00]	.045	
Alternation Bias						0.23	0.19	[-0.13; 0.60]	.212						
R^2	0.43					0.34					0.42				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A14

Differential Intervention Effects of Fluid Intelligence on Motivational Beliefs Including Baseline Differences

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	0.61**	0.19	[0.23; 0.98]	.001		0.37	0.28	[-0.19;0.93]	.194		0.50**	0.15	[0.20; 0.80]	.001	
Fluid Intelligence × Intervention Condition	-0.12	0.28	[-0.68; 0.44]	.669	.946	0.01	0.21	[-0.39;0.41]	.963	.990	-0.22	0.17	[-0.56; 0.12]	.198	.566
Data-based Argumentation	-0.09	0.09	[-0.26; 0.09]	.328		-0.04	0.14	[-0.32;0.24]	.789		-0.33*	0.13	[-0.58;-0.07]	.013	
Views on Variability	0.12	0.13	[-0.13; 0.37]	.347		0.14	0.08	[-0.01;0.30]	.070		-0.05	0.15	[-0.33; 0.24]	.750	
Draws to Decision	0.04	0.11	[-0.18; 0.25]	.730		0.04	0.11	[-0.17;0.25]	.732		0.03	0.17	[-0.31; 0.36]	.865	
Decision Threshold	-0.10	0.11	[-0.32; 0.11]	.346		0.02	0.14	[-0.25;0.29]	.904		-0.07	0.12	[-0.31; 0.16]	.552	
Confirmation Bias	0.00	0.10	[-0.19; 0.19]	.975		0.02	0.17	[-0.32;0.35]	.926		0.11	0.10	[-0.09; 0.31]	.275	
Intrinsic Value	-0.08	0.17	[-0.42; 0.26]	.635		0.43**	0.15	[0.13;0.73]	.005		-0.05	0.14	[-0.33; 0.23]	.744	
Attainment Value	-0.07	0.20	[-0.46; 0.33]	.733		-0.09	0.18	[-0.44;0.27]	.626		0.42**	0.15	[0.12; 0.72]	.006	
General Self-efficacy	0.17	0.14	[-0.11; 0.44]	.228		0.18*	0.09	[0.01;0.35]	.038		-0.03	0.11	[-0.24; 0.18]	.774	
Gender (1 = male)	-0.43**	0.16	[-0.74;-0.12]	.007		-0.45	0.32	[-1.07;0.17]	.158		-0.15	0.29	[-0.72; 0.42]	.602	
Age	-0.02	0.10	[-0.22; 0.18]	.842		-0.15	0.11	[-0.37;0.07]	.186		-0.01	0.13	[-0.27; 0.25]	.935	
Fluid Intelligence	0.03	0.26	[-0.47; 0.54]	.900		-0.04	0.25	[-0.52;0.45]	.886		0.12	0.20	[-0.26; 0.51]	.532	
Self-concept	0.56***	0.12	[0.33; 0.79]	<.001											
R^2	0.51					0.36					0.38				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A14

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Intervention Condition	-0.12	0.21	[-0.54;0.30]	.573	
Fluid Intelligence × Intervention Condition	0.07	0.17	[-0.27;0.40]	.688	.946
Data-based argumentation	-0.26	0.16	[-0.57;0.05]	.101	
Views on Variability	-0.06	0.11	[-0.27;0.15]	.580	
Draws to Decision	-0.05	0.13	[-0.31;0.22]	.726	
Decision Threshold	0.01	0.10	[-0.19;0.21]	.897	
Confirmation Bias	0.06	0.12	[-0.18;0.30]	.615	
Intrinsic Value	0.09	0.18	[-0.27;0.44]	.634	
Attainment Value	-0.06	0.16	[-0.37;0.25]	.691	
General Self-efficacy	0.53***	0.09	[0.35;0.71]	<.001	
Gender (1 = male)	0.01	0.23	[-0.45;0.46]	.973	
Age	-0.11	0.11	[-0.31;0.10]	.319	
Fluid Intelligence	0.00	0.13	[-0.26;0.25]	.974	
R^2	0.46				

Note. Two-tailed significance levels are reported. The p-values of the interaction predictor were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A15

Effects of Adherence and Quality of Delivery on Statistical Literacy Outcomes

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Adherence	-0.44	0.16	[-0.76;-0.12]	.007	.070	-0.03	0.26	[-0.54;0.48]	.907	.981	0.49*	0.15	[0.19;0.79]	.001	.020
Pretest Value	0.66***	0.18	[0.30; 1.02]	<.001		0.33	0.17	[-0.01;0.66]	.058		0.34	0.23	[-0.13;0.80]	.154	
R^2	0.59					0.12					0.30				
Model 2															
Quality of Delivery	0.12	0.17	[-0.21;0.45]	.462	.924	-0.21	0.16	[-0.52;0.10]	.188	.800	0.08	0.27	[-0.45;0.61]	.769	.961
Pretest Value	0.67***	0.17	[0.34;1.01]	<.001		0.39**	0.14	[0.12;0.66]	.004		0.36	0.27	[-0.18;0.89]	.191	
R^2	0.43					0.18					0.13				

Note. Two-tailed significance levels are reported. The p-values of the adherence and quality of delivery predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A15

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Adherence	0.17	0.14	[-0.12;0.45]	.250	.800	0.02	0.19	[-0.36;0.40]	.906	.981	-0.03	0.08	[-0.18;0.12]	.702	.961
Pretest Value	0.21*	0.10	[0.01;0.41]	.038		0.07	0.21	[-0.35;0.49]	.757		0.39*	0.16	[0.08;0.70]	.013	
R^2	0.08					0.00					0.16				
Model 2															
Quality of Delivery	0.11	0.17	[-0.23;0.46]	.515	.936	-0.08	0.23	[-0.54;0.38]	.733	.961	-0.08	0.17	[-0.40;0.25]	.642	.961
Pretest Value	0.09	0.14	[-0.19;0.37]	.532		0.07	0.21	[-0.35;0.49]	.741		0.39*	0.15	[0.09;0.68]	.010	
R^2	0.04					0.01					0.16				

Note. Two-tailed significance levels are reported. The p-values of the adherence and quality of delivery predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A16

Effects of Adherence and Quality of Delivery on Motivation for Data-related Tasks

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Adherence	0.06	0.10	[-0.15;0.26]	.579	.961	0.01	0.16	[-0.30;0.32]	.949	.981	0.18	0.14	[-0.10;0.46]	.209	.800
Pretest Value	0.56***	0.10	[0.37;0.75]	<.001		0.72***	0.13	[0.46;0.97]	<.001		0.75***	0.14	[0.48;1.03]	<.001	
R^2	0.47					0.42					0.41				
Model 2															
Quality of Delivery	-0.10	0.08	[-0.26;0.05]	.181	.800	0.00	0.19	[-0.37;0.36]	.981	.981	-0.09	0.10	[-0.30;0.11]	.378	.924
Pretest Value	0.55***	0.12	[0.32;0.79]	<.001		0.72***	0.16	[0.41;1.03]	<.001		0.70***	0.15	[0.41;0.99]	<.001	
R^2	0.46					0.40					0.40				

Table A16

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Adherence	0.15	0.14	[-0.12;0.43]	.280	.800
Pretest Value	0.61***	0.06	[0.50;0.72]	<.001	
R^2	0.51				
Model 2					
Quality of Delivery	-0.07	0.08	[-0.23;0.10]	.419	.924
Pretest Value	0.60***	0.07	[0.47;0.72]	<.001	
R^2	0.48				

Note. Two-tailed significance levels are reported. The p-values of the adherence and quality of delivery predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A17

Effects of Instructor-rated Core Components on Statistical Literacy Outcomes

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Predict-observe-explain	0.54***	0.13	[0.29;0.79]	<.001	<.001	-0.11	0.17	[-0.44;0.22]	.519	.769	-0.75***	0.08	[-0.92;-0.59]	<.001	<.001
Pretest Value	0.73***	0.16	[0.41;1.05]	<.001		0.37*	0.17	[0.04;0.70]	.030		0.54*	0.22	[0.11; 0.97]	.013	
R^2	0.67					0.14					0.56				
Model 2															
Cooperative Learning	-0.26	0.24	[-0.72;0.21]	.276	.581	-0.30*	0.11	[-0.52;-0.08]	.007	.047	0.04	0.20	[-0.35;0.43]	.849	.951
Pretest Value	0.64**	0.20	[0.25;1.02]	.001		0.29	0.16	[-0.02; 0.60]	.064		0.37	0.25	[-0.12;0.86]	.134	
R^2	0.47					0.23					0.13				

Note. Two-tailed significance levels are reported. The p-values of the predict-observe-explain and cooperative learning predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A17

