Fostering Students' Epistemic Competences when Dealing with Scientific Literature

Die Förderung epistemischer Kompetenzen von Studierenden im Umgang mit wissenschaftlicher Literatur

Inaugural-Dissertation

zur Erlangung der Doktorwürde der

Fakultät der Humanwissenschaften

der Julius-Maximilians-Universität Würzburg

Vorgelegt von

Sarah von der Mühlen

aus Berlin

Würzburg 2018

Betreuer: Professor Dr. Tobias Richter

Erstgutachterin: Professor Dr. Gerhild Nieding

Zweitgutachter: Professor Dr. Wolfgang Lenhard

Tag des Kolloquiums:

“Education is not the learning of facts, but the training of the mind to think.”

Albert Einstein

For my mother

Table of Contents

Acknowledgments

I would like to thank many people that were directly or indirectly involved in my work over the last few years.

First and foremost, I would like to thank my primary supervisor, Prof. Dr. Tobias Richter, for his continuous support and encouragement during and after my time as a research associate at the University of Kassel. He has been incredibly helpful at every stage of my dissertation, and I very much appreciate his expertise, fairness, and constant availability, as well as his constructive feedback and the many ideas and discussions that encouraged me to think carefully about my work. I would also like to thank Prof. Richter for his patience, trust, and understanding in more challenging times. I also thank Prof. Dr. Gerhild Nieding and Prof. Dr. Wolfgang Lenhard for their time, effort, and interest in this dissertation.

Moreover, I would like to thank my other co-authors, Prof. Dr. Kirsten Berthold, Dr. Sebastian Schmid, and Elisabeth Marie Schmidt, for the close collaboration, many helpful and creative ideas for the development of the testing materials, productive meetings and discussions, and for their valuable feedback on the articles that are part of this dissertation. I also thank Dr. Kathrine Bruns for many fruitful ideas and discussions and her great support in the creation of the pilot study.

Many thanks go to the student assistants for helping with the construction and programming of the materials used in the reported studies, and with the data collection. I can't thank Panagiotis Karageorgos enough for his incredible support, enthusiasm, and dedication to the project. A very special thanks also goes to Anna Helfers, who has been enormously helpful with the construction of the materials for the argument structure training. I am also grateful to all the student assistants who helped with the testing. It has been a pleasure to work with these students. In addition, I would like to thank all the students and scientists who were willing to participate in our studies, despite the sometimes challenging and long lasting tasks.

Furthermore, I am very grateful to the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) for supporting the research presented in this dissertation, and to the coordinators of the Research Initiative *Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor (KoKoHs)* for stimulating workshops and meetings. I would also like to thank the reviewers who provided valuable feedback for the empirical studies reported in this dissertation. In a similar vain, I would like to thank everyone who showed interest in my work during numerous conference meetings and the helpful and encouraging feedback they provided.

I would also like to thank my former colleagues from the Department of Psychology at the University of Kassel, who were always very helpful with their scientific and methodological expertise or with practical matters. A special thanks goes to my former roommate Bettina Müller, who I could always talk to whenever I needed some advice, especially in the first phase of my dissertation, to Johanna Maier and Maj-Britt Isberner for useful advice, feedback, and discussions, and to Julia Schindler who has always been there to help with practical matters, even after I stopped working at the University of Kassel.

Moreover, I would like to thank my friend Christian Zohren for the proof reading of this dissertation, and all my friends who showed understanding for my limited time during the final phase of the dissertation.

Finally, I would like to give my dearest thanks to my parents and family, for their infinite love, appreciation, and support, including the vast number of babysitting hours and encouragement in more difficult times. Without the incredible support from my mother, Gabi, and the vast number of babysitting hours she was willing to take without ever complaining, this would not have been possible! My farther, Michael, has also been very supportive and I certainly owe some of my interest in thinking and reasoning to the many, often deep discussions we had at the dinner table throughout my life. Apart from my parents, I would like to thank my brother Manuel, who is often many miles away, yet always takes part in my

life. I am very grateful to my wonderful husband, Steven, who is the most patient and tolerant person I know, who has also been incredibly supportive, is always there for me and who loves me unconditionally. I also thank him for the proof reading of this dissertation. Last, but not least, I would like to thank my son, Oscar Samuel, who brings so much happiness into my life, and whose birth has broadened my mind in many ways.

Abstract

The abilities to comprehend and critically evaluate scientific texts and the various arguments stated in these texts are an important aspect of scientific literacy, but these competences are usually not formally taught to students. Previous research indicates that, although undergraduate students evaluate the claims and evidence they find in scientific documents to some extent, these evaluations usually fail to meet normative standards. In addition, students' use of source information for evaluation is often insufficient. The rise of the internet and the increased accessibility of information have yielded some additional challenges that highlight the importance of adequate training and instruction.

The aim of the present work was to further examine introductory students' competences to systematically and heuristically evaluate scientific information, to identify relevant strategies that are involved in a successful evaluation, and to use this knowledge to design appropriate interventions for fostering epistemic competences in university students. To this end, a number of computer-based studies, including both quantitative and qualitative data as well as experimental designs, were developed. The first two studies were designed to specify educational needs and to reveal helpful processing strategies that are required in different tasks and situations. Two expert-novice comparisons were developed, whereby the performance of German students of psychology (novices) was compared to the performance of scientists from the domain of psychology (experts) in a number of different tasks, such as systematic plausibility evaluations of informal arguments (Study 1) or heuristic evaluations of the credibility of multiple scientific documents (Study 2). A think-aloud procedure was used to identify specific strategies that were applied in both groups during task completion, and that possibly mediated performance differences between students and scientists. In addition, relationships between different strategies and between strategy use and relevant conceptual knowledge was examined. Based on the results of the expert-novice comparisons, an

intervention study, consisting of two training experiments, was constructed to foster some competences that proved to be particularly deficient in the comparisons (Study 3).

Study 1 examined introductory students' abilities to accurately judge the plausibility of informal arguments according to normative standards, to recognise common argumentation fallacies, and to identify different structural components of arguments. The results from Study 1 indicate that many students, compared to scientists, lack relevant knowledge about the structure of arguments, and that normatively accurate evaluations of their plausibility seem to be challenging in this group. Often, common argumentation fallacies were not identified correctly. Importantly, these deficits were partly mediated by differences in strategy use: It was especially difficult for students to pay sufficient attention to the relationship between argument components when forming their judgements. Moreover, they frequently relied on their intuition or opinion as a criterion for evaluation, whereas scientists predominantly determined quality of arguments based on their internal consistency.

In addition to students' evaluation of the plausibility of informal arguments, Study 2 examined introductory students' competences to evaluate the credibility of multiple scientific texts, and to use source characteristics for evaluation. The results show that students struggled not only to judge the plausibility of arguments correctly, but also to heuristically judge the credibility of science texts, and these deficits were fully mediated by their insufficient use of source information. In contrast, scientists were able to apply different strategies in a flexible manner. When the conditions for evaluation did not allow systematic processing (i.e. time limit), they primarily used source characteristics for their evaluations. However, when systematic evaluations were possible (i.e. no time limit), they used more sophisticated normative criteria for their evaluations, such as paying attention to the internal consistency of arguments (cf. Study 1). Results also showed that students, in contrast to experts, lacked relevant knowledge about different publication types, and this was related to their ability to correctly determine document credibility. The results from the expert-novice comparisons

also suggest that the competences assessed in both tasks might develop as a result of a more fundamental form of scientific literacy and discipline expertise. Performances in all tasks were positively related.

On the basis of these results, two training experiments were developed that aimed at fostering university students' competences to understand and evaluate informal arguments (Study 3). Experiment 1 describes an intervention approach in which students were familiarised with the formal structure of arguments based on Toulmin's (1958) argumentation model. The performance of the experimental group to identify the structural components of this model was compared to the performance of a control group in which speed reading skills were practiced, using a pre-post-follow-up design. Results show that the training was successful for improving the comprehension of more complex arguments and relational aspects between key components in the posttest, compared to the control group. Moreover, an interaction effect was found with study performance. High achieving students with above average grades profited the most from the training intervention. Experiment 2 showed that training in plausibility, normative criteria of argument evaluation, and argumentation fallacies improved students' abilities to evaluate the plausibility of arguments and, in addition, their competences to recognise structural components of arguments, compared to a speed-reading control group. These results have important implications for education and practice, which will be discussed in detail in this dissertation.

Zusammenfassung

Die Fähigkeit, wissenschaftliche Texte und die darin enthaltenen Argumente zu verstehen und kritisch zu beurteilen, ist ein zentraler Aspekt wissenschaftlicher Grundbildung, wird jedoch in der Schule kaum vermittelt. Obwohl Studierende die Behauptungen und Befunde, denen sie in der wissenschaftlichen Literatur begegnen, zu einem gewissen Grad kritisch bewerten, zeigen verschiedene Forschungsergebnisse, dass sie dies nicht in ausreichendem Maße tun und diese Evaluationen oft nicht den normativen Standards entsprechen. Darüber hinaus nutzen Studierende Quellenmerkmale nur unzureichend zur Beurteilung. Die Entstehung des Internets und die damit verbundene zunehmende Verfügbarkeit von Informationen stellen uns zudem vor einige wichtige Herausforderungen im Umgang mit diversen Informationsquellen und unterstreichen die Relevanz entsprechender Trainings und Förderungsprogramme.

Ziel der vorliegenden Arbeit war es, die Kompetenzen beginnender Studierender, wissenschaftliche Informationen heuristisch und systematisch zu bewerten sowie wesentliche Strategien, die für eine erfolgreiche Beurteilung wissenschaftlicher Informationen benötigt werden, weiter zu erforschen und auf dieser Grundlage Interventionen zu entwickeln, um diese Kompetenzen bei Universitätsstudierenden gezielt zu fördern. Dazu wurden mehrere computergestützte Studien entwickelt, die sowohl qualitative, als auch quantitative Daten, sowie experimentelle Untersuchungsdesigns beinhalten. Die ersten beiden Studien wurden konzipiert, um Förderbedarf gezielt zu ermitteln und Verarbeitungsstrategien zu identifizieren, die in verschiedenen Aufgaben und unter verschiedenen Bedingungen hilfreich sind. Dazu wurden zunächst zwei Experten-Novizen-Vergleiche entwickelt, in denen die Leistungen von deutschen Psychologiestudierenden (Noviz(inn)en) in einer Reihe unterschiedlicher Aufgaben, z.B. bei der systematischen Bewertung der Plausibilität informeller Argumente (Studie 1) oder der heuristischen Bewertung der Glaubwürdigkeit multipler wissenschaftlicher Texte (Studie 2), mit den Leistungen von Wissenschaftler(inn)en

aus dem Bereich der Psychologie (Expert(inn)en) verglichen wurden. Die Verwendung von Protokollen lauten Denkens diente dazu, die während der Aufgabenbearbeitung verwendeten Strategien, die die Leistungsunterschiede zwischen Studierenden und Wissenschaftler(inn)en möglicherweise mediieren, in beiden Gruppen genau zu erfassen. Darüber hinaus wurde untersucht, inwiefern unterschiedliche Strategien und die Nutzung bestimmter Strategien sowie relevantes konzeptuelles Wissen zusammenhängen. Basierend auf den Ergebnissen der Experten-Novizen-Vergleiche wurde anschließend eine Interventionsstudie, bestehend aus zwei Trainingsexperimenten, entwickelt, um einige Kompetenzen, die sich in den Vergleichen als besonders defizitär erwiesen hatten, gezielt zu fördern (Studie 3).

In Studie 1 wurde untersucht, inwiefern beginnende Studierende in der Lage sind, die Plausibilität informeller Argumente normativ angemessen zu beurteilen und gängige Argumentationsfehler zu erkennen, sowie verschiedene strukturelle Bestandteile von Argumenten zu identifizieren. Die Ergebnisse der Studie 1 legen nahe, dass es vielen Studierenden im Vergleich zu Wissenschaftler(inne)n an relevantem Wissen über die Struktur von Argumenten fehlt und die angemessene Bewertung ihrer Plausibilität für viele von ihnen eine große Herausforderung darstellt. Gängige Argumentationsfehler wurden häufig nicht richtig erkannt. Diese Leistungsunterschiede wurden teilweise durch eine unterschiedliche Strategienutzung mediiert: Studierende zeigten große Schwierigkeit darin, Beziehungen zwischen Argumentbestandteilen ausreichend Beachtung zu schenken. Darüberhinaus verließen sie sich bei der Beurteilung häufig auf ihre Intuition oder Meinung zum Textinhalt, während Wissenschaftler(innen) die Qualität der Argumente in erster Linie auf der Grundlage ihrer internen Konsistenz beurteilten.

Neben Plausibilitätsbeurteilungen informeller Argumente untersuchte Studie 2 die Kompetenz beginnender Studierender, die Glaubwürdigkeit multipler wissenschaftlicher Texte angemessen zu beurteilen und dabei auch Quellenmerkmale zur Beurteilung heranzuziehen. Die Ergebnisse zeigen, dass es Studierenden nicht nur schwerfiel, die

Plausibilität von Argumenten angemessen zu beurteilen, sondern auch die Glaubwürdigkeit wissenschaftlicher Texte heuristisch zu bewerten. Die Defizite auf Studierendenseite wurden dabei vollständig durch eine unzureichende Nutzung von Quellenmerkmalen mediiert. Wissenschaftler(innen) waren dagegen in der Lage, Strategien zur Beurteilung flexibel zu nutzen. Wenn eine systematische Verarbeitung nicht möglich war (Zeitlimit), griffen sie vor allem auf Quellenmerkmale zurück. Wenn eine systematische Verarbeitung jedoch möglich war (kein Zeitlimit), nutzten sie komplexere normative Kriterien zur Beurteilung, wie etwa die Bewertung der internen Konsistenz der Argumente (Vgl. Studie 1). Die Ergebnisse zeigen außerdem, dass es Studierenden an relevantem Wissen über verschiedene Publikationsarten fehlte und diese Schwierigkeiten waren korreliert mit der Fähigkeit, die Glaubwürdigkeit von Texten angemessen zu beurteilen. Die Befunde der Experten-Novizen-Vergleiche liefern zudem Hinweise darauf, dass sich die in den unterschiedlichen Aufgaben erfassten Kompetenzen auf der Basis einer allgemeineren Form der wissenschaftlichen Grundbildung und disziplinären Expertise entwickeln könnten. Die Leistungen in unterschiedlichen Aufgaben waren positiv korreliert.

Auf der Grundlage dieser Ergebnisse wurden zwei Trainingsexperimente entwickelt, um die Kompetenzen Studierender in Bezug auf das Verständnis und die kompetente Bewertung informeller Argumente, gezielt zu fördern (Studie 3). Experiment 1 beschreibt einen möglichen Interventionsansatz, um Studierende, basierend auf Toulmins (1958) Argumentationsmodell, besser mit der Struktur von Argumenten vertraut zu machen. Die Leistungen der Versuchsgruppe, verschiedene Argumentbestandteile dieses Modells korrekt zu identifizieren, wurden dabei in einem Prä-Post-Follow-up Design mit den Leistungen einer Kontrollgruppe verglichen, in der die Fähigkeit des schnellen Lesens trainiert wurde. Die Ergebnisse zeigen, dass das Training vor allem für das Verständnis komplexer und weniger typischer Argumente hilfreich war und Elemente, die die Beziehung zwischen verschiedenen Bestandteilen deutlich machten, im Posttest besser verstanden wurden als in einer

Kontrollgruppe. Darüber hinaus konnte ein Interaktionseffekt mit der Studienleistung gezeigt werden. Besonders „gute“ Studierende mit hohen Durchschnittsnoten konnten am meisten von diesem Training profitieren. Die Ergebnisse von Experiment 2 zeigten, dass ein Training, in dem das Konzept der Plausibilität, normative Kriterien der Argumentbewertung, sowie Argumentationsfehler vermittelt wurden, die Kompetenzen Studierender, die Plausibilität informeller Argumente normativ angemessen zu beurteilen, im Vergleich mit einer Kontrollgruppe, deutlich verbessern konnte. Die Ergebnisse der genannten Studien liefern wichtige Implikationen für die wissenschaftliche Praxis an den Hochschulen, welche in dieser Arbeit ausführlich diskutiert werden.

Chapter I

Introduction

Chapter I: Introduction

> "Research shows that good students study, with or without an exam, and bad students do not study, with or without an exam. Therefore, exams are unnecessary."

What do you think about the quality of the argument stated above? Regardless of your opinion about exams, is this a strong argument in the sense that the stated reasons provide sufficient support for the conclusion that exams are expendable? What about the average student? We do not know anything about the relationship between studying and the presence of an exam in this group. The author obviously has not considered all relevant options and, therefore, commits the fallacy of false dichotomy (or false dilemma). The example illustrates how people sometimes find it hard to deal with scientific evidence. The ability to think critically about scientific claims and evidence requires various epistemic competences and people, including university students, are not always very accurate at evaluating the information they encounter. However, appropriate training and instruction can help to improve this skill. An attempt to demonstrate this was undertaken in this dissertation.

Thinking and Reasoning in a Modern World: The Challenges Facing our Universities

Good thinking and reasoning skills are essential to success in virtually all areas of our life and include important decisions, such as considering the right career or having children, or day-to-day decisions, such as purchasing a car or booking a holiday. If we want to develop a justified point of view about a certain political, environmental, or societal topic (e.g., immunisation, global warming, or digital media consumption), it is important to be able to understand and evaluate different claims and evidence. People who do not adequately consider evidence or fail to link such evidence to the claims being made are susceptible to heteronomy and persuasion attempts by charlatans or conglomerates of interest. Reflective thinkers, on the

other hand, have more autonomy, freedom, independence, and flexibility when encountering new and unfamiliar situations, and are more resistant to these persuasion attempts (Petty, Haugtvedt, & Smith, 1995).