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Predict-observe-explain	-0.27	0.12	[-0.50; -0.03]	.025	.100	-0.07	0.20	[-0.46;0.33]	.739	.896	0.24	0.10	[0.05;0.43]	.014	.062
Pretest Value	0.18**	0.06	[0.08; 0.29]	.001		0.05	0.22	[-0.39;0.48]	.827		0.46**	0.14	[0.18;0.74]	.001	
R^2	0.14					0.01					0.23				
Model 2															
Cooperative Learning	-0.01	0.11	[-0.23;0.22]	.954	.990	-0.02	0.11	[-0.23;0.19]	.832	.951	-0.10	0.10	[-0.29;0.09]	.308	.616
Pretest Value	0.16*	0.07	[0.02;0.30]	.029		0.08	0.22	[-0.36;0.51]	.731		0.42**	0.15	[0.12;0.72]	.006	
R^2	0.04					0.01					0.19				

Note. Two-tailed significance levels are reported. The p-values of the predict-observe-explain and cooperative learning predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A18

Correspondences of Instructor-rated Core Components on Motivational Beliefs within Intervention Condition

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Predict-observe-explain	0.20	0.08	[0.04;0.36]	.012	.060	0.32***	0.08	[0.16;0.48]	<.001	<.001	0.07	0.04	[-0.02;0.15]	.129	.344
Pretest Value	0.53***	0.09	[0.36;0.71]	<.001		0.71***	0.16	[0.40;1.02]	<.001		0.71***	0.15	[0.42;1.01]	<.001	
R^2	0.46					0.49					0.40				
Model 2															
Cooperative Learning	0.08	0.11	[-0.14;0.29]	.492	.757	0.01	0.17	[-0.31;0.34]	.932	.978	0.19	0.10	[0.00;0.39]	.046	.167
Pretest Value	0.54***	0.10	[0.35;0.74]	<.001		0.71***	0.16	[0.39;1.02]	<.001		0.70***	0.14	[0.43;0.98]	<.001	
R^2	0.45					0.41					0.42				

Table A18

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Predict-observe-explain	0.09	0.11	[-0.12;0.31]	.393	.655
Pretest Value	0.62***	0.06	[0.49;0.75]	<.001	
R^2	0.48				
Model 2					
Cooperative Learning	0.17	0.09	[-0.01;0.35]	.065	.217
Pretest Value	0.61***	0.06	[0.48;0.73]	<.001	
R^2	0.51				

Note. Two-tailed significance levels are reported. The p-values of the predict-observe-explain and cooperative learning predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A19

Correspondences of Children-rated Core Components on Statistical Literacy Outcomes within Intervention Condition

Variable	Data-based Argumentation					Views on Variability					Draws to Decision				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Predict-observe-explain	-0.14	0.15	[-0.43;0.15]	.351	.217	-0.03	0.15	[-0.32;0.27]	.856	.951	0.27	0.20	[-0.11;0.66]	.165	.388
Pretest Value	0.69***	0.16	[0.38;1.00]	<.001		0.33*	0.15	[0.03;0.63]	.028		0.43	0.25	[-0.07;0.92]	.092	
R^2	0.45					0.12					0.17				
Model 2															
Cooperative Learning	-0.25	0.14	[-0.53;0.02]	.072	.222	0.00	0.13	[-0.26;0.26]	.998	.998	0.21	0.27	[-0.32;0.75]	.433	.693
Pretest Value	0.70***	0.16	[0.38;1.02]	<.001		0.33*	0.14	[0.05;0.60]	.020		0.33	0.27	[-0.19;0.86]	.213	
R^2	0.49					0.12					0.17				

Note. Two-tailed significance levels are reported. The p-values of the predict-observe-explain and cooperative learning predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A19

(continued)

Variable	Decision Threshold					Alternation Bias					Confirmation Bias				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Predict-observe-explain	0.10	0.22	[-0.34;0.54]	.657	.844	-0.22	0.15	[-0.52;0.07]	.139	.348	0.30	0.11	[0.08;0.52]	.009	.051
Pretest Value	0.17*	0.07	[0.02;0.31]	.023		0.06	0.22	[-0.37;0.49]	.793		0.31	0.16	[-0.01;0.63]	.056	
R^2	0.04					0.05					0.26				
Model 2															
Cooperative Learning	0.18	0.21	[-0.23;0.59]	.384	.655	-0.09	0.16	[-0.40;0.22]	.581	.801	0.29	0.19	[-0.08;0.66]	.122	.344
Pretest Value	0.14	0.08	[-0.02;0.30]	.081		0.05	0.22	[-0.37;0.47]	.828		0.30	0.18	[-0.06;0.66]	.106	
R^2	0.09					0.01					0.25				

Note. Two-tailed significance levels are reported. The p-values of the predict-observe-explain and cooperative learning predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

Table A20

Correspondences of Children-rated Core Components on Motivational Beliefs within Intervention Condition

Variable	Self-concept					Intrinsic Value					Attainment Value				
	B	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}	*B*	*SE*	95% CI	*p*	p_{BH}
Model 1															
Predict-observe-explain	0.33*	0.11	[0.11;0.55]	.003	.030	0.49*	0.18	[0.14;0.84]	.006	.047	0.14	0.34	[-0.52;0.80]	.675	.844
Pretest Value	0.38**	0.12	[0.16;0.61]	.001		0.35	0.19	[-0.03;0.73]	.074		0.58	0.30	[-0.01;1.17]	.052	
R^2	0.57					0.51					0.41				
Model 2															
Cooperative Learning	0.06	0.11	[-0.14;0.27]	.545	.779	0.21	0.22	[-0.22;0.64]	.338	.638	0.11	0.26	[-0.40;0.62]	.660	.844
Pretest Value	0.52***	0.11	[0.30;0.75]	<.001		0.60**	0.22	[0.17;1.02]	.006		0.68***	0.15	[0.39;0.97]	<.001	
R^2	0.46					0.44					0.41				

Table A20

(continued)

Variable	General Self-efficacy				
	B	*SE*	95% CI	*p*	p_{BH}
Model 1					
Predict-observe-explain	-0.01	0.18	[-0.36;0.33]	.934	.978
Pretest Value	0.61***	0.09	[0.44;0.78]	<.001	
R^2	0.48				
Model 2					
Cooperative Learning	0.17	0.15	[-0.12;0.46]	.254	.564
Pretest Value	0.57***	0.06	[0.45;0.69]	<.001	
R^2	0.51				

Note. Two-tailed significance levels are reported. The p-values of the predict-observe-explain and cooperative learning predictors were corrected for multiple testing with the Benjamini-Hochberg procedure.
* p < .05. ** p < .01. *** p < .001.

5

General Discussion

5. General Discussion

Statistical literacy is assumed to enable people to understand and evaluate statistical information in daily life, to enable people to be independent from other people's interpretations of data, and for democracies to function (Bodemer, 2014; United Nations, 2012, Wallman, 1993). In today's societies, adults as well as students are struggling with understanding basic statistical concepts (Batanero & Serrano, 1999; delMas & Liu, 2005; delMas et al., 1999; Galesic & Garcia-Retamero, 2010; Gigerenzer et al., 2007), a situation that highlights the need for the promotion of statistical literacy. As an understanding of statistical concepts is already beginning to develop during the primary school years (English & Watson, 2013; Piaget & Inhelder, 1975; Watson & Moritz, 2000), promoting statistical literacy from early on is crucial for building a statistically literate society. However, despite the relevance of the topic on personal and societal levels, research on the promotion of statistical literacy is scarce, and existing studies show methodological weaknesses (e.g., Bakker, 2004; Ben-Zvi, 2004, 2006; Krause et al., 2009).

The present dissertation addressed this need for methodologically strong intervention studies by developing and evaluating a statistical literacy intervention. To this end, a statistical literacy intervention was developed to promote gifted primary school children's data-based argumentation, understanding of the concept of variability, and motivation in data-related tasks on the basis of existing research studies (see Chapter 1). In Study 1, the efficacy of the intervention was tested in a standardized setting by having university staff carry out the intervention. In Study 2, an instrument for measuring the decision threshold was developed to be able to assess whether children's need for certainty in decision-making was promoted appropriately or too extensively by the intervention. In Study 3, this new measurement instrument was included as a complement to existing ones, and the effectiveness of the intervention was assessed under less standardized conditions by training course instructors from the field to carry it out.

In the following discussion, I put the research from this dissertation into context. First, I summarize the general findings across all three studies. Second, I outline the strengths and limitations of the three studies. Third, I derive theoretical and practical implications as well as future directions. And lastly, I draw a conclusion from all the research.

5.1 General Findings Across the Studies

The present dissertation was aimed at closing the research gap on whether aspects of statistical literacy can be promoted in primary school children. To answer this question, an extracurricular statistical literacy intervention was developed for gifted primary school children from the third and fourth grades. This dissertation was conducted to guide readers across the steps of detecting the needs of the target group to the effectiveness study. This process reflects the gradual realization of the real-world effectiveness of the intervention that was the focus of this work (Herbein et al., 2018b; Humphrey et al., 2016; Lendrum & Wigelsworth, 2013). Study 1 focused on the efficacy of the intervention under standardized conditions. Study 3 focused on the effectiveness under conditions closer to the real-world application of the intervention. When moving from an efficacy study to an effectiveness study, a reduction in the effects of an intervention can often be observed (Hulleman & Cordray, 2009; Lendrum & Wigelsworth, 2013). To still be able to observe the intervention's effects on the children's understanding of the concept of variability, Study 2 added a new instrument to complement the existing measures in the evaluation of the intervention. Accordingly, I summarize and discuss the three papers in this dissertation with a focus on the development of the new measurement instrument to assess the decision threshold and the intervention's main effects. Additionally, I summarize and discuss the differential effects of the intervention.

5.1.1 Measuring the decision threshold

Aspects of children's statistical literacy were assessed with a variety of different measurement instruments. In the first study, we assessed children's jumping to conclusions bias with the beads task (Dudley et al., 1997a, 1997b; Garety et al., 2005; Huq et al., 1988). This bias resembles decision-making on the basis of too little evidence. In the efficacy study, we found that children in the intervention group showed a reduction in their jumping to conclusions bias and had more draws to decision after the intervention compared with the waitlist control group. However, the draws to decision measure had several shortcomings (see Chapter 1.1.4). Amongst others, it was not quite clear whether more draws to decisions are always better, as more draws could also reflect a decision-making style that is too cautious. Thus, in Study 2, we evaluated a decision threshold instrument that was based on objective probability values. This instrument measures the decision threshold as a percentage, which describes the minimum probability a person needs to be certain enough to make a decision. It can be compared with the concept of the alpha level. The common alpha level of 5% would be reflected by a decision threshold of 95%. In this study, the resulting mean decision threshold

of all the children was below 70%, which can be assumed to be quite low. In the effectiveness study, the decision threshold of the children from the intervention group increased from approximately 80% to approximately 90%. As a decision threshold of 95% is considered optimal, this development seemed good. However, future studies need to investigate how this measure is related to actual behavioral outcomes and which decision thresholds can be viewed as appropriate for different contexts.

5.1.2 Effects on data-based argumentation

The intervention effects on children's data-based argumentation were assessed in the efficacy and effectiveness studies. The same measurement instrument was used at pretest and posttest. Children were presented several items that included an interpretation of a situation and the corresponding data. They were then prompted to decide whether the statement fit the data or not and to explain why they gave that answer. Responses were then rated as correct only when the child made the right decision and had a data-related explanation for it. Interrater reliability was high, which speaks in favor of the measurement instrument. Also, the intervention effects on data-based argumentation were positive and statistically significant in both studies. Compared with the waitlist control group, the children in the intervention condition had a gain of around two thirds of a standard deviation on their data-based argumentation. According to the guidelines suggested by Kraft (2020), the effects we found were large in the context of educational interventions. Even though previous studies had already found positive effects of interventions on data-based argumentation (Ben-Zvi, 2006; Paparistodemou & Meletiou-Mavrotheris, 2008; van Dijke-Droogers et al., 2024), to my knowledge, this study is the first randomized controlled field trial to find evidence for the promotion of data-based argumentation. This finding falls in line with the expected results. This intervention was built to enhance the data-based argumentation skills of the participating children by letting them come up with predictions and then presenting them with data that they had to use to explain the correctness of their predictions (POE approach; Gunstone & White, 1981) and by providing a cooperative learning environment (e.g., Capar & Tarim, 2015) in which they could verbalize and exchange their points of view.

5.1.3 Effects on the children's understanding of the concept of variability

The understanding of the concept of variability was assessed with several different measurement instruments. Two of these instruments were assessed in the efficacy and effectiveness studies. The measures were views on variability and draws to decision. Both

measures were positively affected by the intervention in the efficacy study. However, they were not significantly affected in the effectiveness study. More measurement instruments were added in the effectiveness study to gain deeper knowledge of how the intervention affects children's understanding of the concept of variability. One of those instruments was developed in Study 2. The decision threshold measure was found to have high accuracy and convergent validity and offers a new way to measure the jumping to conclusions bias just as the draws to decision measure did. This point is relevant because the jumping to conclusions bias is related to superstitious and conspiracy-related beliefs (Georgiou et al., 2021; Kuhn et al., 2022; Sanchez & Dunning, 2021), and it resembles the tendency to rely on insufficient evidence. The effectiveness study found a positive and statistically significant effect on the decision threshold. Therefore, the intervention has an effect on children's jumping to conclusions bias. This effect suggests that children in the intervention condition look for more evidence before making a decision than the control group does. Additionally, instruments for measuring the alternation bias and the confirmation bias were added in the effectiveness study. A significant reduction in the alternation bias was found. This finding means that the intervention study was successful in making the children generate sequences with properties that are more likely to be produced by random processes. However, the confirmation bias was not significantly affected by the intervention. It was assessed with a self-assessment questionnaire. Thus, the participants did not appear less likely to cherry-pick datapoints to reduce variability in favor of their preexisting beliefs.

In summary, the results were mixed. Some measures were affected in the intended direction, whereas others were not affected. On a positive note, there were no effects that went contrary to the expected direction. The nonsignificant results in the effectiveness study can be explained by the fact that usually weaker effects of interventions in studies with less standardized conditions can be expected (Hulleman & Cordray, 2009; Lendrum & Wigelsworth, 2013). They could also be explained by changes in the course manual that were made between the two studies. However, only a few activities were changed or taken out. Perhaps some aspects of the intervention related to variability were not communicated well in the training of the course instructors. This idea could explain why there were smaller and nonsignificant effects in the effectiveness study. The statistically significant effects can be explained by the children understanding the concept of variability better. They may have learned that variability is often present in actual data and that it resembles noise through which they have to search for a central tendency that resembles a signal (Konold & Pollatsek, 2002). Such changes might be explained by the children working with real data as shown in previous

research (Bakker, 2004; Ben-Zvi, 2004; Dierdorp et al., 2017) in which the children were confronted with natural variability in real data sets. Another explanation might be the intervention sessions that focused on the concept of randomness and the law of large numbers. In these sessions, the children learned about variability in random data and that results from larger samples present a better picture of the population than results from smaller samples.

5.1.4 Effects on motivation in data-related tasks

Motivation in data-related tasks was assessed in the efficacy and effectiveness studies with self-assessment questionnaires. Motivation was generally already high at pretest. However, although most of the effects were not statistically significant, there were some statistically significant positive effects on motivation. Both studies found positive and statistically significant effects on children's self-concept in data-related tasks. With an effect size of around two thirds of a standard deviation, these effects can be interpreted as large (Kraft, 2020). Thus, the intervention was successful in promoting children's self-concept in data-related tasks. This finding falls in line with previous research that showed that motivation in statistics can be fostered with cooperative learning settings (Krause et al., 2009) and working with real data (Gopal et al., 2018). However, not all motivational variables were positively affected. In both studies, intrinsic value was not statistically significantly altered by the intervention. Attainment value was positively affected only in the effectiveness study, a finding that was unexpected, as effects usually turn out to be smaller in intervention studies that are closer to their real-world application (Hulleman & Cordray, 2009; Lendrum & Wigelsworth, 2013). In the effectiveness study, we additionally tested whether the intervention affected the non-domain-specific motivational aspect of general self-efficacy. Similarly, this construct did not appear to be affected by the intervention.

In summary, the effects on motivation were partly positive. The most consistent effect was the positive effect on children's self-concept in data-related tasks. This effect meant that children in the intervention condition thought of themselves as more competent in data-related tasks after the intervention compared with the control group. This finding might be due to cooperative learning methods (e.g., Capar & Tarim, 2015), but it could also be due to other factors. For example, children might not have known much about what data are and what can be done with them before the intervention. For this reason, they could have observed an enhancement in their abilities relatively easily compared with other more prominent subjects such as mathematics and German. The effect could also be due to other parts of the intervention design. For example, at the end of each session, the children were asked to recap what they

learned and to give examples from their own lives of what they learned. Thus, at the end of each session, they were left with the knowledge that they had learned something useful that day. However, in general, a positive finding is that there were no negative effects on the children's motivation. The children could potentially have been negatively affected by the big-fish-little-pond effect (Marsh, 1987; Marsh & Parker, 1984). When switching from their normal classrooms to the intervention settings, in which generally more able children tend to take part, they could have compared themselves with children who were more able than their usual peers. On the one hand, the big-fish-little-pond effect could explain why there were not more positive effects on various motivational aspects, but on the other hand, this effect might not have unfolded because most children did not know much about data at the beginning of the course.