The ability to think critically about information can be traced back to the early philosophies introduced by Plato and Aristotle, and it has become even more important in a highly complex and rapidly changing technological world (Voss, Perkins, & Segal, 1991). Over the next decades, smart machines and computers will replace humans in many domains, forcing us to confront ourselves with the question of what humans are good at but machines are not. According to the University of Phoenix Research Institute (UPRI), computers will take over many routine or manufacturing jobs, but one thing that they are not capable of is the ability to reason about complex issues that require higher-order, flexible cognitive processes, such as insight or critical thinking (UPRI, 2011). The UPRI (2011) reports employment growth for jobs that require a high level of proficiency in reasoning and encourages educational institutions to place special emphasis on fostering students' reasoning skills. Many professional trainings already include tests for argumentation skills in their entrance exams (e.g., SAT, LSAT; Larson, Britt, & Kurby, 2009).

In the academic domain, the ability to reason about the claims and evidence presented in various scientific documents forms a crucial aspect of scientific literacy (Britt, Richter, & Rouet, 2014). Analysing informal arguments is considered a central skill across domains that students should learn (Kuhn, 1991; Kuhn & Udell, 2003), as it encourages critical reflection and conceptual understanding (De Vries, Lund, & Baker, 2002; Duschl & Osborne, 2002). Environments that encourage students to question and evaluate their own and others' line of argumentation have been shown to foster knowledge construction, understanding of a topic, and important metacognitive abilities (e.g., Cobb, 1994; Kelly & Crawford, 1997).

Despite the relevance of critical thinking, however, this competence is often not formally taught to students. In high school, students usually study textbooks, in which scientific

information is typically presented in the form of highly structured descriptions of isolated, certain facts (Penney, Norris, Phillips, & Clark, 2003), whereby underlying relationships and evidence for claims are often neglected (Luke, de Castell, & Luke, 1989; Paxton, 1997). Educational programmes often explain scientific phenomena superficially and lack detailed evidence or controversial discussions (Duschl & Osborne, 2002). Controversies, however, are crucial in science, because they are an important part of how knowledge is derived in scientific discourse (Britt et al., 2014). They allow the reader to reason about different claims and evidence, consider different perspectives, or detect whether a claim lacks support.

When students enter university, they are suddenly confronted with a broad range of more complex scientific literature (e.g., empirical journal or review articles), including many abstract concepts and theories, descriptions of methodological procedures, technical vocabulary, and an impersonal writing style (Fang, 2008). Unlike textbooks, primary scientific texts have a canonical structure (i.e., abstract, introduction, methods, results, discussion; Suppe, 1998) in which scientific topics are discussed in much more detail. Moreover, these texts represent uncertain aspects of science and controversies are omnipresent (Swales, 1990). Given that students are usually not prepared for dealing competently with such documents and the (sometimes conflicting) scientific claims and evidence presented in these documents, it is hardly surprising that many of them find this sudden exposure challenging. When learning about a scientific topic, students not only need to establish a profound understanding of multiple texts (Goldman et al., 2011), but also compare and integrate information from multiple texts (Bråten, Britt, Strømsø, & Rouet, 2011; Perfetti, Rouet, & Britt, 1999). Moreover, they need to critically evaluate what they read (Britt et al., 2014) in order to form a deeper understanding and justified point of view (Richter, 2011).

The rise of the World-Wide Web has made students' learning environments even more complex (Bromme & Goldman, 2014; Goldman & van Oostendorp, 1999). Evidence shows that students use the internet increasingly for educational purposes (Flanagin & Metzger,

2001), even as much as they read academic journals (Metzger, Flanagin, & Zwarun, 2003). The internet provides a tempting selection of a nearly infinite number of different documents, such as popular news or magazine articles, official publications, Wikipedia excerpts, science blogs, (self-published) e-books, or self-help forums, which can be found quickly by powerful search-engines. Although the accessibility of knowledge is generally a blessing, for it enables spreading of scientific knowledge among the general population, it has imposed additional cognitive demands (Britt et al., 2014). Imagine, for example, a student studying geology who has acquired some basic knowledge in this domain and now takes a course about the causes of global warming. This student might consult the internet to learn more about this topic for a course assignment. He or she might use a search engine to look for relevant information sources and these engines will present a long list of different websites containing different information about this topic. The student needs to choose a number of websites for further reading from a large list of information sources. Whereas some websites will state that global warming has mainly natural origins, others will argue that it is primarily human-made, and some may even deny its existence. The student not only needs to understand and integrate information from different information sources, but also critically evaluate the credibility and plausibility of the claims and evidence presented in different information sources, in order to form an accurate understanding of the topic.

One of the most serious challenges when dealing with scientific information on the World-Wide Web might be the exposure to inaccurate, implausible, or biased information. For example, consider the following argument about global warming that has been proposed by Dr. Roy Spencer from the University of Alabama in a publically available electronic paper for the Texas Public Policy Institute (Spencer, 2016, p.2):

> "Whether we use thermometers, weather balloons, or Earth-orbiting satellites, the measurements must be adjusted for known sources of error. This is difficult if not impossible to do accurately. As a result, different scientists come up with different global warming trends—

> or no warming trend at all. So, it should come as no surprise that the science of global warming is not quite as certain as the media and politicians make it out to be."

In this argument, the author states that measurements of the weather can never be entirely accurate and uses this information to justify his critical attitude towards common beliefs about the (manmade) causes of global warming. Although Mr. Spencer's argument should, of course, be judged by its own merits, it is important to note that the Texas Public Policy Institute seems to have received substantial funding, including money from the tobacco and fossil fuel industries (e.g., Exxon Mobile Corporation, 2008), a possible bias that should be considered by the student.

Several studies have shown that information sources that can be found on the internet vary significantly in their quality (e.g., Allen, Burke, Welch, & Rieseberg, 1999; Chung, Oden, Joyner, Sims, & Moon, 2012; Holtzman et al., 2005). For example, Allen et al. (1999) investigated the reliability of science information on the internet using search engines to determine the quality of information about controversial scientific topics, such as evolution or GM technology. The authors found that about one-fifth of the documents were inaccurate, about one-third were interpretatively misleading or biased, and more than three-fourths were not referenced. Many sites presented opinions or social commentary rather than factual information.

One of the reasons for this problem may be that information that can be found on the internet is rarely verified (Britt & Gabrys, 2002; Flanagin & Metzger, 2008). Authors who want to publish their work in peer-reviewed journals generally need to justify different claims and theories by providing specific (empirical) evidence to other scientists from the same domain. The majority of documents that can be found on the internet, however, lack such quality controls (Flanagin & Metzger, 2008). These documents are often published by journalists or lay people rather than scientists (Nwogu, 1991) and their authors do not need to

justify their claims before they are allowed to publish (Goldman & Bisanz, 2002). When scientific information is published in the media, such as popular science journals, popular science books, or online newspapers, it is published with the goal of raising public awareness of a scientific issue rather than understanding, which is reflected in simplification, less detailed information (e.g., about research methods), and fewer critical reflections or alternative viewpoints (Goldman & Bisanz, 2002; Nwogu, 1991; Secko, Amend, & Friday, 2013; Zimmermann, Bisanz, Bisanz, Klein, & Klein, 2001). The information that *is* published is often inaccurate (Holtzman et al., 2005), sensationalised (Ransohoff & Ransohoff, 2000), or influenced by selection and framing (Jarman & McClune, 2010), such as exaggeration (e.g., Mountcastle-Shah et al., 2003; Tabak, 2016). In addition, many media-co-operations are owned by or receive funding from companies, which may increase the tendency to publish arguments that serve the interests of these companies, whereas disconfirming evidence is neglected (Tittle, 2011).

Thus, a student who is interested in learning about a scientific topic, needs to actively evaluate the quality of the information he or she encounters, including both aspects of its content as well as characteristics of its source (e.g., author and publisher information). Given that false beliefs are relatively stable (e.g., Anderson, Lepper, & Ross, 1980; Johnson & Seifert, 1994), it seems important for readers to take a critical stance towards the information they encounter.

How do University Students Deal with Scientific Information?

Given that the competence to reason about scientific issues is usually not formally taught in secondary education, the question arises whether university students are able to deal competently with scientific information, form competent judgements about their quality, and discriminate between credible information sources, containing accurate, trustworthy and plausible information, and those containing inaccurate, biased, or interpretatively misleading information. Perkins (1985) was one of the first to notice that educational practices do rather

little to encourage students' reasoning skills. Although such skills have gained more attention in recent years, evidence from large-scale assessments, such as the Programme for International Student Assessment (PISA), or, in the Unites States, The National Assessment of Educational Progress (NAEP) and the National Center for Education Statistics (NCES), indicate that a large number of students still leave high school with insufficient skills to understand and evaluate scientific texts and arguments (e.g., NAEP, 1996, 1998; NCES, 2010; OECD, 2011, 2014). Results from recent PISA assessments revealed that German students score somewhat higher than the OECD average for both reading literacy and scientific literacy (OECD, 2014). However, although the majority of students were able to use basic scientific knowledge to identify a valid conclusion or scientific evidence for a claim, only a small number of them were able to identify and evaluate more complex arguments, consider alternative views or limitations, link different knowledge, or discriminate between relevant and irrelevant information. Students seem to have particular difficulty to determine whether a stated reason supports a claim or whether a reason is unsupportive or irrelevant to a claim (see also NCES, 2010), and very few are able to evaluate the reliability of sources (OECD, 2014).

Although the number of empirical studies that have examined students' competences to reason about scientific texts is still rather limited, findings from existing studies seem to confirm these results. For example, many students lack sufficient skills to fully understand more complex scientific arguments (e.g., Larson, Britt, & Larson, 2004), and to use more sophisticated criteria for evaluation, such as the relevance of evidence for a claim (e.g., Britt, Kurby, Dandotkar, & Wolfe, 2008; Larson et al., 2009; Larson et al., 2004; Shaw, 1996; Wolfe & Kurby, 2017). Moreover, university students seem to generally trust the information they find on the World-Wide Web (Flanagin & Metzger, 2000; Metzger et al., 2003) and rarely spontaneously attend to features of the document's source (e.g., Barzilai, Tzadok, & Eshet-Alkalai, 2015; Britt & Aglinskas, 2002; Goldman, Braasch, Wiley, Graesser, &

Brodowinska, 2012; Wiley et al., 2009). These challenges among students, which will be discussed in more detail in **Chapter 3**, highlight the need for explicit training and instruction (Kuhn, 1991; Norris, Phillips, & Korpan, 2003; Perkins, Farady, & Bushey, 1991; Phillips & Norris, 1999).

Overview

The central goal of this thesis was to further examine introductory students' competences to systematically and heuristically evaluate scientific information, to identify relevant strategies that are involved in their successful evaluation, and to use this knowledge to design appropriate interventions for fostering epistemic competences in university students. The strategies required for a successful comprehension and evaluation of scientific texts were examined by comparing the performance of scientists and students in a number of empirical studies. **Chapter 2** provides a theoretical overview of current views and theories about how readers understand and evaluate scientific texts and arguments. Furthermore, important requirements and normative criteria that are relevant in this context are described. **Chapter 3** provides a short review of studies that have been performed to assess students' competences to systematically and heuristically evaluate scientific texts and arguments. Moreover, some existing interventions for fostering students' critical thinking skills are reviewed. On this basis, questions and aims for the present research are formulated in **Chapter 4**. **Chapters 5** and **6** present the empirical work that was done to further explore introductory students' epistemic competences. **Chapter 7** presents empirical data from two intervention studies that aimed at fostering students' competences to understand and evaluate informal arguments. The findings and practical implications of the empirical studies reported in this dissertation are summarised and discussed in **Chapter 8**.

Chapter II

Theoretical Background

Chapter II: Theoretical Background

This chapter provides a theoretical overview of current theories about how readers deal with textual information and, in particular, the different arguments presented in these texts. To begin with, I will provide a definition of scientific literacy and explain why critical thinking and reading scientific documents are important requirements for developing scientific literacy. Subsequently, I will present an overview of different goals and strategies that can be applied during reading, and explain why epistemic strategies are particularly important when dealing with scientific texts. Next, I will give an outline of how knowledge is constructed and applied during reading according to modern theories in cognitive psychology and discourse processing, as these processes provide the basis for understanding how students deal with scientific literature. Furthermore, I will elaborate on the special role of informal arguments in scientific discourse. In addition, important requirements, normative criteria for evaluation, and useful settings for instruction will be described. Finally, I will explain why the strategies and heuristics employed by discipline experts for designing interventions are worth looking at, and why individual and contextual factors should also be considered.

What is Scientific Literacy?

If we talk about educating students and helping them to become scientifically literate, this requires a clarification of what this central goal really means. The Programme for International Student Assessment (PISA) defines *scientific literacy* as "[…] **the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen […], which requires the competencies to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically**." (OECD, 2016, p. 13). Similarly, Britt et al. (2014, p. 5) define scientific literacy as "**the ability of people to understand and critically evaluate scientific content in order to achieve their goals".** Although the definition of scientific literacy varies somewhat from author to author (Majetic

& Pellegrino, 2014), there is agreement that it involves conceptual knowledge and knowledge about common principles of scientific inquiry, and that it has less to do with specific school curricula, but with the real-life effects of education on democratic citizenship (Feinstein, 2011; Yore et al., 2007). Reflecting about scientific issues requires spontaneous adaptation to the requirements of different tasks and situations and is generally a purposeful, *goal-directed* activity (Britt et al., 2014; Facione, 1990; Linderholm, & van den Broek, 2002; McCrudden & Schraw, 2007; Rouet & Britt, 2011).

Critical Thinking as a Prerequisite for Scientific Literacy

Although there are various schools about critical thinking and although the ability to think critically, reflect, or evaluate describe slightly different concepts (see King & Kirchener, 1994), different perspectives share some common, underlying principles. I will use these terms interchangeably to refer to the **competence to reason about (textual) information that involves a *judgement* about what to believe and a decision that follows from that judgement** (cf. Ennis, 1985; Halpern, 1998; Yore & Ford, 2011), such as the decision to accept or reject a scientific claim. Critical thinking is essential for developing scientific literacy (Britt et al., 2014). In scientific discourse, people form judgements about ill-defined problems in controversial or uncertain situations that do not have a definite solution (Halpern, 1998; Voss & Means, 1991). Instead, the credibility and plausibility of stated information needs to be evaluated against other stated information, or against prior beliefs and assumptions held by the reader (Dewey, 1938). These judgements rely on normative criteria and they are not final, but open for (self-)criticism (Lipman, 1988).

Reading Literacy and Scientific Literacy

Critical reflections and evaluations are also essential aspects of reading (Norris & Phillips, 2003). The relationship between reading and problem solving has recently gained increased international attention (OECD, 2014, 2016). *Reading literacy* is defined as "**an individual's capacity to understand, use, and reflect on written texts, in order to achieve**

one's goals, to develop one's knowledge and potential, and to participate in society." (OECD, 2016, p. 13). To develop scientific literacy it is essential to read a substantial amount and variety of scientific documents (Norris & Phillips, 2003; Phillips & Norris, 2009). For example, Tenopir and King (2004) demonstrated that scientists spend on average about 25% of total work time on reading scientific documents. Norris and Phillips (2003, p. 226) wrote in their influential paper on scientific literacy:

> "Reading and writing are inextricably linked to the very nature and fabric of science, and, by extension, to learning science. Take them away and there goes science and proper science learning also, just as surely as removing observation, measurement, and experiment would destroy science and proper science learning."

Norris and Phillips (2003) distinguish between two literacy components: a discipline-specific literacy as the *fundamental sense*, and a content-specific literacy as the *derived sense*. They understand the competence to understand and evaluate scientific information as a mainly generic competence that is acquired relatively independent of a specific content domain and involves cross-disciplinary competences, such as the ability to apply knowledge, to reason, or to interpret and solve real-word problems. For example, Lehman and Nisbett (1990) found that college students who were taught inferential rule systems in different disciplines (i.e. social sciences, natural sciences, and humanities) were able to spontaneously apply these skills several months later with different topics in authentic settings (they were called at home). Halpern (1998) noted that knowledge needs to be represented by the student in a generic form, so that it can be applied in different situations. Thus, the competence to reason about scientific information is assumed to be constructed in a relatively fundamental, rather than derived way (see also Dole, Duffy, Roehler, & Pearson, 1991; Pearson, Roehler, Dole, & Duffy, 1992; Yore, Craig, & Maguire, 1998; Yore, Pimm, & Tuan, 2007). Notwithstanding, a person cannot fully understand scientific discourse without appropriate

genre and domain knowledge (Goldman & Bisanz, 2002). I will elaborate on the role of prior knowledge for the comprehension and evaluation of textual information later in this chapter. First, however, I will provide a short overview of different goals and strategies that can be involved during reading.

Receptive versus Epistemic Goals and Strategies for Dealing with Textual Information

When dealing with a text, successful readers flexibly apply a large number of general processing strategies, depending on the processing goal (Pressley, 2000; Rouet & Britt, 2011; Wyatt et al. 1993). The readers' goals strongly influence how they comprehend and evaluate textual information. Text comprehension researchers distinguish between receptive and epistemic goals. Whereas *receptive goals* focus mainly on accumulating and memorising textual information, *epistemic goals* aim at forming a valid understanding of a text that requires evaluation, and on developing a justified point of view about an issue (Richter, 2003, 2011; Richter & Schmid, 2010).

Different systematic and heuristic strategies serve different processing goals. *Systematic* strategies involve strategic and effortful thinking about the information presented in a text and depend on the readers' motivation and ability, whereas *heuristic* strategies involve rule-based processing of surface characteristics (cf. Petty & Wegner, 1999). For example, readers holding a receptive processing goal can systematically structure or organise textual information (Wild, 2000), or they can scan a text based on general heuristics to quickly locate relevant information (Bazerman, 1985). Readers holding an epistemic systematic processing goal carefully scrutinise the truthfulness or plausibility of the information stated in a text to determine the degree to which the stated evidence supports the main claim (e.g. Shaw, 1996). In psychology, *plausibility* can be defined as the ''**acceptability or likelihood of a situation or a sentence describing it**'' (Matsuki et al., 2011, p. 926). In contrast, readers with an epistemic heuristic goal may form a quick (preliminary) judgement about the credibility of information presented

in a text based on general heuristics, such as paying attention to features of the source (e.g., Korpan, Bisanz, Bisanz, & Henderson, 1997; Zimmerman et al., 2001). *Credibility* (or trustworthiness) reflects **the degree to which a source provides accurate, reliable information** (Petty & Wegener, 1999). Source features can involve document type, author, or publisher information, the publication date, or any other feature that defines by whom and under which circumstances the text content was created (Braten, Strømsø, & Britt, 2009; Britt & Aglinskas, 2002; Britt & Rouet, 2012; Scharrer & Salmerón, 2016). The term *sourcing* is used to refer to the **reader's explicit attention to source information *prior* to reading, and her use of source information for judging a document's trustworthiness** (Wineburg, 1991; Rouet, 2006). Traditional two-process models of information processing, such as the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986) or the Heuristic-Systematic Model (HSM; Chen & Chaiken, 1999), assume that strategy use is largely dependent on motivation and ability and that heuristic processes are applied only when recipients lack the motivation or the ability (i.e. general cognitive ability or prior knowledge) to process a message systematically. However, more recent research acknowledges that heuristics, such as paying attention to source features, can serve as important cues for credibility when systematic processes cannot be applied, for example, when relevant domain-specific prior knowledge is missing or when motivational and cognitive resources are limited (Bromme & Goldman 2014; Richter, Schroeder, & Wöhrmann, 2009; Schroeder, Richter, & Hoever, 2008). Although sources can also be evaluated systematically (Petty & Wegener, 1999), most information searches are time limited, and it is not possible to consider all aspects of a document for evaluation. Paying attention to the source and quickly applying sourcing heuristics can also help to initiate subsequent systematic evaluation processes (see Petty & Wegener, 1999 for a review). Useful systematic and heuristic epistemic strategies for evaluating scientific texts are described in the section "Normative Approaches for Evaluation".

Receptive strategies can be useful for memorising information from a text that is consistent, entirely accurate, and plausible (Richter, 2011). However, as pointed out in **Chapter 1**, when university students learn about a scientific topic by studying multiple scientific texts, the documents they encounter usually contain different and often inaccurate, biased or implausible information (e.g., Allen et al., 1999; Chung et al., 2012). Therefore, students will not be able to achieve a full understanding of a topic by purely applying receptive strategies when learning about a scientific issue, especially when they read (multiple) documents from the internet. For a successful understanding of a scientific issue, they need to evaluate different texts and, thus, apply epistemic strategies (Mayer, 1989; Richter, 2011). In addition, applying epistemic processing strategies can protect the reader against fallacious or biased arguments (Johnson, Smith- McLallen, Killeya, & Levin, 2004; Park, Levine, Kingsley Westerman, Orfgen, & Foregger, 2007).

Whereas both epistemic systematic and epistemic heuristic strategies involve a processing goal, evaluations can also involve automatic processes that are not based on specific goals. Such automatic processes are discussed next.

Non-strategic and Effortful Processing of Information: Epistemic Monitoring versus Epistemic Elaboration

Recent evidence using event-related potentials, reading times, or eye-tracking data suggests that epistemic processes can, in addition to slow, resource-demanding, and strategic validation processes (*epistemic elaboration*, Richter, 2003, 2011, 2015), involve routine, efficient, and non-strategic validation processes (*epistemic monitoring*, Richter et al., 2009; Richter, 2011, 2015, Richter & Maier, 2017) that occur as an early part of comprehension (e.g., Feretti, Singer, & Patterson, 2008; Isberner & Richter, 2013; Matsuki et al., 2011; Richter et al., 2009; Singer, 2006; see Isberner & Richter, 2014 for a review). For example, in a Stroop-like stimulus response compatibility task (cf. Stroop, 1935), Isberner and Richter

(2013) were able to show that students reading sentences of varying plausibility responded more slowly when they encountered implausible sentences, as compared to when the sentences were plausible, indicating that readers cannot ignore plausibility even when it is not relevant for a task. Such epistemic monitoring processes do not require activation of a learning goal, because they rely on information that is already part of working memory (e.g., information that is part of the current mental model), or on information that is easily accessible from long-term memory (Gerrig & McKoon, 1998; Richter & Singer, 2017). Perceived plausibility is used as a cue to automatically select and weigh information (Richter, 2015; Richter & Maier, 2017).

Under certain circumstances, however, readers actively elaborate on textual information, particularly if this information is inconsistent with prior beliefs (Richter, 2015). These epistemic elaboration processes are optional and depend on specific reading goals, such as the goal to develop a justified point of view (Richter, 2003, 2011; Schroeder et al., 2008). Such processes may be initiated when inconsistent information is detected, but they also require enough cognitive resources, motivation, and prior knowledge that can be used to actively elaborate on textual information (Richter, 2011, 2015). Readers can particularly profit from active elaboration, rather than simply monitoring of information, because elaborative processes facilitate understanding (Richter, 2011).

In sum, when processing textual information, readers can hold receptive or epistemic processing goals, and apply receptive or epistemic processing strategies. These strategies can be systematic or heuristic. Epistemic strategies are particularly relevant for dealing with scientific texts and arguments of varying quality. Epistemic processes can be both routine and effortless (epistemic monitoring), or strategic and effortful (epistemic elaboration). Whereas routine monitoring processes seem to occur spontaneously, elaborative processes require the activation of an epistemic learning goal, sufficient cognitive resources, and relevant prior

knowledge that can be used for evaluation and these processes are particularly relevant for fostering a deeper understanding of a topic.

To understand how goal-directed epistemic processes can help to create richer and more flexible mental models, it is important to understand how readers represent textual information, and how these representations guide the reader to comprehend and evaluate textual information.

How Readers Represent Textual Information

When people read a text, they construct different levels of representations that are hierarchically organised (Kintsch, 1988; Kintsch & Welsch, 1991; van Dijk & Kintsch, 1983). The lowest level (i.e. *linguistic level*) is a verbatim or surface representation of the text itself that includes words and grammar. The medium level (i.e. *semantic level* or *textbase*) is a representation of its propositional structure and includes local and global meaning units. The highest level of representation is a *situation model* (van Dijk & Kintsch, 1983) or *mental model* (Johnson-Laird, 1983; Henceforth the term mental model will be used). The mental model represents the state of affairs described in the text (Johnson-Laird, 1983; van Dijk & Kintsch, 1983) and integrates new information with previously stated information and with the readers' prior knowledge and is important to gain a deeper understanding of the information described in a text (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Coté & Goldman, 1999; Magliano & Millis 2003; Mayer, 1989; Shaw, 1996). Successful readers connect ideas in the text with relevant prior knowledge, explain these ideas, and actively evaluate these ideas as they attempt to comprehend the information stated in the text. Thus, mental models strongly depend on the readers' knowledge (Kintsch & Welsch, 1991). Representations that are formed during reading may be explicitly stated information, such as empirical evidence or the main claim, or inferred information, such as the relevance of stated evidence for the claim (Britt et al., 2014). Central elements of mental representations in scientific documents are non-human entities (e.g., vaccinations), states (e.g. inoculations contain immunising vaccines), and dynamic events (e.g., children are immunised or not immunised; Britt et al., 2014). The more entities, states and

events are provided in a text, the more representations need to be constructed by the reader. Ideally, a mental model fully captures the meaning of the described state of affairs by connecting it to relevant knowledge and related inferences (Kintsch & van Dijk, 1978). Whereas linguistic and semantic level representations decay quickly, mental models are relatively stable (Kintsch, Welsch, Schmalhofer, & Zimny, 1990). Mental models are the most interesting in psychological research, because they strongly depend on reader characteristics (i.e. prior knowledge) that influence how people comprehend and evaluate information.

Before I describe how readers use their mental models for evaluation in more detail, it is important to understand how people construct knowledge representations and how these representations can help the reader to understand, interpret, and draw inferences about the content of a text.

The Role of Prior Knowledge for the Comprehension of Scientific Texts: Construction Integration Model, Schema Theory, and Constructivist Approach

Several theories have tried to explain how readers construct and represent textual information, and how they rely on their prior knowledge to comprehend, interpret, or draw inferences about such information. Among the most influential approaches are the Construction-Integration Model, the Schema Theory, and the Constructivist Approach, which will be described briefly in the next paragraphs.

Construction-Integration Model

The *Construction-Integration (CI) Model* (Kintsch, 1988) is a bottom-up approach of text comprehension and describes how knowledge is constructed and integrated when readers encounter written information. In the first phase, (i.e. construction phase), a network representation of the text is constructed, whereby information stated in the current text is linked to previous discourse and (both relevant and irrelevant) knowledge is automatically and non-

strategically retrieved from long-term memory into working memory (*memory-based processing*; Gerrig & McKoon, 1998; O'Brien, Lorch, & Myers, 1998). Textual cues guide the reader in forming a mental model. In the second phase (i.e. integration phase), strongly related pieces of information that are particularly relevant for the topic are strengthened by a constraint-satisfaction mechanism, whereas irrelevant information is excluded from the representation, resulting in a consistent representation (Kintsch, 1988). The CI-Model has been very influential in text comprehension research and illustrates the importance of memory-related processes for understanding textual information. Although the integration phase has recently attracted more attention for explaining updating or validation processes as well (Richter & Singer, 2017), it was mainly designed for how readers comprehend or recall textual information (Graesser et al., 2007).

Schema Theory

Another major theory in text comprehension research, the *Schema Theory* (Anderson, 1984) is based on the assumption that readers use pre-existing mental schemes to guide and interpret textual information (cf. Bartlett, 1932). According to this top-down view, readers interpret textual information in light of pre-existing structural schemata that exist as relatively stable representations in long-term memory (Anderson, 1984). When activated, these schemes allow the reader to connect ideas from the text with these representations and draw inferences about the text, including inferences that facilitate the reconstruction of details that have been forgotten. As pre-existing schemes influence how readers interpret textual information, inaccurate or inflexible schemes can sometimes lead to biased processing (e.g., Anderson et al., 1980; Anderson, Spiro, & Anderson, 1978; Johnson & Seifert, 1994, see **Chapter 3**).

Constructivist Approach

The *Constructivist Approach* (Graesser, Singer, & Trabasso, 1994) explains how readers construct knowledge-based inferences. The constructivist approach is based on the assumption that readers try to construct meaning from a text (Bartlett, 1932). In cognitive

psychology, the extent to which a concept is meaningful depends on the number of connections it shares with other concepts in memory (Halpern, 1998). These concepts can all serve as retrieval cues. According to the constructivist theory, readers try to construct a meaning representation that addresses the reader's goals, that is both locally and globally coherent, and that explains why certain states or events are mentioned in the text (Graesser et al., 1994). Although the model was originally established to explain inference-making in narrative texts, it can be applied to science texts as well. For example, striving for coherency of information (e.g., resolving conflicts between conflicting claims; Rouet et al., 1999), goals (e.g., developing a justified point of view; Richter, 2003), and causal relations are all important aspects for how readers interpret scientific information.

Thus, readers use their prior knowledge for comprehension, interpretation, and inference-making. The approaches described above indicate that understanding of a text is usually an interaction between text and reader characteristics. However, these approaches often neglect the implication that readers also use their mental models to *evaluate* textual information (e.g., Isberner & Richter, 2013; Richter et al., 2009; Richter, 2015; Richter & Singer, 2017; Schroeder et al., 2008; Singer, 2006, 2013).

The Role of Prior Knowledge for Validation of Scientific Texts: Mental Model Theory

As mentioned ealier in this chapter, readers not only use their prior knowledge to form an understanding of a text, but also to continuously assess the accuracy or plausibility of the information stated in text (*epistemic validation*; Schroeder et al., 2008). The *Mental Model Theory* (Johnson-Laird, 1983, 1994) can account for this finding. For example, it has been shown that readers use their mental models to evaluate the validity of formal arguments (Johnson-Laird & Byrne, 1992) and the plausibility and informal arguments (Galotti, 1989; Perkins, 1986; Shaw, 1996) by generating knowledge-based inferences (e.g., Johnson-Laird & Byrne, 1992; Shaw, 1996). The Mental Model Theory implies that skilled readers imagine

different situations when dealing with textual information (Shaw, 1996). For example, to determine the quality of an argument, they imagine a situation in which a stated claim and reason are true, a situation in which one or both of them are false, a situation in which the reason does not support the claim, and a situation in which a counter-argument possibly discredits the claim (see section **Normative Approaches for Evaluation**).

In the next section, I will present another important model that is particularly relevant when readers deal with multiple, rather than single scientific texts, and that includes source information as a criterion for evaluation.

Understanding and Evaluating Multiple Texts: The Documents Model

When readers read about a scientific topic, they usually consult multiple texts (Bråten, Stadtler, & Salmerón, in press; Britt et al., 2014; Goldman et al., 2011). Across multiple documents, readers create a mental model that includes an integrated understanding of the information stated in different documents which is referred to as *Documents Model* (Britt, Perfetti, Sandak, & Rouet, 1999; Britt & Rouet, 2012; Britt, Rouet, & Braasch, 2013; Perfetti et al., 1999; Rouet, 2006). The Documents Model includes information about the text's content, the documents' sources (e.g., author or publication type), and relationships between the content of different documents, between sources and content, and between sources. It can be regarded as an extension of Kintsch's (1988) construction-integration model of single-text comprehension, to which it adds two layers: One of them, the *Situations model,* represents an integrated mental representation of the content in the different documents. The other, the *Intertext model,* represents source information and relationships between different documents. The Documents Model is particularly relevant in the presence of conflicting information (Bråten et al., 2011; Britt et al., 2013; Strømsø, Bråten, & Britt, 2010) and takes account of the finding that skilled readers use source information for evaluating the trustworthiness of a text (Bromme & Goldman, 2014).

Thus, readers use their prior knowledge for comprehension, interpretation, and inference-making, but also for validation of textual information. The Mental Model Theory can explain how representations are used for evaluation. The Documents Model adresses readers' comprehension and evaluation of multiple texts.

In the next section, I will describe the normative criteria that are required for a successful evaluation of scientific texts and arguments and present a number of useful strategies and heuristics that are relevant in different contexts.

Normative Approaches for Evaluation

When we talk about helping students to form accurate judgements about the quality of scientific information, we assume that they do not think as well as they might. This requires a clarification of the normative standards that provide the basis for evaluation.

Normative theories provide standards against which actual human performance can be compared (Hahn & Oaksford, 2007) and are closely related to the principle of rational analysis that aims to understand human behaviour as an approximation to an ideal behaviour (Anderson, 1991). Generally, good thinking requires the presence of relevant goals and (knowledge-based) inferences that can guide the reader in examining a problem and in forming a decision about what to believe (Baron, 1991). Richter (2011) stresses the importance of *cognitive flexibility* for evaluating textual information, which he defines as the skill to (spontaneously) adapt and restructure one's knowledge to deal with different demands from various learning materials, based on rational grounds (see also Spiro & Jehng, 1990). Thus, different epistemic strategies need to be applied in different situations. I will begin with a description of the strategies involved in the systematic evaluation of the content of informal arguments. In a second step, I will describe the relevance of sourcing as a heuristic for evaluating scientific texts.

Evaluating the Plausibility of Informal Arguments

Dealing with scientific literature typically means dealing with arguments. When scientists make a new discovery or present a new theory, they need to explain their findings to their community to convince them of their accuracy. In scientific discourse, an argument is **the attempt to persuade a reader that a scientific claim is true** (Britt et al., 2014). To achieve this, people need to provide evidence for their claims. In contrast to explanations, **arguments, per definition, require (factual, theoretical, or empirical) support** (Osborne & Patterson, 2011). If they did not need support, they would be self-evident.

Traditionally, cognitive psychologists have focused on formal reasoning, including well-defined, deductive arguments that have a single, verifiable conclusion that follows logically from the premises, if the argument is valid (Evans & Thompson, 2004). However, the great majority of arguments found in scientific documents are *informal*, rather than formal, in nature. Although informal reasoning in the domain of science *is* rational, in the sense that it is goal-directed, often highly systematic, and justifiable on pragmatic grounds (Anderson, 1991), the problems are typically more ill-structured and cannot be solved sorely by applying logical rules (Johnson-Laird & Byrne, 1992; Rips 1983; Walter, 1996).

The Structural Components of Informal Arguments. Arguments can be represented as an argument scheme (Britt & Larson, 2003; Chambliss, 1995; Chambliss & Murphy, 2002; Voss, 2005), whereby the claims holds the top position and activates relevant knowledge and beliefs about the topic before related reasons or reason-claim-connections are activated (Voss, Fincher-Kiefer, Wiley, & Silfies, 1993).

Like formal arguments, informal arguments consist of at least one or more *reasons* (or data) that support a main claim (Toulmin, 1958; Voss, 2005; Voss & Means, 1991). The *claim* is the main statement being argued for and is open for debate. Scientific claims can include various types of claims, such as policy claims (e.g., possession of weapons should be prohibited), value claims (e.g., it was the right decision to prohibit possession of weapons),

factual claims (e.g., weapons are dangerous), or causal claims that refer to an explanatory mechanism (Rottenberg, 1988). For example, consider the following statements:

1a. Nuclear power plants should be abolished.

1b. Nuclear power plants should be abolished, because nuclear power plants should be abolished.

1c. Nuclear power plants should be abolished, because they produce energy.

1d. Nuclear power plants should be abolished, because their waste poses a problem for the environment.

Statement 1a is an *unsupported assertion*, because it does not have a reason. Statement 1b is deductively valid, but, from an informal reasoning perspective, it is an *insufficient argument*, because it does not provide sufficient evidence for the claim. Statement 1c provides a reason (i.e., because they produce energy), but this reason does not support the claim. Such a statement is referred to as an *unwarranted argument*. Only statement 1d meets the requirements of a minimal argument (Toulmin, 1958), because it provides a reason (i.e., because their waste poses a problem for the environment) that is relevant for the claim.