5.1.5 Differential effects

Different subgroups of children may benefit more or less from the same intervention on the basis of gender, age, or previous knowledge. Understanding which interventions work best for specific subgroups allows for a more efficient use of resources and for evaluations of equity in education (e.g., Maden et al., 2017). To test for such differences, differential effects were assessed in the two studies that evaluated the statistical literacy intervention in this dissertation. In the efficacy study, differential effects of previous knowledge, gender, grade level, and fluid intelligence were assessed. Most differential effects were not statistically significant. We did not find any differential effects of previous knowledge or gender. There were differential effects of grade level on views on variability and self-concept in data-related tasks. Also, a differential effect of fluid intelligence on data-based argumentation was found. In the effectiveness study, differential effects of previous knowledge, gender, age, grade level, and fluid intelligence were assessed. None of the observed differential effects were statistically significant.

In summary, there were only three differential effects in the efficacy study and none in the effectiveness study. Considering the number of possible differential effects, the number we found is relatively small. This finding speaks in favor of the equity of the statistical literacy intervention. Most children were affected equally by the intervention. The fact that the effectiveness study did not replicate the differential effects of the efficacy intervention might have several reasons. It could be the case that the changes made to the intervention contents optimized the impact they had on the children in terms of equity. It is possible that the differential effects in the efficacy study were incidental findings. It is also possible that the differential effects in the effectiveness study could not be found because the main effects were

smaller and the statistical power might have been too small to detect them. However, in general, the overall picture seems to be that the intervention works equally well for all subgroups of children.

5.2 Strengths and Limitations

Several strengths and limitations must be considered when interpreting the results of the present dissertation. The first strength of this dissertation is that the statistical literacy intervention that was at the heart of this dissertation was conceptualized on the basis of evidence from other studies and was evaluated to be effective. The intervention went through the first five steps to gradually develop and evaluate the real-world effectiveness of an intervention, from detecting the needs of the target group to the effectiveness study (Herbein et al., 2018b; Humphrey et al., 2016; Lendrum & Wigelsworth, 2013). This dissertation is an example of how such projects can be carried out and can result in practice-oriented course concepts for sustainable use in the field.

Next, the study designs of the efficacy and effectiveness studies are a strength of this dissertation. Both studies were conducted as randomized controlled field trials. This study design is assumed to be the *gold standard* in educational and psychological research (Lendrum & Wigelsworth, 2013; Torgerson & Torgerson, 2013). In both studies, the participating children took a pretest, and then they were randomly allocated to the intervention group or the waitlist control group. Subsequently, children from the intervention group received the intervention, and after that, both groups participated in a posttest. For reasons of fairness, the children in the control group received the intervention after the posttest. This design provides the opportunity to find out exactly which changes in the observed variables can be attributed to the intervention instead of other sources of variability from outside the intervention (Towne & Hilton, 2004). Especially in the field of statistical literacy, such studies are scarce, as they are more difficult to conduct, a fact that emphasizes the value of these studies.

The wide variety of measurement instruments that were used is another strength of this dissertation. As explained in Chapter 1.1, statistical literacy consists of various skills and dispositions. An overarching measurement with one instrument is difficult and might not be appropriate. Through the many different instruments used in Studies 1 and 3, the changes in data-based argumentation, understanding the concept of variability, and motivation could be examined more closely. In addition to the many existing measurement instruments, Study 2 also developed another instrument for measuring the decision threshold, thereby allowing us to take a closer look at how jumping to conclusions can manifest itself in children. Altogether,

the variety of measurement instruments allowed conclusions to be drawn about the effectiveness of the intervention and how it might be adapted to be even more effective.

Another strength of this dissertation is the use of appropriate statistical methods. To evaluate the measurement instrument in Study 2, we employed a signal detection theory approach, which is appropriate for understanding decision-making under conditions of noise and uncertainty (Green & Swets, 1966; Stanislaw & Todorov, 1999). In Studies 1 and 3, hierarchical multiple regression analyses were used to assess the intervention effects while controlling for baseline differences. These models are fit for nested data structures (e.g., children in groups) and for settings in which various factors may influence the intervention's outcomes (Hoyt et al., 2008; Raudenbush, 1988). In addition to the main effects of the intervention, differential effects were assessed to further understand the impact of the intervention on subgroups of children. This exploration of differential effects revealed that the intervention had similar effects in most subgroups of children.

As a further strength, implementation fidelity was assessed in this dissertation with several different measures. Implementation fidelity is assessed to ensure that an intervention is carried out as intended and that the intervention's effect can be traced back to the contents of the intervention (Smith et al., 2007). In the efficacy study, the fidelity of the implementation was secured by having university staff who were close to the development of the course conduct the intervention. In the effectiveness study, there were several different measures. The adherence to the course manual and the quality of delivery were assessed by the course instructors in every session. Both were rated very highly. Also, course instructors rated the implementation of the core components during every session, thereby revealing that most of the time they were implemented well. Additionally, at posttest, the children from the intervention group rated how well several aspects of the core components helped them learn during the intervention's sessions. These ratings revealed that both received high ratings, but the POE approach (Gunstone & White, 1981) was rated as more helpful than cooperative learning methods (e.g., Capar & Tarim, 2015). All measures of implementation fidelity were also used as predictors in separate regression analyses to detect whether they might be related to the promotion of certain aspects of statistical literacy. Most of the predictors were not significantly related to the outcomes. However, one thing stood out. The POE approach rated by both the instructors and the children positively predicted the intrinsic value of data-related tasks at posttest when the pretest values were controlled for. Even if these analyses do not allow any conclusions to be drawn about causality, they could indicate that the POE approach was related to the enjoyment of the intervention.

Besides these strengths, some limitations have to be considered. First, all three studies were conducted in the context of the HCAP, which is an extracurricular enrichment program for gifted children (for more information, see Trautwein et al., 2023). These children were used as an adequate starting point for researching statistical literacy interventions, as statistical concepts might be difficult to understand and abilities in this domain might be connected to intelligence (Sproesser et al., 2018). Nonetheless, this specific target group limits the generalizability of the findings. Less able children might have more problems understanding statistical concepts and might therefore not be affected in the same way by the intervention. Thus, future research could test whether the intervention works in the same way for a broader target group.

Next, even though children's statistical literacy was assessed with many measurement instruments, a question that remains is whether there are effects on aspects of statistical literacy that are conceptually further away from the intervention's contents. More assessment tools could be fruitful for assessing whether there are transfer effects on abilities and dispositions that demonstrate their relevance for children's daily lives. For example, children's motivation in data-related tasks was assessed only with self-assessment questionnaires. It would be interesting to know whether their actual behavior changes such that they become more motivated to use data in relevant situations.

Another limitation was the relatively small sample sizes used in the intervention studies. The small samples limit the generalizability of the studies, as the samples might not represent the broader population. Also, smaller sample sizes reduce the study's power to detect statistically significant effects, especially if the intervention's impact is small or subtle. Especially the differential effects are hard to detect with small samples. Furthermore, with fewer participants, the results are more susceptible to effects of outliers or random fluctuations, which can distort findings and make them less reliable. Such fluctuations can lead to unstable estimates of effect sizes, making it harder to draw accurate conclusions about the intervention's effects. In conclusion, the effectiveness that the intervention demonstrated would profit from evidence from studies with larger sample sizes.

Lastly, even though the intervention was based on the POE approach (Gunstone & White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015), the studies cannot show that these defined core components were indeed responsible for the effectiveness of the intervention. To demonstrate causal effects, a research design that tests an intervention with the core components would need to be tested against an intervention without them. However, some evidence for their effectiveness can be drawn from the implementation fidelity measures

that were built around the core components and showed some relationships to the intervention's outcomes. Thus, it stands to reason that these core components could contribute to the effectiveness of the intervention, but more evidence is needed to confirm this idea.

5.3 Implications and Future Directions

On the one hand, the present dissertation had the goal of determining whether a new measurement instrument for detecting individual decision thresholds on the basis of objective probability values had appropriate levels of accuracy and validity. On the other hand, the present dissertation was conducted to determine whether a newly developed statistical literacy intervention could be successfully implemented in practice. However, there are questions that remain open. From these results, future research questions can be derived, and I discuss them in the following. Afterwards, I summarize future directions for educational practice based on the results of the present dissertation.