Although the minimal requirement for an informal argument is the inclusion of at least one claim and one or more reasons, it may include additional components, such as warrants, backing evidence, and rebuttals (Toulmin, 1958). If not explicitly stated, *warrants* are underlying assumptions that the reader already holds and these assumptions allow the reader to determine the strength of the evidence for the main claim (Toulmin, 1958). In argument 1d stated above, the unstated warrant might be "things that pollute the environment should be prohibited". In everyday arguments or popular science texts, news briefs, or science blogs, warrants are often not explicitly stated, but need to be inferred by the reader (e.g., Chambliss, 1995; Greene, 1994). In scientific documents, however, it is important to state warrants explicitly and explain precisely why reported evidence justifies a certain conclusion.

Moreover, in scientific discourse, it is important to provide additional elaborations and evidence for why the evidence is relevant for the main claim (*backing,* Toulmin, 1958). For example, in our argument about nuclear power it should be explained why atomic waste is a thread to the environment. The claim can also be criticised generating *rebuttals* (Toulmin, 1958) or counter-arguments. For example, although the claim (i.e., nuclear power plants should be abolished) is supported with a relevant reason (i.e., because their waste poses a problem for the environment), one might argue that nuclear power plants should still be operated, because they produce less carbon dioxide than fossil fuels (which are another thread to the environment). Rebuttals, in turn, can be countered again with other arguments (e.g., arguments concerning the safety of nuclear power plants) and so forth. Figure 1 displays the components of the Toulmin (1958) model with a different example.

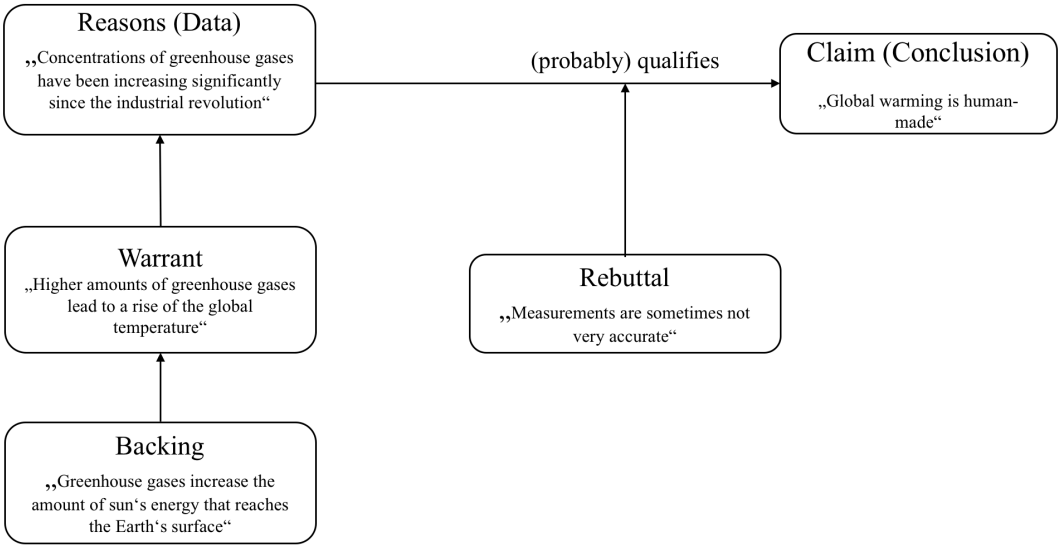

Figure 1. The structure of an informal argument according to the Toulmin (1958) model, illustrated with an example.

Scientific texts are often structured like full-fledged arguments. For example, Suppe (1998) examined more than 1000 data-based papers in science and found that papers from different disciplines had a common structure. They present data (i.e. reasons), show the relevance of the observations to a scientific problem, provide a detailed description of data

collection and analysis methods, and justify their claims and interpretations of the evidence. Finally, they identify, acknowledge, and generate alternative explanations.

Normative Criteria for the Evaluation of Informal Arguments. Toulmin (1958) proposed his model in reaction to the traditional formal reasoning perspective. However, it was only in the 1980s, when Blair and Johnson postulated their informal reasoning approach, that the discipline successfully established as a research field (van Eemeren et al., 1996). For a competent evaluation of informal arguments, Blair & Johnson (1987) name three important criteria. These include evaluations of the truthfulness of the information stated in the claim and reasons (*acceptability*), evaluations of whether the reasons support the claim (*relevance*), and evaluations of whether all relevant information has been considered, or, in other words, whether the reasons provide enough support for the claim (*sufficiency*). Similarly, Shaw (1996) proposed that when people evaluate the acceptability of arguments, these evaluations can be assertion-based, argument-based judgements, or alternative-based. *Assertion-based* judgements evaluate the truth of the claim or the data (cf. *acceptability*, Blair & Johnson, 1987), *argument-based* judgements consider whether the data provide relevant support for the claim (cf. *relevance*, Blair & Johnson, 1987), and *alternative-based* judgements focus on whether all information relevant for the truth of the claim has been considered (cf. *sufficiency*, Blair & Johnson, 1987). Several others have adopted this approach to determine the quality of an argument (e.g., Voss, Blais, Means, Greene, & Ahwesh, 1989; Voss & Means, 1991).

Baker (1985) distinguished between two standards against which readers evaluate the acceptability of a claim: an *external consistency* standard, whereby readers validate information against their prior knowledge and beliefs (e.g., accuracy of information), and an *internal consistency* standard, whereby they validate information against other information in the text (e.g. relevance of reason for the claim).

Whereas a formal argument is considered valid if the conclusion follows deductively from the premises, an informal argument is considered ***strong*****, if the conclusion *probably***

follows from the evidence (e.g., Ikuenobe, 2004; Johnson & Blair, 1977; Siegel & Biro; 1997). A strong argument that also has true premises is referred to as *cogent (*Johnson & Blair, 1977) or *sound* (Voss & Means, 1991). Thus, an ideal argument meets all the criteria described above. However, the strength of an argument can only be assessed by focussing on aspects of its internal consistency, such as relevance or sufficiency (Blair & Johnson, 1987). Although it is possible to routinely evaluate the truthfulness of information (Isberner & Richter, 2014), evaluations of argument strength require cognitive effort and can only be assessed strategically (Britt et al., 2014). Thus, if the reader wants to detect a weak argument based on normative evaluation criteria, it is essential to engage in systematic evaluation processes.

Competent readers use their prior knowledge about the rhetorical structure of arguments, along with conceptual and genre knowledge, to form accurate judgements about their quality (Britt et al., 2014; Britt & Larson, 2003; Halpern, 1998; Wolfe, Britt, & Butler, 2009). Activation of an argument scheme can provide important retrieval cues and helps the reader to construct coherent mental representations, understand relations between argument components, and shift their attention towards the internal consistency of the argument (Britt et al., 2014). Moreover, such knowledge can help to generate possible alternative explanations. Structural aspects can also facilitate knowledge transfer to unfamiliar situations (Halpern, 1998). It has been argued that the difficulty to interpret scientific documents often lies in failures to see its structure and the connections between different statements and arguments or argument components (e.g., Myers, 1991; Norris & Phillips 1994, 2009).

To be helpful, however, an argument scheme must be activated during reading, i.e. the reader needs to notice that he or she is dealing with an argument (Britt et al., 2014). Moreover, strategic epistemic elaboration processes need to be activated to allow the reader to actively evaluate the plausibility of an argument, attend to its internal consistency, and generate rebuttals, additional evidence, and alternative explanations, rather than simply monitoring the accuracy of information (Richter, 2011). When a person elaborates on a concept, many

meaningful connections are activated that can be retrieved and used for evaluation (Halpern, 1998). Activation of an argument scheme and relevant elaborative processes may be aided by including linguistic markers (e.g., Britt & Larson, 2003; Larson et al., 2004), or by including implausible information (e.g., argumentation fallacies) in a text (e.g., Braasch, Rouet, Vibert, & Britt, 2012; Richter, 2011; Richter & Maier, 2017; Staub, Rayner, Pollatsek, Hyönä, & Majewski, 2007, see **Chapter 3**).

Argumentation Fallacies. Richter (2011) argues that including implausible information in a text can help to initiative elaborative processes, which, in turn can help to detect an argument scheme, because it can provide cues that activate relevant knowledge from long-term memory (see also Richter & Maier, 2017). There is some disagreement among researchers concerning the precise nature and definition of fallacies, depending on the perspective (e.g., dialectical, rhetoric, informal, epistemic, see Slob, 2002 for a discussion). A wide-spread definition has been provided by Walton (1995) as **an instance of a general argument that does not meet a standard of correctness but that appears to be correct, and that poses a problem to the realisation of the reader's goal**. Other definitions also include the notion that, from a normative perspective, a fallacy should not be persuasive, but often persuades in actual practice (Hahn & Oaksford, 2007; Van Eemeren, Garssen, & Meuffels, 2009). Following Blair and Johnson's (1987) approach, any argument that violates one (or more) of the criteria of a cogent argument can be considered a fallacy, but usually the fallacies that occur in scientific discourse are arguments that violate the assumptions of relevance and sufficiency, because they might have true premises (and thus appear strong on the first glance), but lack (sufficient) support. Therefore, they conflict with the reader's goal to achieve a valid understanding of a topic. For example, consider the following statement (adapted from Thomas, 1991; Schroeder et al., 2008):

2. According to the balance theory, interaction partners are more likely to feel attracted to each other if they agree in their opinions about certain persons, issues, objects, or events, because their interpersonal liking is increased.

The sentence stated above is an excerpt from a scientific text about interpersonal attraction that was adapted for our training intervention (**Chapter 7**) to describe a *circular reasoning* fallacy (Dauer, 1989). In the example, the information stated in the reason (i.e. "because their interpersonal liking is increased") is basically a repetition of the information that was already stated in the claim (i.e. "[…] interaction partners are more likely to feel attracted to each other […]"). Although the author uses slightly different wording, the reason does not provide any independent evidence for the claim and there is no way to accept the reason unless the reader already believed in the claim. In other words, the claim lacks support. Such fallacies are very common in the literature students come across when reading about a scientific topic.

Other examples of fallacious arguments that are common in scientific discourse include *false conclusions* (Dauer, 1989), such as drawing causal conclusions from correlational evidence or misinterpreting research results. Moreover, different research results have different implied truth values (e.g., true, probable, uncertain, false) and different epistemic status or role (e.g., cause, effect, observation, hypothesis, method, motivation), which often results in failures to draw accurate conclusions (Norris & Phillips, 2009).

Frequently, it can be found that conclusions are drawn more broadly than they should be drawn (*overgeneralisation*, Dauer, 1989). Common examples include conclusions that are drawn from very few observations to a whole population, or conclusions that are drawn with too much certainty, thereby over representing the statistical evidence.

When authors of scientific texts falsely state that there are only two mutually exclusive options, whereas in fact, these options overlap, or other options exist that are neglected, they commit another common fallacy, the fallacy of *false dichotomy* (Dauer, 1989). For example,

the claim that some psychological trait is either inherited or it must be acquired misses the option that it can also be a combination of both.

Sometimes inappropriate examples are used to justify scientific claims (*wrong example*, Dauer, 1989). Meaningless examples do not provide sufficient evidence for a claim and, therefore, weaken the argument. There are many more examples of fallacious arguing, but the fallacies described above appear to be very prominent in scientific discourse (Dauer, 1989; Johnson & Blair, 1977). Good reasoners, however, should be able to detect such fallacies based on the principles described above (Blair & Johnson, 1987; cf. Weinstock, Neuman, & Tabak, 2004).

In sum, arguments require support and scientific texts are often structured like arguments. They not only need to provide evidence for a particular claim, but also explain precisely how findings have been established and why they justify a certain conclusion (warrant). Moreover, authors of such texts need to include detailed information and elaborations (backing), define the conditions and limits in which their theoretical claims apply, and consider alternative explanations and counterarguments (rebuttals). Understanding how an argument is structured is highly important, because this equips the reader with relevant knowledge for determining its quality. For a successful evaluation, readers need to evaluate the accuracy, relevance, and sufficiency of an argument, but its strength can only be assessed by paying attention to aspects of its internal consistency. Including linguistic markers and argumentation fallacies can be useful for stimulating elaborative processes that are needed for a competent evaluation.

In addition to evaluations of the content of a text, successful readers also need to pay attention to features of the source when evaluating their quality, especially when a situation does not allow systematic processing. The importance of source information for evaluation is described next.

Using Source Information for Evaluation

As mentioned in **Chapter 1**, different documents, especially those found on the internet, often vary in quality. Paying attention to sources can help with the construction of accurate mental models, because it directs the reader's attention towards more relevant and trustworthy information sources, especially when he or she is dealing with multiple documents (Bazerman, 1985; Bråten, Salmerón, & Strømsø, 2016; Bråten, Strømsø, & Salmerón, 2011; Goldman et al., 2012; Lundeberg, 1987; Rouet, Britt, Mason, & Perfetti, 1996; Strømsø & Bråten, 2014; Tabak, 2016; Wineburg, 1991). Sourcing can serve as an "anticipatory framework" for drawing inferences about the content of a text (Wineburg, 1991), whereby readers activate relevant textual schemata (Anderson et al., 1978) that can be used to form a first impression about the credibility of the document. Wineburg (1991) refers to this process as *sourcing heuristic*. This, again, can help readers to construct richer representations about the document (Bråten et al., 2009; Bråten et al., 2011; Rouet et al., 1999).

The sourcing heuristic can be useful for selecting relevant and reliable sources from a nearly endless list of documents on the internet (Bråten et al., 2011). For example, if an introductory student of Medicine wants to learn something about possible health effects of genetically modified food, it would be important to attend to whether the author of a text is qualified (e.g., is it a scientist in the field of Medicine, a Medical doctor, politician, a journalist, a student…?) and whether the author holds any conflicts of interest (e.g., did a researcher receive substantial funding from certain companies?). Moreover, the type of publication (e.g., peer-reviewed journal, blog entry, newspaper article) and the quality and motivation of the publisher should be examined. Importantly, basing one's judgements on document type has been considered an advanced strategy for credibility evaluations based on source (Braten et al., 2009; Rouet et al., 1996; Wineburg, 1991). In addition, searching for the presence of current and relevant references in the text can help to determine whether a source

is qualified. Finally, the student should check whether a document is current (or still relevant. Based on these criteria, the student can select a number of documents that seem trustworthy enough for further reading. In a second step, the claims and evidence presented in these documents can be examined more carefully. Table 1 provides an overview of epistemic strategies that can be used for the evaluation of informal arguments in scientific texts, including both systematic strategies for evaluating content and heuristics for evaluating a documents' source.

Table 1

Epistemic strategies for the evaluation of informal arguments in scientific documents

Step 1: Set epistemic goal (systematic or heuristic)	• e.g., form a justified point of view; disregard information that is unsupported or inaccurate • e.g., select sources that seem trustworthy for futher inspection
Step 2: Identify source features	• e.g., author /publisher information, date of publication, text genre, funding institution
Step 3: Evaluate trustworthiness of source	• Does the author have enough expertise? Is the stated information current (or relevant today)? Has the publication been peer-reviewed? Are there any biases or conflicts of interest?
Step 4: Identify argument structure	• Find linguistic markers (e.g., modals, connectors) • Identify claim, reason(s), (unstated) warrant, (unstated) backing evidence, (unstated) rebuttal(s)) • Identify inconsistencies/ contradictions / implausible information
Step 5: Assess	
• **acceptability**	• Check for truthfulness / accuracy of stated information

	• Consider counterarguments
• **relevance**	• Do the reasons support the stated claim? • Does the conclusion (likely) follow from the premises?
• **sufficiency**	• Do the reasons provide enough support the claim? • Seek additional evidence
Step 6: Evaluate Plausibility of Argument	• Does the conclusion (likely) follow from the premises? • Is there (unstated) information that does not support the claim and possibly discredits it? • Identify possible fallacies
Step 7: Form judgement	• Based on Step 6, accept or reject the argument

To conclude, in order to determine the quality of an argument and detect argumentation fallacies, readers not only need to evaluate the truthfulness of stated information, but also consider the relevance and sufficiency of stated evidence for the claim being argued for. In addition, readers should pay attention to features of the document's source to evaluate its the trustworthiness.

Scientists seem to apply the strategies and heuristics described above routinely when they deal with scientific texts. I will elaborate on how they might develop their expertise in the next section and explain why it might be worth looking at their strategy use when designing interventions to foster students' epistemic competences.

Scientists as "Discipline Experts"

Usually, experts are routinised at applying different strategies flexibly when trying to comprehend and evaluate scientific information (Wyatt et al., 1993). For example, scientists critically consider the contribution of research results and evaluate the methodology used to draw inferences about the acceptability of a claim, whereby they rely on prior domain and

topic knowledge as well as standards of disciplinary practice (Bazerman, 1985; Leinhardt & Young, 1996; Shanahan & Shanahan, 2008; Wineburg, 1998; Yore, Bisanz, & Hand, 2003). Moreover, although scientific thinking was traditionally thought of as a slow, iterative process, experts seem to make regular use of different heuristics. For example, they often use analogies to formulate hypotheses, solve unfamiliar problems, or explain unexpected research results (Dunbar, 2000). Furthermore, paying attention to source features is considered important for evaluation by experts in different disciplines (e.g., Bazerman, 1985; Chinn & Malhotra, 2002; Leinhardt & Young, 1996; Lundeberg, 1987; Shanahan & Shanahan, 2008; Shanahan, Shanahan, & Misischia, 2011; Wineburg, 1991, 1998; Wyatt et al., 1993). Rouet, Favart, Britt, and Perfetti (1997) suggest that an increase in domain knowledge and familiarity with different types of documents may lead to an increased awareness of source information. In addition, scientists work in an environment that cherishes controversies as a means for promoting understanding, which might implicitly provide them with relevant knowledge about normative criteria for evaluation, such as thinking of alternative explanations, rebuttals, or the relevance of stated evidence (Britt et al., 2014).