5.3.1 Implications for future research

In Study 2, a measurement instrument was developed and evaluated to measure the decision threshold. To do so, we used an SDT approach (Green & Swets, 1966; Stanislaw & Todorov, 1999) that was based on objective probability values of the items to derive individual decision thresholds for each participant and to analyze the measurement instrument's accuracy. The instrument was found to be accurate and showed good convergent validity. However, some questions remain open. Individual decision thresholds were derived, but it is not yet clear whether there is an optimal decision threshold. Future studies could explore how the decision threshold is related to behavior (e.g., health decisions) or how it is related to well-being. Additionally, it would be useful to explore more of the nomological net around the decision threshold in children. In adults, the jumping to conclusions bias was found to be related to superstitious and conspiracy-related beliefs (Georgiou et al., 2021; Kuhn et al., 2022; Sanchez & Dunning, 2021). Does a similar relationship already exist in children? Knowing the answer to this question could aid the design of early interventions for children to enhance their decision threshold and potentially make them more resistant to superstitious beliefs and fake news.

Studies 1 and 3 evaluated the efficacy and effectiveness of a newly developed statistical literacy intervention. We evaluated the intervention's effects under standardized conditions and more field-like conditions. The next step in gradually developing and evaluating the real-world effectiveness of an intervention would be a scaling-up study (Herbein et al., 2018b; Humphrey et al., 2016; Lendrum & Wigelsworth, 2013), which entails conducting another randomized

controlled field trial with a broader target group and a larger sample. Such a study could help to raise more evidence of the effectiveness of the program. Another way to find out more about the intervention's effectiveness would be to test whether there are long-term effects. In the efficacy and effectiveness studies, posttests were administered only 1 day to 1 week after the last session of the intervention. Adding a follow-up test weeks or months after the posttest could determine whether the intervention has lasting effects.

There are also some open questions about why the statistical literacy intervention was effective. The intervention's effectiveness was based on two defined core components: the POE approach (Gunstone & White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015). Such core components have been demonstrated to reveal how the instructional goals of an intervention are achieved (Blase & Fixsen, 2013; Nelson et al., 2012). However, the study design used in the efficacy and effectiveness studies can disclose only whether the intervention is effective but not why it is effective. Using implementation fidelity measures of the core components as predictors in the effectiveness study's analysis revealed to some degree how the core components are related to some of the outcomes of the intervention. However, to determine whether they contribute to the intervention's effectiveness, an intervention with core components would need to be tested against an intervention without them. Another way to find out more about the intervention's effectiveness would be to test single activities on their effectiveness. The intervention consists of many different activities that are included in the intervention's manual. Some of those activities might contribute more to the effectiveness of the intervention than others. Extracting the activities that are thought to be most effective and testing them in a smaller randomized controlled trial could enhance the understanding of why the intervention is effective and make individually effective activities more accessible to a broader audience. Thus, such an approach could have a larger impact on statistical literacy in our society. For instance, some of these activities could potentially find their way into regular school settings to promote students' statistical literacy.

5.3.2 Implications for educational policy and practice

First, the results of the present dissertation provide evidence for the efficacy and effectiveness of the statistical literacy intervention that was developed as part of a statewide extracurricular enrichment program for gifted primary school children in the German state of Baden-Württemberg. These findings point to the effectiveness of funding and support for such programs. In this program, the statistical literacy intervention is one of several evidence-based interventions called HCCs. These courses are tailored to the needs of the target group and have

been scientifically evaluated. These tested interventions serve to ensure the quality of the program and to ensure that the gifted children, who are potential future decision-makers, are well-equipped (Lee et al., 2021). To provide everything necessary to implement this statistical literacy intervention, a course manual and instructor training are permanent components of the HCAP.

Second, the intervention was shown to promote aspects of primary school children's statistical literacy. This finding provides evidence for the malleability of statistical literacy at an early age when children start to develop statistical concepts (English & Watson, 2013; Piaget & Inhelder, 1975; Watson & Moritz, 2000). The promotion of statistical literacy is recommended and desired by statisticians such as in the *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II* by the American Statistical Association (Bargagliotti et al., 2020). The present dissertation provides evidence that statistical literacy can be developed from early on and shows how it can be promoted. The results indicate that an extracurricular promotion of aspects of statistical literacy is potent with the POE approach (Gunstone & White, 1981) and cooperative learning methods (e.g., Capar & Tarim, 2015). Similar approaches might be useful in regular school contexts.

Third, the present dissertation used a wide variety of instruments to measure aspects of statistical literacy and developed and evaluated a new measurement instrument for assessing children's decision thresholds. To evaluate the effectiveness of educational policies and practices, appropriate measurement instruments are needed. The present dissertation derived three fundamental aspects of statistical literacy (data-based argumentation, understanding the concept of variability, and motivation in data-related tasks) and provided adequate instruments for measuring them. However, these instruments are restricted to fundamental aspects of statistical literacy that are situated primarily in the context of primary school education. Nonetheless, they can be extended or supplemented to serve a broader context.

Finally, the findings of this dissertation can be viewed in the light of current PISA results. In the last decade, performance in mathematics and science has been decreasing in OECD countries (OECD, 2023). To strengthen resilience in performance, the OECD identified several potentially helpful actions, one of which is securing high-quality educational material. The present dissertation may provide such material, as the statistical literacy intervention was shown to be effective in providing several aspects of statistical literacy. The underlying learning materials may be useful in a more general educational context. However, evidence for this context is needed. Regardless, evaluating learning material and its implementation in

randomized controlled field trials could be a useful approach for providing effective high-quality educational material.

5.4 Conclusion

This dissertation focused on developing and evaluating the statistical literacy intervention “Luck or genius? Understanding data and making predictions.” The intervention’s efficacy was demonstrated under standardized conditions with university staff as the course instructors. The intervention's effectiveness was demonstrated under conditions closer to real-world implementation with trained course instructors from the field. In between the two studies, a new measurement instrument was developed and evaluated to assess children’s decision thresholds. In general, the intervention was found to be effective in promoting primary school children’s data-based argumentation and partly effective in promoting their understanding of the concept of variability as well as their motivation in data-related tasks.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process.

During the preparation of this work the author used ChatGPT (OpenAI, 2025) in order to improve style and grammar of the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of this dissertation.

References

Acee, T. W., & Weinstein, C. E. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *The Journal of Experimental Education*, *78*(4), 487-512. https://doi.org/10.1080/00220970903352753

Arens, A. K., Trautwein, U., & Hasselhorn, M. (2011). Erfassung des Selbstkonzepts im mittleren Kindesalter: Validierung einer deutschen Version des SDQ I [Assessment of self-concept in middle childhood: Validation of a German version of the SDQ I]. *Zeitschrift für Pädagogische Psychologie.* https://doi.org/10.1024/1010-0652/a000030

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, *3*(2), 64-83. https://doi.org/10.52041/serj.v3i2.552

Balzan, R. P., Ephraums, R., Delfabbro, P., & Andreou, C. (2017). Beads task vs. box task: The specificity of the jumping to conclusions bias. *Journal of Behavior Therapy and Experimental Psychiatry*, *56*, 42-50. https://doi.org/10.1016/j.jbtep.2016.07.017

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education. *American Statistical Association*. https://www.amstat.org/docs/default-source/amstat-documents/gaiseiiprek-12_full.pdf

Barron, G., & Leider, S. (2010). The role of experience in the Gambler's Fallacy. *Journal of Behavioral Decision Making*, *23*(1), 117-129. https://doi.org/10.1002/bdm.676

Batanero, C., & Serrano, L. (1999). The meaning of randomness for secondary school students. *Journal for Research in Mathematics Education*, *30*(5), 558-567. https://doi.org/10.2307/749774

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42-63. https://doi.org/10.52041/serj.v3i2.547

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute. https://www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/2D1_BENZ.pdf

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In: Ben-Zvi, D. & Garfield, J. (Eds.). *The Challenge of*

Developing Statistical Literacy, Reasoning and Thinking. Dodrecht: Kluwer Academic Publishers, 3-15. https://link.springer.com/chapter/10.1007/1-4020-2278-6_1

Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. *School Science and Mathematics*, *108*(8), 355-361. https://doi.org/10.1111/j.1949-8594.2008.tb17850.x

Blase, K., & Fixsen, D. L. (2013). *Core intervention components: Identifying and operationalizing what makes programs work. ASPE research brief.* US Department of Health and Human Services. https://eric.ed.gov/?id=ED541353

Blume, J. D., & Royall, R. M. (2003). Illustrating the law of large numbers (and confidence intervals). *The American Statistician*, *57*(1), 51-57. https://doi.org/10.1198/0003130031081

Bodemer, N. (2014). Gesundheitsrisiken verstehen: ein Bildungsproblem; Überlegungen zum risikokompetenten Bürger. *DIE Zeitschrift für Erwachsenenbildung*, *21*(2), 33-35. https://doi.org/10.3278/DIE1402W033

Capar, G., & Tarim, K. (2015). Efficacy of the cooperative learning method on mathematics achievement and attitude: A meta-analysis research. *Educational Sciences: Theory and Practice*, *15*(2), 553-559. https://doi.org/10.12738/estp.2015.2.2098

Cobb, G.W. (1993). Reconsidering statistics education: A national science foundation conference. *Journal of Statistics Education*, *1*(1), 1-28. https://doi.org/10.1080/10691898.1993.11910454

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801-823. https://doi.org/10.2307/2975286

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, *29*(2), 186-204. https://doi.org/10.1016/j.cedpsych.2004.01.004

Cruise, R., Cash, R., and Bolton, D. (1985). Development and validation of an instrument to measure statistics anxiety. In *Proceedings of the American Statistical Association, Statistical Education Section* (pp. 92-97). Alexandra, VA: American Statistical Association.

delMas, R. C., Garfield, J. B., & Chance, B. L. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3). https://doi.org/10.1080/10691898.1999.12131279

delMas, R. C., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *2*(1), 55–82.