Interestingly, experts from different domains (e.g., scientists working in history, law, or psychology) seem to apply similar strategies when they evaluate scientific documents (Brand-Gruwel et al., 2017; Rouet & Britt, 2011; Rouet et al., 1997). Rouet et al. (1997) refer to such domain experts as *discipline experts* (or discipline specialists), whereby their expertise does not refer to the result of an extensive effort of deliberated practice (cf. Ericsson, Krampe, & Tesch-Römer, 1993), but to their amount of expertise in a particular field, such as the domain of psychology. Such knowledge may also familiarise them the structure of scientific texts and arguments (Suppe, 1998). Results from studies showing that scientists spend a great deal of their time reading (e.g., Tenopir & King, 2004) may provide some evidence for this line of thought.

Thus, if we want to help students to improve their ability to learn from scientific texts, it seems plausible to look more carefully at the strategies and heuristics used by domain experts, to encourage readers to acquire and adopt these strategies, and to design interventions accordingly. Constructive learning environments may be a promising setting for such interventions.

Using Constructive Leaning Environments for Instruction

Constructivist conceptions of learning assume that knowledge cannot be transmitted but is actively constructed by the learner and that instruction should allow learners to experience a problem in a *constructive learning environment* (*CLE*, Jonassen, 1999). It has been shown that students who actively construct their own information show deeper, conceptual understanding (e.g., Chi et al., 1989; Marsh, Edelman, & Bower, 2001; Mayer, 2001). To facilitate self-regulated learning and allow the construction of adequate mental models, authentic, relevant, and interesting problems should be displayed in an environment that enables the learner to interact with the materials, manipulate the content, receive feedback, and correct their response (Jonassen, 1999). Important information should be easily accessible. Feedback has been shown to increase motivation (Deci, 1971) and foster learning (Jonassen, 1999). Use of learning goals and examples can help to reduce cognitive load and help the learner to deeply process information during the practice phase (Renkl, 2009). Different examples or cases of a problem should be included to enable the formation of rich and flexible mental models and content should be organized and visualized in a way that is appealing to the learner (Jonassen, 1999). Experts can serve as *cognitive models* who present such examples (Jonassen, 1999; Renkl, 2009). Video tutorials are useful for reducing complexity, because they stimulate both visual and auditory channels (Mousavi, Low, & Sweller, 1995). Finally, instructional prompts, in which learners are required to self-explain stated information, have been shown to be particularly useful for the acquisition of knowledge

in CLEs, because they encourage active, deep processing of information (e.g., Berthold & Renkl, 2010).

In sum, CLEs use different elements, such as feedback, examples, learning goals, and prompts, to encourage active construction of knowledge by the reader. However, it is important to note that different individual and contextual variables can influence the amount of learning.

Individual and Contextual Factors: Aptitude-Treatment Interactions

Snow (1989) provided a detailed description of how individual differences and contextual factors, including text and task dimensions, can influence how readers deal with textual information. He argued that personal characteristics, such as prior knowledge, cognitive ability, learning styles, or thinking dispositions, always contribute to how a person deals with textual information and that interventions should be designed to match the characteristics of learners. So-called *aptitude-treatment interactions* (ATI; Cronbach & Snow, 1969; Snow, 1989) occur when different instructional treatments (i.e. *treatments*) result in different learning outcomes, depending on such personological variables (i.e. *aptitudes*). For example, whereas better learners may profit from constructivist settings, struggling readers seem to profit more from guided and more structured interventions (see Kalyuga, 2007 for a review). Thus, CLEs may be useful for some, but not all students.

Conclusion

The competence to comprehend and evaluate scientific information is an important aspect of scientific literacy. It is assumed that this competence depends on the activation of relevant goals and knowledge and develops in a relatively fundamental, rather than derived sense. Whereas receptive goals can be helpful for dealing with plausible texts, epistemic goals and strategies are highly relevant when dealing with controversial topics and documents of varying quality, such as scientific texts. The CI-Model, the Schema Theory, and the

Constructivist Approach can explain how readers apply knowledge to comprehend, interpret, and draw inferences about scientific texts. The Mental Model Theory accounts for the finding that readers also use their prior knowledge to validate scientific information. Epistemic processes can be routine and effortless (epistemic monitoring) or strategic and effortful (epistemic elaboration). To achieve an adequate understanding and justified point of view towards a scientific issue, readers need to activate elaborative processes. Including linguistic markers and fallacies in scientific documents can stimulate such processes and increase awareness of an argument scheme, which is needed for a competent evaluation of the quality of stated information. Students should use acceptability, relevance, and sufficiency as criteria for determining the plausibility of arguments and they should also consider features of the document's source for judging its trustworthiness. Sourcing is particularly important when systematic processes cannot be applied. Scientists seem to apply different strategies competently and looking at how these domain experts apply their knowledge can help to design appropriate interventions. CLE's that allow the learner to construct knowledge in an interactive environment, including feedback and examples, can be promising environments for developing such interventions. Finally, it is important to be aware of individual and contextual differences, which can influence whether learning is successful or not.

Chapter III

Literature Review

Chapter III: Literature Review

The present chapter reviews existing research of studies that have examined students' competences to evaluate the plausibility of scientific texts and arguments, and their ability to use source information for evaluation. Individual and contextual influences are described. In addition, existing interventions are reviewed.

Students' Competences to Evaluate the Plausibility of Informal Arguments

As described in **Chapter 2**, readers seem to routinely and non-strategically evaluate textual information, whereby they rely on prior knowledge to validate or monitor textual information (Isberner & Richter, 2014; Richter, 2011, 2015; Richter & Maier, 2017; Richter & Singer, 2017; Schroeder et al., 2008; Singer, 2013). Several lines of research show that such spontaneous evaluations sometimes result in biased processing of textual information (e.g., Anderson et al., 1980; Chinn & Brewer, 1993; Eagly & Chaiken, 1993; Johnson & Seifert, 1994; Maier & Richter, 2013a, 2013b; Norris et al., 2003; Schroeder et al., 2008; Wiley, 2005; Wolfe et al., 2009; Wolfe & Kurby, 2017). For example, students tend to rely on information that is consistent with their prior beliefs *(text-belief-consistency effect*, e.g., Eagly & Chaiken, 1993; Maier & Richter, 2013a, 2013b, Wiley, 2005), even when they encounter new information that outdates this information (*continued influence effect*, e.g., Anderson et al., 1980; Chinn & Brewer, 1993; Johnson & Seifert, 1994). They also tend to show a *plausibility bias* when they evaluate arguments in scientific texts (Maier & Richter, 2013b; Schroeder et al., 2008). For example, Schroeder et al. (2008) found that information was more likely to be judged as plausible, if it was consistent with the reader's mental model, irrespective of objective plausibility.

Students may be able to strategically control their evaluations and critically (re)consider stated claims and evidence to some extent. For example, students sometimes correctly distinguish strong from weak arguments (e.g., Hoeken, Timmers, & Schellens, 2012;

Hornikx & Hoeken, 2007; Hornikx & ter Haar, 2013; van Eemeren et al., 2009; see Johnson, Smith-McLallen, Killela, & Levin, 2004 for a review), whereby they seem to better at detecting some, but not all fallacies. For example, Ramasamy (2011) found that Malaysian undergraduates were fairly able to detect overgeneralisations, but not false analogies. Many studies, however, indicate that undergraduate students' evaluations of informal arguments often do not meet the criteria for normatively accurate evaluations (e.g., Britt et al., 2008; Larson et al., 2009; Larson et al., 2004; Manuel, 2002; Norris & Phillips, 1994; Norris et al., 2003; Perkins et al., 1991; Shaw, 1996, Wolfe et al., 2009; Wolfe & Kurby, 2017; Wu & Tsai, 2007). For example, Norris et al. (2003) examined whether university students from different disciplines were able to evaluate scientific media reports about current scientific topics. Although most students were able to read these reports without problems, many were unable to interpret their central findings correctly. In particular, they struggled with the distinction between explanation and description, confused cause and correlation, and tended to view all statements as equally true or justified.

Using Argument Schemes for Evaluation

High school and university students seem to have particular difficulty evaluating scientific arguments (Duschl & Osborne, 2002; Kelly, Druker, & Chen, 1998; Kuhn, 1991; Osborne, Erduran, & Simon, 2004; Shaw, 1996). Findings from several studies suggest that they may not be able to sufficiently activate relevant argument schemes that could be used for evaluation (Britt et al., 2008; Britt & Larson, 2003; Larson et al., 2004; Manuel, 2002; Norris & Phillips, 1994; Shaw, 1996). For example, Norris and Phillips (1994) and Manuel (2002) found that students struggled to identify central evidence for a claim. Similarly, Britt et al. (2008) showed that students often have difficulty precisely recalling argumentative claims. Moreover, although even younger students seem to use argument schemes to guide comprehension to some extent, if their structure is made explicit and the argument does not contain any misleading information (e.g., Chambliss, 1995; Chambliss & Murphy, 2002),

undergraduate students still seem to struggle with the recognition of more complex arguments (Larson et al., 2004; cf. Toulmin, 1958). For example, Larson et al. (2004) found that undergraduates identified only 30% of claims and reasons correctly when they were presented with complex argumentative texts. They often misidentified uncontroversial and unsupported statements, data, and even counter-arguments as the main claim, when the structure was untypical (e.g., when the rebuttal was stated first).

Students seem to have particular difficulty with adequately representing relations between argument components (Britt & Kurby, 2005; Larson et al., 2009; Larson et al., 2004; Shaw, 1996; Wolfe & Kurby, 2017; Wolfe et al., 2009). For example, Larson et al. (2004) found that many students in their study were not able to identify warrants. Britt and Kurby (2005) observed similar results when they explicitly asked undergraduate and graduate students to judge whether a stated reason supported a claim. Whereas most graduate students successfully rejected unwarranted arguments, undergraduate students were significantly less able to identify unwarranted arguments (97% vs. 68%, respectively).

Further evidence for the notion that students often seem to neglect the claim-reason connection was found in a study by Shaw (1996) who showed that undergraduate and graduate students were much more likely to object to the truth of the premises and the conclusions of an argument than to violations of its internal consistency. Similarly, Wolfe et al. (2009) found that college and high school students struggled to evaluate the internal consistency of two-clause (claim, reason) arguments. In addition, students also seem to have difficulty generating rebuttals (e.g., Wu and Tsai, 2007). These results indicate that students tend to focus on the acceptability of a claim, but neglect the criteria of relevance and sufficiency for evaluation (cf. Blair & Johnson, 1987).

Thus, although students seem to evaluate arguments in scientific discourse to some extent, students often seem to struggle with a normatively accurate evaluation of (more complex) scientific arguments in scientific discourse and often process information in a

biased fashion. Prior research indicates a lack of relevant knowledge about the structure of arguments. In particular, students struggle with attending to relational aspects between argument components, such as warrants.

Several lines of research also indicate that students not only fail to accurately evaluate the content of a text, but also do not pay sufficient attention to the document's source for evaluating its trustworthiness. Students' use of source-related criteria for evaluation is reviewed next.

Students' Use of Source Information for Evaluating Credibility

Although university undergraduates and even high school students are able to use source information for evaluation under optimal conditions (e.g., Geriets, Kammerer, & Werner, 2011; Perfetti, Britt, & Georgi, 1995; Stadtler & Bromme, 2007), a variety of research demonstrates that they are rarely unable to do so spontanously and without explicit instruction or training (e.g., Barzilai et al., 2015; Bråten, Strømsø, & Andreassen, 2016; Brem, Russels, & Weems, 2001; Britt & Aglinskas, 2002; Gerjets, Kammerer, & Werner, 2011; Goldman et al., 2012; Korpan et al., 1997; Stadtler & Bromme, 2007; Stahl, Hynd, Britton, McNish, & Bosquet, 1996; Wiley et al., 2009; Wineburg, 1991).

One of the first, and probably most influential, studies showing that students often neglect source information was conducted by Wineburg (1991), who compared expert historians and advanced high-school students' use of source information while working on multiple historical documents, using think-alouds. In this study, almost all historians (98%) carefully considered the source of each document, such as author, document type, or date of publication, before reading them, and they used this information to make inferences about the content. In contrast, only 31 % of students attended to features of the document's source. The finding that high school and undergraduate students often fail to sufficiently attend to source information and their inability to use this information for forming judgements about the credibility of a document has been corroborated by several other studies, mostly in the domain

of history (e.g., Britt & Aglinkas, 2002; Stahl et al., 1996), and in the domain of natural sciences (e.g., Barzilai et al., 2015; Bråten et al., 2016; Sanchez, Wiley, & Goldman, 2006).

Individual Differences and Contextual Variables

The degree to which students are able to comprehend and evaluate scientific texts and arguments seems to be influenced by several individual and contextual factors (see Barzilai & Strømsø, 2018, for a review). For example, differences in cognitive ability seem to contribute to an increased ability to identify argumentation fallacies (Weinstock, Neuman, & Glassner, 2006) and differences in working memory can influence how readers understand more complex argumentative texts (e.g., Just & Carpenter, 1992). Furthermore, poor readers are often less accurate in retrieving relevant information from memory (e.g., Johns, Matsuki, & van Dyke, 2015; Perfetti, 1985). In addition, strong prior beliefs about a topic and differences in people's argument schemes can reinforce biased information processing (e.g. Baron, 1995; Stanovich & West, 1997; Wolfe, 2012; Wolfe et al., 2009). Readers show even stronger biases for belief-consistent information when this information is self-relevant (e.g., Maier & Richter, 2013a). They also seem to be particularly susceptible to the continued influence of incorrect information if it has been integrated as a cause or outcome of a situation described in the text (e.g., Anderson et al., 1980; Johnson & Seifert, 1994), probably because causal information is more easily accessible from memory than non-causal information (O'Brien & Myers, 1987; Richter & Singer, 2017). However, if a text contains enough causal explanations, readers are less susceptible to this effect (Kendeou, Smith, & O'Brien, 2013). Mature epistemological beliefs, such as views of knowledge as complex, uncertain, and justified by various sources, are associated with more elaborative information processing (e.g., Mason & Scirica, 2006; Ricco, 2007; Weinstock, 2009). For example, Ricco (2007) found that mature epistemological beliefs were associated with a reduced number of informal reasoning fallacies. Generally, prior knowledge is considered a crucial factor for how readers understand and evaluate scientific arguments (Barzilai & Strømsø, 2018). More recently,

motivational and affective differences, such as topic interest (e.g., Bråten et al., 2014), self-beliefs (e.g., Wigfield et al., 2016), or need for cognition (e.g., Winter & Krämer, 2012) have gained increased attention. The ability to evaluate the quality of arguments also seems to be culture-specific to some extent (e.g., Hornikx & Hoeken, 2007; Hornikx & ter Haar, 2013; Wolfe, 2012; Wolfe et al., 2009). For example, Hornikx and ter Haar (2013) found that Dutch students, but not German students, were sensitive to the quality of statistical evidence. Gender differences in reading ability or topic-specific prior knowledge and interest may also influence performance, but relevant research is still lacking (Barzilai & Strømsø, 2018).

Students' use of source information also depends on several individual and contextual variables. For example, Goldman et al. (2012) found that better learners were more likely than poorer learners to evaluate the trustworthiness of reliable documents using source information when reading multiple documents about a complex scientific issue. Furthermore, prior knowledge seems to be an essential predictor of students' use of source information (e.g., Bråten et al., 2011; Rouet et al., 1996; Rouet et al., 1997; Strømsø et al., 2010). For example, Rouet et al. (1997) found that graduate students showed more source awareness than undergraduates. In a similar study, it was found that graduate students based their evaluations more often on document type, whereas undergraduate students who had only little experience with various types of documents based their ratings on the content of the texts (Rouet et al., 1996). In addition, people's familiarity with the topic can influence their sourcing activities (e.g., Bråten et al., 2011; McCrudden, Stenseth, Bråten, & Strømsø, 2016). For example, Bråten et al. (2011) showed that readers low in topic knowledge were more likely to trust less credible sources. Similiarly, students' prior beliefs, values, and motivations seem to play an important role for the extent to which they use source information for evaluation (e.g., Braasch, Bråten, Britt, Steffens, & Strømsø, 2014; Gottlieb & Wineburg, 2012; Strømsø et al., 2010; van Strien, Brand-Gruwel, & Boshuizen, 2014). For example, readers are more likely to pay attention to the source, if the information in a text is judged as implausible in light of their

prior beliefs about the topic (de Pereyra, Britt, Braasch, & Rouet, 2014). People's epistemological beliefs about knowledge and how it is derived seem to be particularly important and can strongly influence evaluation processes (e.g., Barzilai et al., 2015; Bråten, Ferguson, Strømsø, & Anmarkrud, 2014). Moreover, students' use of source characteristics seems to depend on conceptual factors, such as the types of documents presented (Rouet et al., 1996), or the salience of sources (Bråten et al., 2016). In addition, students' source awareness seems to increase with the amount of conflicting information (e.g., Braasch et al., 2012; Rouet, Le Bigot, de Pereyra, & Britt, in press). Individual and contextual factors may also interact. For example, Barzilai and Eseth-Alkalai (2015) showed that the presence of conflicting information between sources influenced students' source awareness, but this depended on their beliefs in the uncertainty of knowledge and the need to justify claims with evidence.

Thus, although there is some evidence indicating that students evaluate scientific information to some extent, these evaluations are often biased and do not always meet normative standards. In particular, students seem to struggle to activate accurate argument schemes and to use these schemes for evaluation. Moreover, students seem to base their evaluations about a document's trustworthiness mainly on content rather than features of the source. However, several individual and contextual factors can influence students' evaluations. Existing interventions to improve students' skills to reason about scientific texts and arguments are described next.