Dierdorp, A., Bakker, A., Ben-Zvi, D., & Makar, K. (2017). Secondary students' considerations of variability in measurement activities based on authentic practices. *Statistics Education Research Journal*, *16*(2), 397-418. https://doi.org/10.52041/serj.v16i2.198

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997a). The effect of self-referent material on the reasoning of people with delusions. *British Journal of Clinical Psychology*, *36*, 575–584. https://doi.org/10.1111/j.2044-8260.1997.tb01262.x

Dudley, R. E. J., John, C. H., Young, A. W., & Over, D. E. (1997b). Normal and abnormal reasoning in people with delusions. *British Journal of Clinical Psychology*, *36*, 243–258. https://doi.org/10.1111/j.2044-8260.1997.tb01410.x

Dumont, H.; Neumann, M.; Maaz, K. & Trautwein, U. (2013). Die Zusammensetzung der Schülerschaft als Einflussfaktor für Schulleistungen. *Internationale und nationale Befunde. Psychologie in Erziehung und Unterricht*, 60, 163-183. http://dx.doi.org/10.2378/peu2013.art14d

Ellenberg, J. H. (1994). Selection bias in observational and experimental studies. *Statistics in Medicine*, *13*(5-7), 557-567. https://doi.org/10.1002/sim.4780130518

Engel, A. (1973). Outline of a Problem Oriented, Computer Oriented and Applications Oriented High School Mathematics Course, *International Journal of Mathematical Education in Science and Technology*, 4 (4), 455-492. https://doi.org/10.1080/0020739730040408

English, L., & Watson, J. (2013). Beginning inference in fourth grade: Exploring variation in measurement. In V. Steinle, L. Ball, & C. Bardini (Eds.), *Mathematics education: Yesterday, today and tomorrow* (pp. 274–281). Melbourne, Australia: MERGA. https://files.eric.ed.gov/fulltext/ED572843.pdf

English, L. D., & Watson, J. M. (2015). Exploring variation in measurement as a foundation for statistical thinking in the elementary school. *International Journal of STEM Education*, *2*, 1-20. https://doi.org/10.1186/s40594-015-001-x

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301. https://doi.org/10.1037/0033-295X.104.2.301

Fife, J.H., James, K. and Peters, S. (2020), A Learning Progression for Variability. ETS Research Report Series, 2020: 1-22. https://doi.org/10.1002/ets2.12286

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education*, 28(1), 96-105. https://doi.org/10.2307/749665

Gal, I. (2002). Adult statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70* (1), 1-25. https://doi.org/10.2307/1403721

Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine*, *170*(5), 462-468. https://doi.org/10.1001/archinternmed.2009.481

Garety, P. A., Freeman, D., Jolley, S., Dunn, G., Bebbington, P. E., Fowler, D. G., ... & Dudley, R. (2005). Reasoning, emotions, and delusional conviction in psychosis. *Journal of Abnormal Psychology*, *114*(3), 373. https://doi.org/10.1037/0021-843x.114.3.373

Garfield, J.B., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, *4*(1), 92-99. http://dx.doi.org/10.52041/serj.v4i1.527

Garfield, J.B., & Ben-Zvi, D. (2008). Learning to reason about variability. In J.B. Garfield, & D. Ben-Zvi (Eds.), *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice* (pp. 201-214). Springer. https://doi.org/10.1007/978-1-4020-8383-9_10

Gaspard, H., Dicke, A. L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., & Trautwein, U. (2015a). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, *51*(9), 1226-1240. https://doi.org/10.1037/dev0000028

Gaspard, H., Dicke, A. L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015b). More value through greater differentiation: gender differences in value beliefs about math. *Journal of Educational Psychology*, *107*(3), 663-677. https://doi.org/10.1037/edu0000003

Georgiou, N., Delfabbro, P., & Balzan, R. (2021). Conspiracy theory beliefs, scientific reasoning and the analytical thinking paradox. *Applied Cognitive Psychology*, *35*(6), 1523-1534. https://doi.org/10.1002/acp.3885

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the*

Public Interest, *8*(2), 53-96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Gopal, K., Salim, N. R., & Ayub, A. F. M. (2018, November). Influence of self-efficacy and attitudes towards statistics on undergraduates' statistics engagement in a Malaysian public university. In *Journal of Physics: Conference Series* (Vol. 1132, No. 1, p. 012042). IOP Publishing. http://dx.doi.org/10.1088/1742-6596/1132/1/012042

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. Prevention Science: *The Official Journal of the Society for Prevention Research*, *16*(7), 893–926. https://doi.org/10.1007/s11121-015-0555-x

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1, pp. 1969-2012). New York: Wiley.

Gunstone, R. F., & White, R. T. (1981). Understanding of gravity. *Science Education*, *65*(3), 291-299. http://dx.doi.org/10.1002/sce.3730650308

Gustina, R., Hastuti, I., Nizaar, M., & Syaharuddin, S. (2023). Predict observe explain learning model: Implementation and its influence on students' critical thinking ability and learning outcomes (a meta-analysis study). *Jurnal Kependidikan: Jurnal Hasil Penelitian dan Kajian Kepustakaan di Bidang Pendidikan, Pengajaran dan Pembelajaran*, *9*(2), 706-718. https://doi.org/10.33394/jk.v9i2.7388

Hanna, D., Shevlin, M., & Dempster, M. (2008). The structure of the statistics anxiety rating scale: A confirmatory factor analysis using UK psychology students. *Personality and Individual Differences*, *45*(1), 68-74. https://doi.org/10.1016/j.paid.2008.02.021

Herbein, E., Golle, J., Tibus, M., Schiefer, J., Trautwein, U., & Zettler, I. (2018a). Fostering elementary school children's public speaking skills: A randomized controlled trial. *Learning and Instruction*, *55*, 158-168. https://doi.org/10.1016/j.learninstruc.2017.10.008

Herbein, E., Golle, J., Tibus, M., Zettler, I., & Trautwein, U. (2018b). 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. http://dx.doi.org/10.1016/j.cedpsych.2018.09.003

Hood, M., Creed, P. A., & Neumann, D. L. (2012). Using the expectancy value model of motivation to understand the relationship between student attitudes and achievement in statistics. *Statistics Education Research Journal*, *11*(2), 72-85. https://doi.org/10.52041/serj.v11i2.330

Hoyt, W. T., Imel, Z. E., & Chan, F. (2008). Multiple regression and correlation techniques: Recent controversies and best practices. *Rehabilitation Psychology*, *53*(3), 321. https://doi.org/10.1037/a0013021

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88–110. https://doi.org/10.1080/19345740802539325

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature*. Retrieved from Education Endowment Foundation website: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Review_Final.pdf

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology Section A*, *40*(4), 801-812. https://doi.org/10.1080/14640748808402300

Johnson, D. W., Johnson, R. T., and Stanne, M. E. (2000). *Cooperative Learning Methods: A Meta-Analysis*. Minneapolis, MN: University of Minnesota Press. http://www.tablelearning.com/uploads/File/EXHIBIT-B.pdf

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning*, *2*(4), 269-307. http://dx.doi.org/10.1207/S15327833MTL0204_3

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454. https://doi.org/10.1016/0010-0285(72)90016-3

Kalaian, S. A., & Kasim, R. M. (2014). A meta-analytic review of studies of the effectiveness of small-group learning methods on statistics achievement. *Journal of Statistics Education*, *22*(1). http://dx.doi.org/10.1080/10691898.2014.11889691

Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic Data Explorations*. Emeryville, CA: Key Curriculum Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289. https://doi.org/10.2307/749741