Improving Students' Competences to Evaluate Scientific Texts and Arguments

In recent years, a growing body of research has explored possibilities to improve students' reasoning skills in science education (e.g., Geddis, 1991; Kuhn, 1991; Means & Voss, 1996; Osborne et al., 2004). Researchers have identified a number of tools and

conditions that may promote critical thinking and reduce biased processing of information. For example, in line with the reasoning in the present work, inducing an epistemic, rather than a receptive goal, seems to be an effective means to reduce (but not eliminate) students' processing biases (e.g., Maier & Richter, 2013a, 2016; McCrudden & Sparks, 2014; Wiley & Voss, 1999), and to create more elaborated mental models (Richter, 2003, 2011; Richter & Maier, in press). Furthermore, Wiley and Voss (1999) found that students were able to write more comprehensive essays including more causal information when they were instructed to write an argumentative essay (i.e., an epistemic reading goal), compared to when they were instructed to write a summary or narrative. Epistemic goals can also direct students' attention towards source features (Stadtler, Scharrer, Skodzik, & Bromme, 2014).

With regard to university students' abilities to comprehend and evaluate informal arguments, there is only a very limited number of studies that have explicitly addressed this issue. One attempt to improve the ability to evaluate the quality of arguments based on teaching argument structure was made by Larson et al. (2009). In particular, the researchers developed a tutorial that included elements from CLEs (Jonassen, 1999), such as immediate feedback, example-based and active learning and, in addition, declarative knowledge about the structure of informal arguments. This approach was successful at reducing students' evaluations of the quality of (relatively simple) claim-reason arguments by shifting their attention towards the internal consistency of the arguments. Similarly, Larson et al. (2004) found that a short tutorial, in which key components of more complex arguments were explained, shifted students' attention towards relations between argument components when immediate feedback was provided. These results also indicate that it seems to be important to include feedback in interventions.

Linguistic markers that signal relations in a text can help to detect the presence an argument scheme (Britt et al., 2014). Readers establish coherence by attending to coherence relations stated in different information units in a text (Noordman, Vonk, & Kempff, 1992).

Britt and Larson (2003) were able to show that modals (e.g., "should") and uncertainty markers (e.g., "probably") could help readers to identify and evaluate a statement as a claim, compared to unmarked statements. Similarly, Larson et al. (2004) showed that providing explicit markers, such as causal connectors (e.g., "because", "but"), improved argument comprehension. Britt and Larson (2003) also demonstrated that arguments that contained markers were processed faster than arguments that did not provide these signals.

Familiarising students with normative criteria for evaluating informal arguments (cf. Blair & Johnson, 1987; Voss & Means, 1991) can also be helpful. For example, Weinstock et al. (2004) found that awareness of normative criteria for evaluating arguments predicted the ability to identify fallacious informal arguments. Furthermore, including implausible information in a text has been shown to encourage critical thinking in students (e.g., Braasch et al., 2012; Staub et al., 2007).

In addition, instructional prompts can be helpful for fostering knowledge about arguments (e.g., Hefter et al, 2014; Hefter et al., 2015). For example, Hefter et al. (2014) developed a short computer-based training intervention in which important principles of argumentation were demonstrated by expert models in short video-based examples, using a CLE (Jonassen, 1999). Knowledge was then prompted and had to be self-explained by the student. The intervention was effective for improving both students' declarative knowledge about arguments and their ability to evaluate these arguments.

Finally, refutations that include a plausible alternative that can be integrated into the reader's mental model along with sufficient causal information are sometimes necessary to change strong misconceptions that have been manifested as schema-like knowledge structures or flawed mental models (e.g., Braasch, Goldman, & Wiley, 2013; Kendeou et al., 2013).

Increasing Students' Source Awareness

Recently, an increasing number of interventions targeting sourcing skills have been established (see Bråten et al., in press, for a review). Moreover, previous research has

identified a number of conditions that may promote students' use of sources for evaluation. These include the inclusion of conflicting information (Braasch et al., 2012; Kammerer, Kalbfell, & Gerjets, 2016; Strømsø & Bråten, 2014), increasing the salience of sources (Bråten et al., 2016), highlighting source information (e.g., Le Bigot & Rouet, 2007), and reading multiple, rather than single texts (e.g., Britt & Aglinskas, 2002; Nokes, Dole, & Hacker, 2007). For example, Braasch et al. (2012) were able to show that including conflicting claims in scientific news briefs reports increased source awareness in undergraduate students compared to when claims were consistent (*Discrepancy-Induced Source Comprehension*, see also Rouet et al., in press). The effect has also been shown across documents (e.g., Kammerer et al., 2016; Strømsø & Bråten, 2014). In addition, Stadtler et al. (2014) showed that explicitly signalling conflicting claims with lexical cue phrases led to a more balanced processing of conflicting information and increased students' sourcing activities.

A number of interventions were successful at improving high school and university students' source awareness and their ability to use such information for evaluating document credibility (e.g., Brand-Gruwel & Wopereis, 2006; Britt & Aglinskas, 2002; Calkins & Kelley, 2010; Goldman et al., 2009; Graesser et al., 2007; Stadtler & Bromme, 2007, 2008; Wiley et al., 2009; Wopereis, Brand-Gruwel, & Vermetten, 2008, see Brante & Strømsø, 2017 and Braten et al., in press, for a review). Computer-based tutorials may be particularly helpful. For example, the SEEK (i.e. Source, Evidence, Explanation, and Knowledge) Web tutor (Graesser et al., 2007; Wiley et al., 2009) could improve university undergraduates' abilities to use source information for evaluation. In this intervention, students were provided with declarative knowledge about different aspects that need to be considered when evaluating multiple web-based documents. Importantly, they were explicitly instructed to evaluate the credibility of sources, the strength of supporting evidence, and the fit of new information to their existing knowledge about an issue *prior* to a prompted task. Stadtler and

Bromme (2007, 2008) provided university students with a computer tool (*meta.a.ware*) that contained monitoring and evaluation prompts to improve students' comprehension and evaluation skills, including both content and source, whereby students were asked to indicate and rate information about the source. The interventions successfully improved students' attention to sources. Another promising approach for fostering critical thinking and sourcing skills that is set in educational practice is the READI (Reading, Evidence, and Argumentation in Disciplinary Instruction) curriculum (Goldman et al., 2009). It teaches important principles of argumentation and sourcing skills in classroom settings and highlights the importance of integrating information from multiple sources and evidence-based argumentation. More recently, Kim and Hannafin (2016) attempted to integrate an intervention that aimed at improving students' document level literacy skills into a university course. However, although the intervention could improve comprehension, it did not significantly improve students' evaluation skills.

Conclusion

In sum, although different individual and contextual variables influence students' evaluations, previous research indicates that a large number of students seems to struggle with normatively accurate evaluations of scientific texts and arguments. Moreover, many of them do not seem to pay sufficient attention to features of a document's source. However, it appears that even short-term interventions that focus on teaching knowledge about the structure of arguments, normative evaluation criteria, or a competent evaluation of multiple documents, can be helpful for improving students' comprehension and evaluation skills. CLEs that include use different elements for active knowledge construction, such as tutorials, examples, immediate feedback, and prompts may be useful settings for such interventions. In the presence of strong misconceptions, refutations should be included.

Chapter IV

The Present Research

Chapter IV: The Present Research

The findings described in **Chapter 3** have important implications for designing relevant studies and interventions. The empirical studies that are portrayed in the following chapters further examined undergraduate students' competences to evaluate the quality of arguments and, in particular, their ability to activate relevant argument schemes and consider the internal consistency of arguments (**Study 1**). In addition, students' ability to use source information for determining document credibility was assessed (**Study 2**). Furthermore, evidence from two training experiments is presented that addressed some of the challenges among students (**Study 3**). In this chapter, I will provide a brief description of the relevance, aims, and central questions of the present research, followed by an overview of the studies that will be described and discussed in **Chapters 5-8**.

Contributions of the Present Research

The present research extends previous research on reading and scientific literacy in several respects. First, although the number of studies addressing students' epistemic competences has been growing in the last years and their importance has been more widely acknowledged (Goldman et al., 2016; OECD, 2014), the majority of research in cognitive and educational psychology is still concerned with receptive processing strategies, such as organising and structuring of information (Wild, 2000) or text scanning (Bazerman, 1985), and with text comprehension, such as inference-making and elaboration (e.g., Alexander & Fox, 2011; Brooks, 2011; Duke & Carlisle, 2011). In addition, previous research has often examined how students deal with narrative texts (e.g., Anderson et al., 1987; Graesser et al., 1994). In contrast, the present work focused on epistemic competences and the strategies that underlie a valid understanding of scientific texts (Richter, 2003, 2011; Richter & Schmid, 2010), as the ability to deal competently with such documents constitutes an essential aspect of scientific literacy (Britt et al., 2014).

Second, higher-order, reflective thinking processes are most likely to occur in ill-structured environments (i.e. environments that contain highly complex, abstract, and irregular information; King & Kirchener, 1994; Richter 2011) and in advanced learning settings where readers already possess some basic knowledge about a content domain (Richter, 2011). Therefore, the empirical work presented in this focused on university undergraduates rather than high school students. Dealing with multiple scientific texts of varying quality is also particularly prevalent in this group.

Third, acknowledging the finding that readers possess a broad number of systematic and heuristic processing strategies, which they use interchangeably depending on the processing goal (Pressley, 2000; Rouet & Britt, 2011; Wyatt et al., 1993), both systematic and heuristic epistemic competences were assessed, whereby students' use of source information was examined in addition to their evaluations of text content. Although there is agreement among researchers that paying attention to source information is important for selecting relevant and reliable information (see Bråten, Stadler, & Salmerón, in press, for a review), the majority of research on reading literacy still focuses predominantly on evaluations of the content of a text rather than the source (e.g., Blanchard & Samuels, 2015; Brooks, 2011; Kamil, Pearson, Moje, & Afflerbach, 2011). Moreover, existing research on students' use of source information focused mainly on systematic processing strategies (e.g., Bazerman, 1985; Lundeberg, 1987; Wineburg, 1991). In contrast, the present work assessed sourcing behaviour when systematic strategies could not be applied. In addition, it was examined whether there are indications that systematic and heuristic competences might develop as a common construct of discipline expertise (cf. Rouet et al., 1997).

Fourth, although previous research has set some useful standards of instruction, these standards usually do not specify in sufficient detail how they should be taught to students (Goldman et al., 2016). As recommended by several others (e.g., Blair, 1995; Larson et al., 2004; Shaw, 1996), our intervention tested the effectiveness of a training that aimed at

teaching normative aspects of argument evaluation (cf. Blair & Johnson, 1987), in combination with conceptual knowledge about common fallacies (cf. Dauer, 1987). In addition, our intervention focused on teaching the structure of more complex, full-fledged arguments (cf. Toulmin, 1958) rather than simpler, two-clause arguments (e.g., Larson et al., 2009). It should also be noted that, whereas several recent interventions have successfully attempted students' sourcing skills (Bråten et al., in press), the number of interventions targeting students' comprehension and evaluation of informal arguments is still very limited (Larson et al., 2004; Larson et al., 2009).

Finally, whereas most of the research on epistemic competences has been performed in the domain of natural sciences (e.g., Barzilai et al., 2015; Bråten, et al., 2016; Sanchez et al., 2006) or history (e.g., Britt & Aglinkas, 2002; Stahl et al., 1996), the present studies examined students' epistemic competences in the domain of psychology.

Aims and Research Questions

The central goal of the empirical work presented in this dissertation was to further examine students' competences to comprehend and evaluate scientific texts and arguments, to identify relevant strategies competent readers use when dealing with such documents, and to use this knowledge for designing suitable training interventions. Furthermore, the following questions (or sub-goals) were explored:

1. **How (well) do students, compared to scientists, evaluate the plausibility of informal arguments and the credibility of multiple scientific texts under different (systematic vs. heuristic) conditions? Which specific cognitive processing strategies mediate possible performance differences between students and scientists and is performance in one task related to performance in the other?**

2. **How familiar are students with the structure of informal arguments, common argumentation fallacies, and different publication types? To which extent is such conceptual knowledge related to task performance?**
3. **Can familiarising students with the propositional structure of arguments help them to better comprehend and evaluate scientific arguments?**
4. **Can teaching normative aspects of argument evaluation and common fallacies improve students' competences to judge the plausibility of arguments?**
5. **Will some students profit more from training in argumentation than others?**
6. **Can training in the ability to identify structural components of arguments improve the ability to evaluate the quality of arguments as well, and vice versa?**

Drawing on the concept of discipline expertise (Rouet et al., 1997), we expected scientists and better learners to demonstrate a superior performance in all assessed tasks, compared to students and less able learners. In particular, we expected students to experience difficulty with representing and evaluating more complex scientific arguments and with attending to relational aspects between argument components (cf. Larson et al.; 2009; Larson et al., 2004; Shaw, 1996; Wolfe et al., 2009). In addition, students were expected to be less familiar with multiple scientific texts (Britt et al., 2014), and to use source information for evaluating document credibility to a lesser degree than scientists (cf. Bazerman, 1985; Lundeberg, 1987; Wineburg, 1991). Scientists, in contrast, were expected to apply different epistemic heuristic and systematic strategies flexibly, depending on the processing goal (Wyatt et al., 1993). We further expected that teaching the structure of arguments would improve students' performances to represent informal arguments, including more complex, scientific arguments (Larson et al.,

2009; Larson et al, 2004), and that training of normative aspects of argument evaluation, along with conveying conceptual knowledge about common fallacies, would improve their performance to evaluate the quality of arguments (cf. Blair, 1995). Better learners were expected to particularly profit from our interventions, assuming that these students would possess a higher amount of relevant prior knowledge that could be applied during the intervention (cf. Rouet et al., 1997). Furthermore, we expected performances in different epistemic tasks to be related (Britt et al., 2014). Finally, it was hypothesised that fostering knowledge about the structure of arguments might improve students' performances to evaluate the quality of arguments as well (Britt & Larson, 2003; Britt et al., 2014), whereas teaching normative aspects of argument evaluation would also increase students' knowledge about structural components of arguments (Britt et al., 2014).

Overview of Studies

To answer the research questions formulated above, a total of four empirical studies, including both quantitative and qualitative analyses, as well as experimental designs, were performed.

Expert-Novice Comparisons. First, two empirical studies were designed to examine university students' competences to judge the plausibility of informal arguments and identify common fallacies (**Chapter 5**) and to evaluate the credibility of multiple scientific documents (**Chapter 6**). To this end, the performance of psychology students (novices) was compared to the performance of scientists from the domain of psychology (experts), using think-alouds, retrospective interviews, and response accuracy and latency measures. Think-aloud approaches have been widely used in text comprehension research to gain insight into how people are processing reading materials, how they interpret information, which goals they activate to guide comprehension and evaluation, and to keep track of their performance (e.g., Chi, Leeuw, Chiu, & La Vancher, 1994; Coté & Goldman, 1999; Goldman et al., 2012; Magliano & Millis, 2003;

Maier & Richter, 2016; Wineburg, 1991). Cognitive interviews were used to assess clarity of instruction, perceived task difficulty, familiarity with argumentation fallacies, and self-reported strategy use (Prüfer & Rexroth, 2000). Furthermore, combining accuracy and latencies of responses has been shown to be a reliable approach for assessing reading processes in adult readers (Richter & van Holt, 2005). Whereas the first study focused on assessing students' abilities to systematically evaluate scientific arguments (cf. Larson et al., 2009; Shaw, 1996), the second study focused on heuristic evaluations of scientific texts and students' use of source information for evaluation (cf. Rouet et al., 1997; Wineburg, 1991). The primary objective of the expert-novice comparison was to identify useful processing strategies that could be adopted for designing appropriate training interventions. In addition, relationships between performances and relevant domain-specific prior knowledge required in different tasks (i.e. knowledge about arguments in Study 1 and different publication types in Study 2) were examined.

Intervention Study. Based on the results of the expert-novice comparisons, an intervention study consisting of two training experiments was constructed to foster a selection of competences in tasks that students particularly struggled with (**Chapter 7**). These included the ability to comprehend the propositional structure of (complex) arguments (cf. Toulmin, 1958), and the ability to use normative criteria, such as judging the internal consistency of arguments (cf. Blair & Johnson, 1987; Shaw, 1996) for evaluation, because both aspects seem to be essential for dealing successfully with arguments in scientific texts (Britt & Larson, 2003; Britt et al., 2014; Halpern, 1998; Wolfe et al., 2009). Implausible information was included along with relevant declarative knowledge about normative aspects of argument evaluation (cf. Blair, 1995) to encourage elaborative epistemic processes (Richter, 2011). Linguistic markers were included to highlight the structure of arguments (Experiment 1, cf. Britt & Larson, 2003) and plausible alternatives were provided to change possible misconceptions (Experiment 2, cf. Braasch et al., 2013). Based on Jonassen's (1999) cognitive modelling approach, the

intervention was set in a stimulating environment that could be manipulated by the learner. Expert models were used to teach relevant strategies (cf. Chi, Glaser, & Rees, 1982; Larson et al., 2009). Immediate feedback (cf. Deci, 1971; Jonassen, 1999), learning goals and instructional prompts (Renkl, 2009) were also important elements. In both experiments, the performance of students who took part in a training of the structure of full-fledged arguments (Experiment 1) or of normative criteria for the evaluations of their plausibility (Experiment 2) was compared to the performance of a control group, using a pre-post-follow-up design. In addition, the role of study performance for training success was examined to account for individual differences in prior knowledge (Rouet et al., 1997).

The studies are presented in detail in **Chapters 5-7** and summarised and further discussed in **Chapter 8**.

Chapter V

Study 1

Judging the Plausibility of Arguments in Scientific Texts: A Student-Scientist Comparison

A version of this chapter was published in:

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E.M., & Berthold, K. (2016). Judging the plausibility of arguments in scientific texts: A student-scientist comparison. *Thinking & Reasoning, 22,* 221-246.

Abstract

The ability to evaluate scientific claims and evidence is an important aspect of scientific literacy and requires various epistemic competences. Readers spontaneously validate presented information against their knowledge and beliefs but differ in their ability to strategically evaluate the soundness of informal arguments. The present research investigated how students of psychology, compared to scientists working in psychology, evaluate informal arguments. Using a think-aloud procedure, we identified the specific strategies students and scientists apply when judging the plausibility of arguments and classifying common argumentation fallacies. Results indicate that students, compared to scientists, have difficulties forming these judgements and base them on intuition and opinion rather than the internal consistency of arguments. Our findings are discussed using the mental model theory framework. Although introductory students validate scientific information against their knowledge and beliefs, their judgements are often erroneous, in part because their use of strategy is immature. Implications for systematic trainings of epistemic competences are discussed.