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., Lipson, A. (1993). Inconsistencies in Students' Reasoning about Probability. *Journal for Research in Mathematics Education*, 24 (5), 392-414. https://doi.org/10.2307/749150

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253. https://doi.org/10.3102/0013189X20912798

Krause, U. M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and Instruction*, *19*(2), 158-170. http://dx.doi.org/10.1016/j.learninstruc.2008.03.003

Krummenauer, J., Emhart, M., & Kuntze, S. (2020). Können Kinder zu Beginn der ersten Klasse bereits mit statistischen Daten argumentieren? – Empirische Befunde aus einer Interviewstudie im Anfangsunterricht [Can children at the beginning of first grade already reason with statistical data? – Empirical findings from an interview study in the early years classroom]. *Mathematica Didactica*, *43*(2), 111-130. https://doi.org/10.18716/ojs/md/2020.1152

Krummenauer, J., Gutensohn, F., Aichele, J., Emhart, M., & Kuntze, S. (2022). Argumentation based on statistical data at the very beginning of primary school–evidence from two empirical studies. http://hdl.handle.net/10045/126620

Krummenauer, J., & Kuntze, S. (2018). Primary student's data-based argumentation—An empirical reanalysis. In Bergqvist, E., Österholm, M., Granberg, C., & Sumpter, L. (Eds.). *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3) (pp. 251–258). Umeå, Sweden: PME.

Krummenauer, J., & Kuntze, S. (2019). Primary students' reasoning and argumentation based on statistical data. In *Eleventh Congress of the European Society for Research in Mathematics Education* (No. 22). Freudenthal Group; Freudenthal Institute; ERME. https://hal.science/hal-02398118v1

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*(4), 674. https://doi.org/10.1037/0033-295X.96.4.674

Kuhn, D. (2011). *What is scientific thinking and how does it develop?* In U. Goswami (Eds.), The Wiley-Blackwell handbook of childhood cognitive development (S. 497–523). Wiley-Blackwell. http://dx.doi.org/10.1002/9781444325485.ch19

Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development*, *1*(1), 113-129. https://doi.org/10.1207/S15327647JCD0101N_11

Kuhn, S. A. K., Lieb, R., Freeman, D., Andreou, C., & Zander-Schellenberg, T. (2022). Coronavirus conspiracy beliefs in the German-speaking general population: endorsement rates and links to reasoning biases and paranoia. *Psychological Medicine*, *52*(16), 4162-4176. https://doi.org/10.1017/S0033291721001124

Kuntze, S., Martignon, L., Vargas, F., Engel, J. (2015). Competencies in understanding statistical information in primary and secondary school levels: An inter-cultural empirical study with German and Colombian students. *Avances de Investigación en Educación Matemática*, *7*, 5-25. http://dx.doi.org/10.35763/aiem.v1i7.103

Lee, S. Y., Matthews, M., Boo, E., & Kim, Y. K. (2021). Gifted students' perceptions about leadership and leadership development. *High Ability Studies*, *32*(2), 219-259. http://dx.doi.org/10.1080/13598139.2020.1818554

Lendrum, A., & Wigelsworth, M. (2013). The evaluation of school-based social and emotional learning interventions: Current issues and future directions. *Psychology of Education Review*, *37*, 70-76. http://dx.doi.org/10.53841/bpsper.2013.37.2.70

Levpuscek M. P., Cukon, M. (2022). That old devil called "Statistics": statistics anxiety in university students and related factors. *Center for Educational Policy Studies Journal*, *12*(1), 147-168. https://doi.org/10.26529/cepsj.826

MacKay, R. J., & Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. *Statistical Science*, *15*(3), 254-278. http://dx.doi.org/10.1214/ss/1009212817

Maden, M., Cunliffe, A., McMahon, N., Booth, A., Carey, G. M., Paisley, S., ... & Gabbay, M. (2017). Use of programme theory to understand the differential effects of interventions across socio-economic groups in systematic reviews—A systematic methodology review. *Systematic Reviews*, *6*, 1-23. https://doi.org/10.1186/s13643-017-0638-9

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W. (1990). *Self Description Questionnaire – I (SDQI). Manual.* Macarthur, N.S.W. Australia: University of Western Sydney.

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*(1), 213–231. https://doi.org/10.1037/0022-3514.47.1.213

Martignon, L., & Wassner, C. (2005). Schulung frühen stochastischen Denkens von Kindern [Training children in early stochastic thinking]. *Zeitschrift für Erziehungswissenschaft*, 8(2), 202-222. https://doi.org/10.1007/s11618-005-0134-1

Mayring, P. (2015). Qualitative Content Analysis: Theoretical Background and Procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. C. Presmeg (Eds.), Advances in mathematics education. *Approaches to qualitative research in mathematics education: Examples of*

methodology and methods (pp. 365–380). Dordrecht, New York: Springer. https://doi.org/10.1007/978-94-017-9181-6_13

McKenzie, J. D., Jr. (2004). Conveying the Core Concepts. In ASA Section on Statistical Education. (pp. 2755 - 2757). http://www.statlit.org/pdf/2004mckenzieasa.pdf

Moore, D. S. (1990). Uncertainty. *On the shoulders of giants: New approaches to numeracy*, 95-137. https://files.eric.ed.gov/fulltext/ED334084.pdf#page=104

Moritz, S., Scheunemann, J., Lüdtke, T., Westermann, S., Pfuhl, G., Balzan, R. P., & Andreou, C. (2020). Prolonged rather than hasty decision-making in schizophrenia using the box task. Must we rethink the jumping to conclusions account of paranoia?. *Schizophrenia Research*, *222*, 202-208. https://doi.org/10.1016/j.schres.2020.05.056

Moritz, S., Woodward, T. S., & Hausmann, D. (2006). Incautious reasoning as a pathogenetic factor for the development of psychotic symptoms in schizophrenia. *Schizophrenia Bulletin*, *32*(2), 327-331. https://doi.org/10.1093/schbul/sbj034

Moritz, S., Van Quaquebeke, N., & Lincoln, T. M. (2012). Jumping to conclusions is associated with paranoia but not general suspiciousness: A comparison of two versions of the probabilistic reasoning paradigm. *Schizophrenia Research and Treatment*, 2012. https://doi.org/10.1155/2012/384039

Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The emergence of scientific reasoning. In *Current topics in children's learning and cognition*. IntechOpen.

Nasser, F. M. (2004). Structural model of the effects of cognitive and affective factors on the achievement of Arabic-speaking pre-service teachers in introductory statistics. *Journal of Statistics Education*, *12*(1). http://dx.doi.org/10.1080/10691898.2004.11910717

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research*, *39*(4), 374–396. https://doi.org/10.1007/s11414-012-9295-x

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220. https://doi.org/10.1037/1089-2680.2.2.175

OECD (2023), *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en.

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments—A comprehensive review of the literature. *Teaching in Higher Education*, *8*(2), 195-209. https://doi.org/10.1080/1356251032000052447

OpenAI. (2025). *ChatGPT* (January 20 version) [Large language model]. https://chat.openai.com/chat

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, *7*(2), 83-106. https://doi.org/10.52041/serj.v7i2.471

Peters, S. A. (2011). Robust understanding of statistical variation. *Statistics Education Research Journal*, *10*(1), 52–88. https://doi.org/10.52041/serj.v10i1.367

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. Oxford, England: Norton.

Radovanović, J., & Sliško, J. (2013). Applying a predict–observe–explain sequence in teaching of buoyant force. *Physics Education*, *48*(1), 28. http://dx.doi.org/10.1088/0031-9120/48/1/28

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, *13*(2), 85-116. https://doi.org/10.3102/10769986013002085

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201–226). Kluwer. https://doi.org/10.1007/1-4020-2278-6_9

Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Kyama (Eds*.), Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89–96). Psychology of Mathematics Education. http://www.igpme.org/publications/current-proceedings/

Rebholz, F., Golle, J., Tibus, M., Ruth-Herbein, E., Moeller, K., & Trautwein, U. (2022). Getting fit for the Mathematical Olympiad: positive effects on achievement and motivation?. *Zeitschrift für Erziehungswissenschaft*, *25*(5), 1175-1198. https://doi.org/10.1007/s11618-022-01106-y

Ruiz-Primo, M. A., Li, M., Tsai, S. P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *47*(5), 583-608. http://dx.doi.org/10.1002/tea.20356

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 1-12. https://doi.org/10.1080/10691898.2002.11910678

Sanchez, C., & Dunning, D. (2021). Jumping to conclusions: Implications for reasoning errors, false belief, knowledge corruption, and impeded learning. *Journal of Personality and Social Psychology*, *120*(3), 789. https://doi.org/10.1037/pspp0000375

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. https://doi.org/10.1037/edu0000630

Schield, M. (1999). Statistical literacy: Thinking critically about statistics. *Of Significance*, *1*(1), 15-20.