Keywords: informal argument evaluation, epistemic competences, mental model theory, think-aloud procedure, competences in higher education

Arguments can affect our daily lives in many ways, whether we think of politicians trying to persuade us to vote for a particular party, a newspaper article providing a certain perspective on a societal issue, or taking decisions about which kind of career we would like to pursue. In scientific discourse, arguments also play a central role, because they link theoretical claims to supporting empirical evidence. Students entering university are confronted with scientific literature that presents different and at times conflicting theories backed up by more or less compelling evidence. The ability to evaluate scientific claims and evidence is an important aspect of scientific literacy and requires various epistemic competences (Britt, Richter, & Rouet, 2014).

The present research investigated how students of psychology, compared to scientists working in psychology, evaluate arguments and which strategies they use to judge their plausibility. Successful readers possess a broad number of general processing strategies that they use in a flexible way, depending on the processing goal (Wyatt et al. 1993). Although argumentation skills are generally not formally taught in higher education, we expect scientists to implicitly acquire these epistemic competences in the course of academic socialisation. Within the scientific community, reading and evaluating scientific texts belongs to a scientists' daily activities, and controversies are valued as a means of fostering and advancing understanding (Britt et al., 2014). Moreover, as a result of reading a broad range of scientific literature, scientists are generally more familiar with the basic structure of arguments (Britt & Larson, 2003). To comprehend, interpret and critically evaluate information presented in the text, scientists form abstract representations of the functional components of arguments and their interrelations. In contrast, students often misinterpret disagreement as a general uncertainty on a scientific issue, leaving them in a position in which they feel the need to either solve or tolerate discrepancies (Britt et al., 2014). In our study, we not only investigated differences between scientists' and students' accuracy in

judging the plausibility of arguments but also explored the strategies that explain the presumed superior performance of scientists with the purpose of identifying successful strategies for the evaluation of arguments.

We start from the assumption that knowledge about the structure of an argument is particularly important for understanding and evaluating arguments (Britt et al., 2014; Britt & Larson, 2003; Wolfe, Britt, & Butler, 2009). In this context, we will present the Toulmin model of argumentation (Toulmin, 1958) to describe the typical structure of an argument and use this model as the background for sketching the basic normative aspects of argument evaluation (e.g., Shaw, 1996). Subsequently, we will discuss the challenges and pitfalls students typically face when trying to accurately judge the plausibility of arguments. The mental model theory (Johnson-Laird, 1983) is used as a framework to explain differences between scientists and students and will provide a basis to make assumptions about the present research.

The Toulmin Model of Argumentation

Competent readers will use their prior knowledge about the structure of an argument along with conceptual and genre knowledge to evaluate its plausibility (Britt et al., 2014; Britt & Larson, 2003; Wolfe et al., 2009). Familiarity with the rhetorical structure allows them to identify the main claim, connect and evaluate the premises supporting the claim, and activate possible alternative explanations.

In 1958, the British philosopher Stephen Toulmin established an influential argumentation model in which the typical structure of an argument was described as syntactic relations among five key components: claim, datum, warrant, backing, and rebuttal (Toulmin, 1958). The *claim* comprises the main statement being argued for and its acceptability is open for debate. It is controversial in the sense that not everyone will agree with it. By definition, an argument requires support, which may be theoretically or

empirically derived, because if the claim were not controversial, its support would be self-evident. This evidence is referred to as *data* (or ground). The *warrant* forms the link between claim and data and determines the strength of the support for the conclusion, while *backing* evidence provides support for the warrant. Finally, *rebuttals* limit the range in which the argument holds true. Consider the following example:

> Harry was born in Bermuda. A man born in Bermuda will generally be a British subject, on account of the British National Acts. Therefore, Harry is presumably a British subject, unless he has become a naturalised American. (Toulmin, 1958, p. 94)

The claim that Harry is a British subject is supported by the datum that Harry was born in Bermuda. The datum lends support to the claim because of the warrant that a man born in Bermuda will generally be a British subject. Backing for the warrant is stated by referring to the British National Acts. However, the argument is only conclusive if Harry has not changed his nationality since birth. This sentence constitutes the rebuttal.

In everyday life, warrants are often not explicitly stated, but implicitly implied so that readers need to make inferences about the relevance of the data for the claim (e.g., Chambliss, 1995). However, in scientific literature, explicitly stating why a particular conclusion is drawn from the results provided in a study, or why the use of a particular statistical method justifies this conclusion, is important. Moreover, rebuttals are especially important in science, because they comprise an essential part of how scientific knowledge is derived in a scientific community (Britt et al., 2014). The order in which the different components are represented is probably hierarchical with the claim holding the top position, followed by the datum and the other components (Britt & Larson, 2003). Markers such as "therefore" or "causes" signal relations between components, and arguments that include markers have shown to be processed faster than arguments without these signals (Britt & Larson, 2003).

The Evaluation of Informal Arguments

Traditionally, the focus in cognitive psychology has been on formal reasoning. However, scientific texts usually comprise informal rather than formal arguments, and arguments in everyday life are rarely deductively valid or have a single verifiable solution. Toulmin (1958) proposed his model in reaction to the traditional formal reasoning perspective. However, not until the 1980s when Blair and Johnson postulated their informal reasoning approach was the discipline successfully established as a research field (van Eemeren et al., 1996). Informal arguments differ from formal arguments in several respects. Like formal arguments*, informal arguments* always contain a claim and a conclusion. However, they may consist of additional components such as warrants (Toulmin, 1958), or components may not be explicitly stated but instead have to be inferred by the reader (Green, 1994). Informal arguments are generally more ill-structured and therefore not necessarily deductively valid. Finally, the conclusion in an informal argument is not logically consistent but it can be criticised by generating rebuttals or counter-arguments. According to Voss and Means (1991), the acceptability of an informal argument depends on three criteria: the truth of data and claim, the quality of the relationship between these components, and whether all aspects of the topic have been considered.

When people evaluate the acceptability of arguments, they can be based on arguments, alternatives, or assertions (Shaw, 1996). *Argument-based* judgements consider the internal consistency of an argument, that is whether the data provide relevant support for a claim (*relevance*, cf. Blair & Johnson, 1987). *Alternative-based* judgements focus on a different aspect of internal consistency by evaluating whether all information relevant for the truth of the claim has been considered (*sufficiency*, cf. Blair & Johnson, 1987). Finally, *assertion-based* judgements evaluate the truth of the claim or the data (*accuracy*, cf. Blair & Johnson, 1987). All three types of judgements are necessary to achieve a complete evaluation of a

claim's acceptability. However, the strength of an argument can only be assessed by argument-based and alternative-based judgements, both of which focus on the internal consistency of the argument.

Mental Model Theory and the Evaluation of Informal Arguments

Successful readers not only need to identify different components of an argument correctly, they also need to establish connections between ideas within the text and with relevant prior knowledge, and actively construct coherent representations of the text and its content (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Magliano & Millis 2003; Shaw, 1996; van den Broek, Risden, & Husebye-Hartman, 1995). Researchers in cognitive psychology (e.g., education, language, and reading) widely accept that readers construct a referential representation of the situation described in a text, which comprises the state of affairs that is described in the message rather than the message itself (Johnson-Laird, 1983; van Dijk & Kintsch, 1983). These representations are crucial to a deeper understanding of the information presented in a text (Mayer, 1989). For example, O'Brien and Myers (1999) suggested that individuals use their general knowledge to construct a *mental model* of the text content.

According to the mental model theory, readers use mental models to also evaluate the truthfulness of formal arguments (Johnson-Laird & Byrne, 1992) and informal arguments in a text (Galotti, 1989; Perkins, 1986; Shaw, 1996). When asked to evaluate the acceptability of a statement, readers generate deductive and inductive inferences. Based on their prior knowledge, they imagine a situation in which premises and conclusion are true and a situation in which either premises and conclusion are both false or premises are true but the conclusion is false while searching for alternative explanations (Johnson-Laird & Byrne, 1992; Shaw, 1996).

Lay Readers' Evaluations of Informal Arguments

Lay readers engage in these evaluations to some extent, and they are able to distinguish strong from weak arguments (see Johnson, Smith-McLallen, Killela, & Levin, 2004 for a review). For example, van Eemeren, Garssen, and Meuffels (2009) recorded participants' ratings of argument reasonableness under several conditions. In one condition, strong arguments were used to defend a certain claim (i.e., a rebuttal against someone else's argument). In another condition, the argumentation was weak, and a fallacy was used to defend the claim. Participants gave higher ratings in the first condition, indicating that they seemed to possess some knowledge about the quality of arguments. Similarly, Hoeken, Timmers, and Schellens (2012) found that the participants in their study were sensitive to violations of a number of argument-specific criteria, for example, arguments from analogy. As another example, Hornikx and Hoeken (2007) found that a claim was perceived as more persuasive by their participants when it was supported by high-quality data (e.g., statistical evidence), although this effect was observed only in Dutch but not French students. In another study (Hornikx & ter Haar, 2013) the same effect was observed in Dutch but not German students.

Although lay readers engage in argument evaluations to some extent, various studies show that they are not always accurate in their evaluations. For example, van Eemeren, Garssen, and Meuffels (2012) found that *ad hominem* attacks were rejected by students as unreasonable, but were perceived as more reasonable when they were presented as if they were critical questions regarding authority argumentation. In another study, Hoeken and van Vugt (2014) provided participants with several strong and weak arguments, and observed them in a debate. They found that the participants in their study processed arguments in a biased way, using evaluation criteria for arguments based on analogy or expertise only when the argument went against the claim they were asked to defend. They neglected to use these

evaluation criteria when the argument supported the claim. Moreover, students often have difficulties generating rebuttals. In a qualitative and quantitative study of high school students' abilities to reason on a socio-scientific issue, less than 40% of participants were able to generate rebuttals (Wu & Tsai, 2007). Similarly, other research in this domain found that the majority of high school and college students asked to write argumentative texts failed to spontaneously include rebuttals in their texts (*my-side bias*, Perkins, Farady, & Bushey, 1991). This shortcoming poses a problem to the generation of assertion-based judgements, because these judgements require a search for alternative explanations (Johnson-Laird & Byrne, 1992; Shaw, 1996).

Lay readers also often fail to consider all relevant components of an argument and how they are related. Research on the evaluation of deductive arguments indicates that people tend to accept invalid arguments if they believe both premises and conclusions are true (*belief bias*, see Evans, Newstead, & Byrne, 1993 for a review). Similar to the belief bias in the processing of deductive arguments, research on the evaluation of informal arguments has shown that lay readers tend to focus on the truthfulness of the claim (acceptability) but neglect the relevance of the data for the conclusion (relevancy) or relevant alternatives (sufficiency). For example, in a study by Larson, Britt, and Kurby (2009), college and high school students struggled with the evaluation of the quality of two-clause (claim, datum) arguments. Teaching the structure of arguments and providing immediate feedback were necessary to encourage assertion-based judgements and to shift attention to the internal consistency of arguments. Similarly, in a study by Shaw (1996), undergraduate and graduate students were far more likely to object to the truth of the premises and the conclusions of an argument than to violations of the internal consistency. In other words, they neglected argument-based and alternative-based judgements in favour of assertion-based judgements.

In terms of mental model theory, argument-based judgements are more effortful, because they require the reader to think of possible alternative explanations and simultaneously keep those explanations in working memory. Specifically, the reader is required to imagine conditions under which both the data (the premises) and the claim (the conclusion) are true and conditions in which the data are true but the claim is false. Moreover, both data and claim need to be activated in working memory for readers to be able to evaluate their link. However, lay readers often fail to keep track of an argument's structural components but instead build a unified mental model in which premises and conclusions are not separately represented (Shaw, 1996). Readers trained in argumentation skills are able to build a correct structural representation of arguments and evaluate not only the acceptability of the data and the claim but also the relevance and the sufficiency of the data for the claim (Shaw, 1996; Voss & Means, 1991). The evaluation of arguments may propose a challenge to lay readers, because the evaluation of the internal consistency of arguments is effortful and often requires domain knowledge that is not explicitly stated in the text (Britt et al., 2014).

Aims and Hypotheses of the Present Study

The present study examined how scientists (post-docs and advanced doctoral students) and first-year students evaluate the plausibility of arguments embedded in expository texts from the domain of psychology. The texts included strong and weak arguments and participants were asked to mark strong arguments as plausible and weak arguments as implausible. The weak arguments represented common types of argumentation fallacies. In a second step, those arguments classified as weak were required to be allocated to a specific fallacy.

The present research extends previous studies by using both on- and offline measures including think-aloud protocols and a retrospective interview. By employing this method, we

can gain insights into the strategies that mediate performance differences between scientists and students. To control for effects of think-aloud protocols, the study materials included two parallel versions of the plausibility judgement task (with different texts), and participants completed one version silently and the other while thinking aloud. The present research further extends previous studies, because we also tested the performance of scientists and students in identifying the structural components of full-fledged arguments (including warrants, Toulmin, 1958).

First, drawing on the notion that knowledge about the typical structure of an argument provides a more accurate representation of information that is essential for evaluating it, we expected the performance in judging the plausibility of arguments to be positively correlated with the performance in the task to identify the structural components of arguments (Britt et al., 2014). Second, as scientists are generally more familiar with the basic structure of arguments (Britt & Larson, 2003), and use this knowledge to evaluate their plausibility (Britt et al., 2014; Britt & Larson, 2003; Wolfe et al., 2009), we expected scientists to be more accurate in identifying the structural components of informal arguments and also to provide more accurate plausibility judgements. Generally, we expected both scientists' and students' judgements to be more accurate when strong compared to weak arguments were presented. Overall, strong arguments are likely to communicate plausible information, which is conceptually coherent with a person's prior knowledge and is readily accessible (Johnson-Laird, 1983). Thus, plausible information is generally more likely to be integrated into the current state of the situation model (Schroeder, Richter, & Hoever, 2008). Third, we expected scientists, compared to students, to use superior strategies for their judgements. We know from expertise research that experts use highly routinized strategies that they apply fast and efficiently (Wyatt et al., 1993). The scientists in our study were not experts in the sense that they achieved exceptional skills at a certain activity as the result of a prolonged

effort of deliberate practice to improve performance (e.g., Ericsson, Krampe, & Tesch-Römer, 1993). However, they completed or were close to completing a doctorate degree in psychology, which requires several years of disciplinary training, whereas the students in our study were new to this field. We use the term *discipline expertise* to refer to the amount of expertise the participants in our study possess in the domain of psychology (cp. Rouet, Favart, Britt, & Perfetti, 1997), with scientists being discipline experts in this context. Following the results of Shaw (1996) that more experienced readers are able to build better structural representations of arguments and to evaluate not only the acceptability of the data and the claim but also the relevance and the sufficiency of the data for the claim, we proposed that scientists would base their plausibility judgements on the arguments and be more attentive to the internal consistency of arguments than students. In contrast, as lay readers often fail to keep track of the argument's structural components, students were expected to focus more on the acceptability of the claim or premise alone (i.e., make assertion-based judgements), or base their judgements purely on intuition. Importantly, we expected the superior judgements of scientists to be mediated by strategy use. Finally, checking the internal consistency of arguments is cognitively demanding (Shaw, 1996). Thus, differences between scientists and students were expected to show not only in the error rates, but also in the response latencies, with scientists showing longer response times. In sum, the following expectations were formulated:

H1: The performance in judging the plausibility of arguments is positively correlated with the performance in the task to identify the structural components of arguments.

H2: Scientists are more accurate in identifying the structural components of informal arguments and provide more accurate plausibility judgements.

H3: Judgements are more accurate when strong compared to weak information is presented.

H4: Scientists, compared to students, use superior strategies, such as argument-based judgements, whereas students more frequently make assertion-based judgements or judgements based on intuition.

H5: The superior judgements of scientists are mediated by their strategy use.

H5: Scientists take more time to evaluate the plausibility of arguments.

Method

Participants

Twenty first-year psychology students and 20 scientists (8 postdocs and 12 advanced doctoral students in psychology who were at least in their third year of their doctoral studies and close to graduation) participated in the study. The sample consisted of 77 % females and 23 % males (students: 80 % females vs. 20% males; scientists: 74% vs. 26 %). The average age of students was 21.70 years (*SD* = 4.18), whereas the average age of scientists was 30.81 (*SD* = 5.08). Participants provided informed consent at the beginning of the experiment and were reimbursed with course credits or financial remuneration (25 Euros per hour for scientists and 8 Euros per hour for students) after its completion.

Plausibility judgements

Text materials. The text materials provided for the plausibility judgements consisted of two expository texts (one in each version of the task, see below) similar to those typically read by psychology undergraduates. Both texts addressed theories on smoking behaviour (371 words in one text, 394 in the other; adapted from Fuchs & Schwarzer, 1997, and Schroeder et al., 2008). Both texts consisted of 22 items including strong and weak arguments. All statements were part of a coherent text. Five sentences in each version were

weak, and the remaining sentences either were strong arguments by themselves or formed strong arguments with the previous sentence. Inconsistencies were always stated in one sentence to avoid global incoherencies. Weak arguments were created by attenuating the justification for the claim and included one of five common argumentation fallacies (false conclusion/contradiction, false dichotomy, wrong example, circular reasoning, overgeneralisation; Dauer, 1989; see Table A1 for examples). For example, the sentence *The theoretical construct of inherent nicotine sensitivity holds that some people react more sensitively to nicotine, because they are more susceptible to nicotine* contains a circular definition of the concept of nicotine sensitivity (the actual text materials were in German, e.g., *Das theoretische Konstrukt der Nikotinsensitivitätstheorie besagt, dass manche Menschen sensitiver auf Nikotin reagieren, weil sie sensibler auf Nikotin ansprechen*). The five fallacies were chosen, because they represent rather blunt instances of weak informal arguments (non-sequitur arguments) and their detection does not require any formal training in argumentation. An important aspect of this method is that meaning and propositional content were not spoiled by the implausibility manipulation. That is, all weak arguments were semantically and syntactically correct. Likewise, sentences in both strong and weak arguments were coherent with previous discourse context and were thus congruent with the current state of the situation model. The only difference between strong and weak items was that the data supporting the claim of an argument were poor or defective.