Schutz, P. A., Drogosz, L. M., White, V. E., & Distefano, C. (1998). Prior knowledge, attitude, and strategy use in an introduction to statistics course. *Learning and Individual Differences*, *10*(4), 291-308. https://doi.org/10.1016/S1041-6080(99)80124-1

Scholtes, V. A., Terwee, C. B., & Poolman, R. W. (2011). What makes a measurement instrument valid and reliable?. *Injury*, *42*(3), 236-240. https://doi.org/10.1016/j.injury.2010.11.042

Shaughnessy, J. M. (1977). Misconceptions of probability: An experiment with a small-group, activity-based, model building approach to introductory probability at the college level. *Educational Studies in Mathematics,* 8, 285–316. https://doi.org/10.1007/BF00385927

Shaughnessy, J., Watson, J., Moritz, J., & Reading, C. (1999, April 22–24). *School mathematics students' acknowledgment of statistical variation* [Paper presentation]. 77th Annual Conference of the National Council of Teachers of Mathematics, San Francisco, CA, United States. https://hdl.handle.net/102.100.100/518171

Smith, S. W., Daunic, A. P., & Taylor, G. G. (2007). Treatment fidelity in applied educational research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education and Treatment of Children*, *30*(4), 121-134. https://dx.doi.org/10.1353/etc.2007.0033.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*(4), 753-766. https://doi.org/10.1111/j.1467-8624.1991.tb01567.x

Sproesser, U., Kuntze, S., & Engel, J. (2018). Using models and representations in statistical contexts. *Journal für Mathematik-Didaktik*, 39(2), 343-367. https://doi.org/10.1007/s13138-018-0133-4

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instrum. Comput. 31*, 137–149. https://doi.org/10.3758/BF03207704

Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest*, *12*(1), 3-54. https://doi.org/10.1177/1529100611418056

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, *12*(2), 147-169. https://doi.org/10.1007/BF03217081

Towne, L., & Hilton, M. (Eds.). (2004). *Implementing randomized field trials in education: Report of a workshop*. Committee on Research in Education, National Research Council. National Academy of Science. Washington, D.C.: The National Academies Press. https://doi.org/10.17226/10943

Trautwein, U., Golle, J., Jaggy, A. K., 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. https://doi.org/10.1111/nyas.15074

Torgerson, C. J., & Torgerson, D. J. (2013). *Randomised trials in education: An introductory handbook.* London: EEF.

Toulmin, S. (2003). *The Use of Argument* (2nd ed.). Cambridge: University Press.

Tremblay, P. F., Gardner, R. C., & Heipel, G. (2000). A model of the relationships among measures of affect, aptitude, and performance in introductory statistics. *Canadian Journal of Behavioural Science*, *32*(1), 40. https://psycnet.apa.org/buy/2000-13432-005

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105. https://doi.org/10.1037/h0031322

United Nations (2012). *Making Data Meaningful Part 4: A Guide to Improving Statistical Literacy*. Geneva: UNITED NATIONS. https://unece.org/fileadmin/DAM/stats/documents/writing/Making_Data_Meaningful_Part_4_for_Web.pdf

van Dijke-Droogers, M., Drijvers, P., & Bakker, A. (2024). Effects of a learning trajectory for statistical inference on 9th-grade students' statistical literacy. *Mathematics Education Research Journal*, 1-24. https://doi.org/10.1007/s13394-024-00487-z

Veckenstedt, R., Randjbar, S., Vitzthum, F., Hottenrott, B., Woodward, T. S., & Moritz, S. (2011). Incorrigibility, jumping to conclusions, and decision threshold in schizophrenia.

Cognitive Neuropsychiatry, *16*(2), 174-192. https://doi.org/10.1080/13546805.2010.536084

Vedejová, D., & Čavojová, V. (2022). Confirmation bias in information search, interpretation, and memory recall: Evidence from reasoning about four controversial topics. *Thinking & Reasoning*, *28*(1), 1-28. https://doi.org/10.1080/13546783.2021.1891967

Wallace, D., Rheinlander, K., Woloshin, S., & Schwartz, L. (2009). Quantitative literacy assessments: an introduction to testing tests. *Numeracy*, *2*(2), 3. http://dx.doi.org/10.5038/1936-4660.2.2.3

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, 88(421), 1-8. https://doi.org/10.1080/01621459.1993.10594283

Warren, P. A., Gostoli, U., Farmer, G. D., El-Deredy, W., & Hahn, U. (2018). A re-examination of "bias" in human randomness perception. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(5), 663. https://doi.org/10.1037/xhp0000462

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129-140. https://doi.org/10.1080/17470216008416717

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273-281. https://doi.org/10.1080/14640746808400161

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, New Jersey: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203053898

Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46. https://doi.org/10.52041/serj.v2i2.553

Watson, J. M. & Callingham, R. (2005). Statistical literacy: From idiosyncratic to critical thinking. In G. Burrill & M. Camden (Eds.), Curricular Development in Statistics Education. International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 2004 (pp. 116–162). Voorburg, The Netherlands: International Statistical Institute. iase-web.org/documents/papers/rt2004/4.1_Watson&Callingham.pdf

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*(1), 1-29. https://doi.org/10.1080/0020739021000018791

Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *The Journal of Mathematical Behavior*, *19*(1), 109-136. http://dx.doi.org/10.1016/S0732-3123(00)00039-0

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Yu, R. Q., Gunn, J., Osherson, D., & Zhao, J. (2018). The consistency of the subjective concept of randomness. *Quarterly Journal of Experimental Psychology*, *71*(4), 906-916. https://doi.org/10.1080/17470218.2017.1307428

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*(3), 319-328. https://doi.org/10.1111/j.2044-8279.1991.tb00989.x

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172-223. https://doi.org/10.1016/j.dr.2006.12.001

Declaration of Authors' Contribution

This publication-based dissertation includes three manuscripts that were written together with other authors. The contributions to the manuscripts are presented in the subsequent sections using the Contributor Roles Taxonomy (CRediT).

Chapter 2

Study 1: Evaluating the Efficacy of a Statistical Literacy Intervention

Author	Author position	Contributor Roles
Lucas Stark	First	Conceptualization, Investigation, Data Curation, Formal Analysis, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing
Jens Krummenauer	Second	Conceptualization, Resources, Writing – Review & Editing
Ann-Kathrin Jaggy	Third	Supervision, Writing – Review & Editing
Fabienne Kremer	Forth	Conceptualization, Resources, Project Administration, Writing – Review & Editing
Sebastian Kuntze	Fifth	Conceptualization, Resources, Writing - Review & Editing
Benjamin Nagengast	Sixth	Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing – Review & Editing
Ulrich Trautwein	Seventh	Conceptualization, Supervision, Funding Acquisition, Writing – Review & Editing
Jessika Golle	Eighth	Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing – Review & Editing

Chapter 3

Study 2: Assessing Decision Thresholds in Primary School Students Using Signal Detection Theory: Validating an Adapted Version of the Beads Task

Author	Author position	Contributor Roles
Lucas Stark	First	Conceptualization, Investigation, Data Curation, Formal Analysis, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing
Benjamin Goecke	Second	Formal Analysis, Methodology, Validation, Supervision, Writing – Review & Editing
Ann-Kathrin Jaggy	Third	Conceptualization, Supervision, Writing – Review & Editing
Jens Krummenauer	Forth	Conceptualization, Writing – Review & Editing
Sebastian Kuntze	Fifth	Conceptualization, Writing - Review & Editing
Jessika Golle	Sixth	Conceptualization, Methodology, Supervision, Funding Acquisition, Writing – Review & Editing
Benjamin Nagengast	Seventh	Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing – Review & Editing

Chapter 4

Study 3: Promoting Primary School Children's Statistical Literacy: Results of a Randomized Controlled Field Trial

Author	Author position	Contributor Roles
Lucas Stark	First	Conceptualization, Investigation, Data Curation, Formal Analysis, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing
Jens Krummenauer	Second	Conceptualization, Resources, Writing – Review & Editing
Ann-Kathrin Jaggy	Third	Supervision, Writing – Review & Editing
Fabienne Kremer	Forth	Conceptualization, Resources, Project Administration, Writing – Review & Editing
Sebastian Kuntze	Fifth	Conceptualization, Resources, Writing - Review & Editing
Jessika Golle	Sixth	Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing – Review & Editing
Benjamin Nagengast	Seventh	Conceptualization, Methodology, Validation, Supervision, Funding Acquisition, Writing – Review & Editing
Ulrich Trautwein	Eighth	Conceptualization, Supervision, Funding Acquisition, Writing – Review & Editing