The sentences communicating strong or weak sentences were selected by a quasi-random procedure. Both types of sentences were comparable in features such as length or semantic complexity. The mean length of strong and arguments was 3.4 clauses (Text 1) and 3.5 clauses (Text 2). Moreover, they had similar readability scores—32 for the strong arguments vs. 38 for the weak arguments—as indexed by the German adaptation of Flesch's Reading Ease Index (Amstad, 1978).

The selection and presentation of test items was slightly adapted for our purpose. Our items were embedded in a larger test battery including different tasks to assess various epistemic competences and their correlations (for further tasks not reported here, see von der Mühlen, Richter, Schmid, Schmidt, & Berthold, 2016). Texts were shortened to contain 22 items and five weak arguments (i.e. five types of argumentation fallacies) in each version.

Validation of text and item materials. The texts used for the plausibility judgements had been normed and validated in a study by Schroeder et al. (2008) who found that strong and weak arguments differed in their plausibility. However, given that the selection and presentation of items were slightly different in our study, all items were pretested again in a pilot study with 101 introductory psychology students. The pilot study served as a basis to select suitable items for the final test battery. Reliability was calculated separately for the two parallel versions of the tests. Cronbach's α was .64 for the response accuracy and .97 for the response times in Version 1, and .64 vs. .94, respectively, in Version 2. The correlation between parallel versions in this test was $r = .49$, $p < .01$.

Argument structure

Text materials. The text materials provided for the identification of argument components, that is, the argument structure task, consisted of short texts (141 words in Argument 1, 117 words in Argument 2) including claim, datum, warrant, backing, and rebuttal (Toulmin, 1958). The texts had similar readability scores (38 in Version 1 vs. 32 in Version 2), and the test consisted of 12 items (six items in each parallel version of the test).

Validation of text and item materials. The texts used for the argument structure task were pretested in the same pilot study as those for the plausibility judgements. Reliability was again calculated separately for the two parallel versions of the tests. Cronbach's α was .86 for the response accuracy and .92 for the response times in version 1, and .87 vs. .95,

respectively, in version 2. The correlation between parallel versions in this test was $r = .81$, $p < .01$.

Software

The testing software used to display the texts and to record responses and response times was Inquisit 3.0.6.0. This software enables response time recordings in the millisecond range.

Screen activity recordings

The software package HyperCam 2.28.01 was used to record screen activity (e.g., mouse movement and clicking). Recording screen activity was important to relate utterances in the think-aloud protocols to the part of the test text that participants were working on.

Procedure

Participants were tested individually in a laboratory. Upon arrival, they were welcomed by the experimenter, seated in front of a computer (HP notebook, 15`` screen) and given a headset. All words were presented in black font (Calibri 12) against white background, with a visual angle of 1.4 degrees. Two exceptions were the reminders for the keys representing plausible and implausible response options and the sentences participants had marked implausible which appeared in red font. Each participant completed two parallel versions of the tests; one task was completed in silence and the other while thinking aloud. The parallel versions contained different texts, but the tasks were identical.

At the beginning of the session, participants were provided with the two short texts and were asked to identify the different components of an argument. The texts were presented both complete and in fragments. Participants were asked to read the complete text first. In the next step, the text was presented to them in fragments consisting of several paragraphs, each representing a different component of the argument (i.e. claim, datum, warrant, backing, and rebuttal). The paragraphs were numbered, and participants were instructed to

assign each number to its corresponding argument component that participants could select from a list appearing at the bottom of the screen. Participants were given as much time as they needed to complete the task.

After completion of the argument structure task, participants were asked to judge the plausibility of different arguments in two texts. They were instructed to read the texts thoroughly, sentence-by-sentence on a computer screen in a self-paced fashion. Participants judged the plausibility of each sentence by pressing a key for *plausible* or another key for *implausible.* They were asked to judge the internal consistency and quality of the arguments and not to base their judgements on their opinion or prior knowledge about the content of the text. Furthermore, they were told that global fallacies (i.e. inconsistencies of a statement with other passages mentioned earlier in the text) were not included. After participants rated the plausibility of all items of a text, they were instructed to allocate all sentences they had marked implausible to specific argumentation fallacies. Completion of both test versions took about 30-45 minutes. At the end of the session, all participants were interviewed. Finally, they were thanked for participation and dismissed. Participants were debriefed a few weeks later.

Think-aloud Protocols

Think-aloud protocols were obtained during plausibility judgements in one version of the test. All participants worked on one version of the test in silence and on the other while thinking-aloud. In the think-aloud version, participants were instructed to say aloud "everything that comes to mind" while they were working on the tests. In particular, they were asked to consistently think aloud while they were inspecting the texts and while they made their judgements to receive online measures and to prevent disruption of a continuous flow of thought. Half of the participants received the silent version first, the other half the think-aloud version. Participants were audiotaped while wearing a headset.

Transcription. The audio-recordings were transcribed verbatim. From the total transcriptions, 30% were cross-checked for accuracy, indicating 100% accuracy. Every item in the test served as one unit of analysis.

Coding. All forty protocols were coded. Based on previous theory and research on argument evaluation and epistemic strategies (Blair & Johnson, 1987; Richter & Schmid, 2010; Shaw, 1996), a coding scheme for the strategies derived from the think-aloud protocols was developed containing three main categories: *intuitive judgements* (e.g., "Somehow this does not seem plausible to me"), judgements based on the *internal consistency* of the arguments (e.g., "There is a clear contradiction between claim and reason"), and plausibility judgements based on *knowledge or opinion regarding the claim* ("This cannot be true. I am a smoker and I know that"). Judgements based on the internal consistency of the arguments (i.e. argument-based judgements, cf. Shaw, 1996) were related to the relevance of data for the claim, whereas judgements based on the participants' knowledge or opinions regarding the claim (i.e. assertion-based judgements, cf. Shaw, 1996) were related to the accuracy of the information stated in only one component of the argument (cf. Blair & Johnson, 1987). The three categories were used to differentiate between those who considered all components of an argument and their relationship, those who considered only one component, and those who failed to evaluate the argument strategically at all (intuitive judgements). Inter-rater reliability approximated almost perfect agreement (Cohen's κ, based on 10 randomly selected protocols coded by two raters was .91).

Interview

A retrospective interview was conducted to gain further insights in the procedural and declarative knowledge used by scientists and students during task completion. Based on Prüfer and Rexroth (2000), different probing techniques (i.e., general probing, category

selection probing, information retrieval probing, and special comprehension probing; Table A2) were used to assess clarity of instruction, perceived task difficulty, familiarity with argumentation fallacies, and self-reported strategy use.

Design

The study was based on a design with pre-existing groups defined by different degrees of discipline expertise (scientists vs. students). The test battery included two parallel versions of the test with one text in each version. All participants read one text in silence and the other while thinking-aloud to control for potential effects of the think-aloud procedure. The order of the two versions and the assignment of test versions to the silent and the think-aloud conditions were counterbalanced across participants. Moreover, we counter-balanced the order of texts appearing first or second. Response latencies, accuracy of answers, and response strategies derived from the think-aloud protocols and interviews were recorded as dependent variables.

Results

Type I error probability was set at .05 for all hypothesis tests. One-tailed tests were applied for testing univariate hypotheses that predicted higher values of scientists compared to students. The hypotheses were tested in a series of analyses. First, we conducted a Multivariate Analysis of Variance (MANOVA) for the response accuracies to account for increased error rates resulting from multiple comparisons between scientists and students. Univariate follow-up tests were performed to interpret group differences. In addition, Holm-Bonferroni corrections (Holm, 1979) were applied in the dependent variables for each group of univariate comparisons (response accuracies, response latencies, strategies derived from the think-aloud protocols, self-reported task difficulties, and strategies derived from the interview). Second, we used univariate comparisons to test for differences in the use of specific strategies as revealed by the think-aloud data. For these univariate tests, the given

sample size yielded a power (1-β) of .80 given the sample size and Type I error probability for detecting large effects (Cohen's $d \geq 0.80$). For medium effects (Cohen's $d \geq 0.50$), power was .47 (power was determined with GPower, Faul, Erdfelder, Lang, & Buchner, 2007). Third, differences between scientists and students regarding understanding of the task, perceived difficulty and strategy use were calculated from the retrospective interview using the χ^2 test. Fourth, a multi-level logistic mediation analysis (test items nested within participants) was conducted to examine the extent that strategy use can explain the expected superior performance of scientists as reflected in the response accuracies. The reported means and standard errors—with the exception of those reported in the mediation analysis—were computed with subjects as the units of observation.

Response Accuracy

The MANOVA revealed a significant main effect of expertise regarding accuracy of answers when judgements of plausibility (implausible and plausible), fallacy allocations, and judgements about argument structure were assessed. As expected, students' answers, compared to those given by scientists, were more often erroneous, $F(4, 34) = 4.87$, $p < .01$, $\eta^2 = .36$. Univariate follow-up analyses were performed to interpret the multivariate group difference. Holm-Bonferroni corrections resulted in a Type I-error probability of .01 for the fallacy allocations, .013 for weak items, .017 for overall plausibility judgements, .025 for the argument structure task, and .05 for strong items.

Plausibility judgements. Scientists ($M = 0.81$, $SE = 0.02$) outperformed students ($M = 0.71$, $SE = 0.03$) when judging the plausibility of the text items, $t(37) = -3.05$, $p < .01$, $d = 0.94$ (Figure 1). In particular, they more often identified weak sentences correctly ($M = 0.76$, $SE = 0.03$ vs. $M = 0.58$, $SE = 0.04$, respectively), $t(37) = -3.73$, $p < .001$, $d = 1.20$. In addition, scientists tended to identify more strong sentences correctly ($M = 0.85$, $SE = 0.02$ vs. $M = 0.81$, $SE = 0.03$, respectively). However, this difference did not reach significance,

$t(37) = -1.11$, $p = .14$. Finally, scientists ($M = 0.60$, $SE = 0.04$) performed significantly better than students ($M = 0.39$, $SE = 0.04$) with allocating weak items to specific types of argumentation fallacies, $t(37) = -4.03$, $p < .001$, $d = 1.27$. No significant differences were found in response accuracy in this task between the silent ($M = 0.69$, $SE = 0.03$) and the think-aloud condition ($M = 0.68$, $SE = 0.02$), $p = .83$.

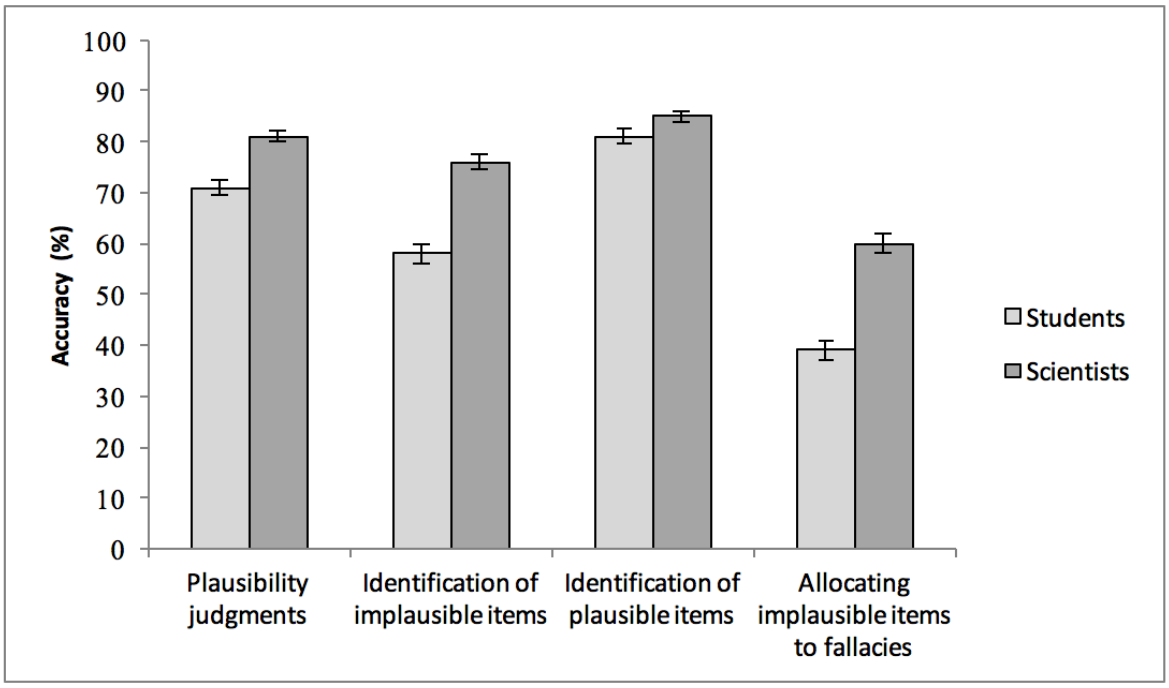

Figure 1. Mean proportion of accurate responses (with standard error of the mean) regarding overall plausibility judgements, identification of plausible and implausible items, and allocation of implausible items to specific fallacies, for scientists and students.

Argument structure. Students, compared to scientists, showed deficits in identifying the functional components of the arguments correctly ($M = 0.42$, SE = 0.06 vs. $M = 0.66$, $SE = 0.07$, respectively), $t(35.2) = -2.62$, $p < .01$, $d = 1.20$. Recognising the warrant ($M = 0.25$, $SE = 0.08$) and backing ($M = 0.20$, $SE = 0.08$) was especially difficult, followed by the claim ($M = 0.43$, $SE = 1.00$), the datum ($M = 0.53$, $SE = 0.09$), and the rebuttal ($M = 0.68$, $SE =$

1.00). Response accuracies in this task were positively correlated with accuracy of responses in the plausibility task ($r = .47$, $p < .01$) and the identification of weak items ($r = .50$, $p = .001$).

Response latencies

We log-transformed the latencies to normalise their distributions and to linearise their relationship with the predictor variable (Ratcliff, 1993) for testing differences between scientists vs. students. Response latencies with a duration of less than one second were discarded from the analysis. Holm-Bonferroni corrections resulted in a Type I-error probability of .01 for strong items, .013 for overall plausibility judgements, .017 for the argument structure task, .025 for weak items, and .05 for the condition (think-aloud vs. silent).

Plausibility judgements. Scientists took significantly longer (M= 34.2 s, SE = 0.32 s) than students (M = 23.9 s, SE = 0.2 s) to evaluate the plausibility of each statement, $t(37) = -2.76$, $p < .01$, $d = 0.92$. This was true for both strong items (M = 33.9 s, SE = 3.12 s in scientists vs. M = 23.7 s, SE = 1.73 s in students, $t(37) = -2.89$, $p < .01$, $d = 0.95$) and weak items (M = 33.4 s, SE = 0.33 s vs. M = 23.9 s, SE = 0.24 s, respectively, $t(37) = -2.36$, $p = .01$, $d = 0.77$). There was a tendency for participants to take more time for their judgements in the think-aloud condition (M = 31.4 s, SE = 3.5 s) than in the silent condition (M = 26.5 s, SE = 1.8 s), but this difference did not reach significance, $p = .11$.

Argument structure. Likewise, scientists (M = 37.1 s, SE = 3.5 s) spent more time in identifying the different argument components than students (M = 26.3 s, SE = 2.1 s), $t(37) = -2.69$, $p < .01$, $d = -0.88$.

Response Strategies

Based on the data derived from the think-aloud protocols, 36.1% of judgements were classified as intuitive, whereas 27.8% were classified as judgements based on the internal

consistency of the statement. For example, a typical intuitive judgement concerning the weak (circular) sentence *The theoretical construct of inherent nicotine sensitivity holds that some people react more sensitively to nicotine because they are more susceptible to nicotine* was: "I don't know why, but that just doesn't sound plausible to me" (student, female), while a typical statement regarding the internal consistency was: "Sensitive and susceptible, that's basically the same! They are trying to use slightly different words, but claim and reason are basically the same here, so the reason is no good for the claim!" (scientist, female). Moreover, 14.3% of all judgements were assertion-based and made references to participants' knowledge or their opinion regarding the truth of either the premise or the conclusion but not references to the link between premise and conclusion. One representative statement in this category was: "Is that so? I'd rather say that some people are simply more addicted, while others are less addicted" (student, male). Other judgements were based on factors such as references provided in the text (6 %), global text coherence (2.6%), and perceived (in)completeness of the information stated (2.3%). In 2.3% of all items, no coding was possible, because participants neglected to think aloud or made utterances unrelated to their judgements.

Holm-Bonferroni corrections resulted in a Type I error probability of .008 for statements based on the internal consistency, .01 for those based on intuition, .013 for those based on knowledge or opinion regarding the claim, .017 for statements based on references, .025 for those based on global text coherence, and .05 for those based on perceived (in)completeness. The response strategies differed between scientists and students (Figure 2). Students relied more often on their intuition ($M = 0.44$, $SE = 0.05$) than scientists ($M = 0.27$, $SE = 0.04$) when making their judgements, $t(33) = 2.42$, $p < .01$, $d = 0.82$), whereas scientists based their judgements more often on evaluations of the internal consistency of the arguments ($M = 0.47$, $SE = 0.04$ in scientists vs. $M = 0.12$, $SE = 0.04$ in students), $t(33) = -5.67$, $p < .001$, d

= 2.00. Students tended to more often base their judgement on knowledge or opinion regarding the premise or claim (M = 0.17, SE = 0.05 in students vs. M = 0.11, SE = 0.03 in scientists). This difference, however, did not reach significance, $t(33) = 1.00$, $p = .16$. No significant group differences were found for the remaining response strategies, $p = .19$ for references, $p = .27$ for global text coherence, and $p = .29$ for perceived (in)completeness.

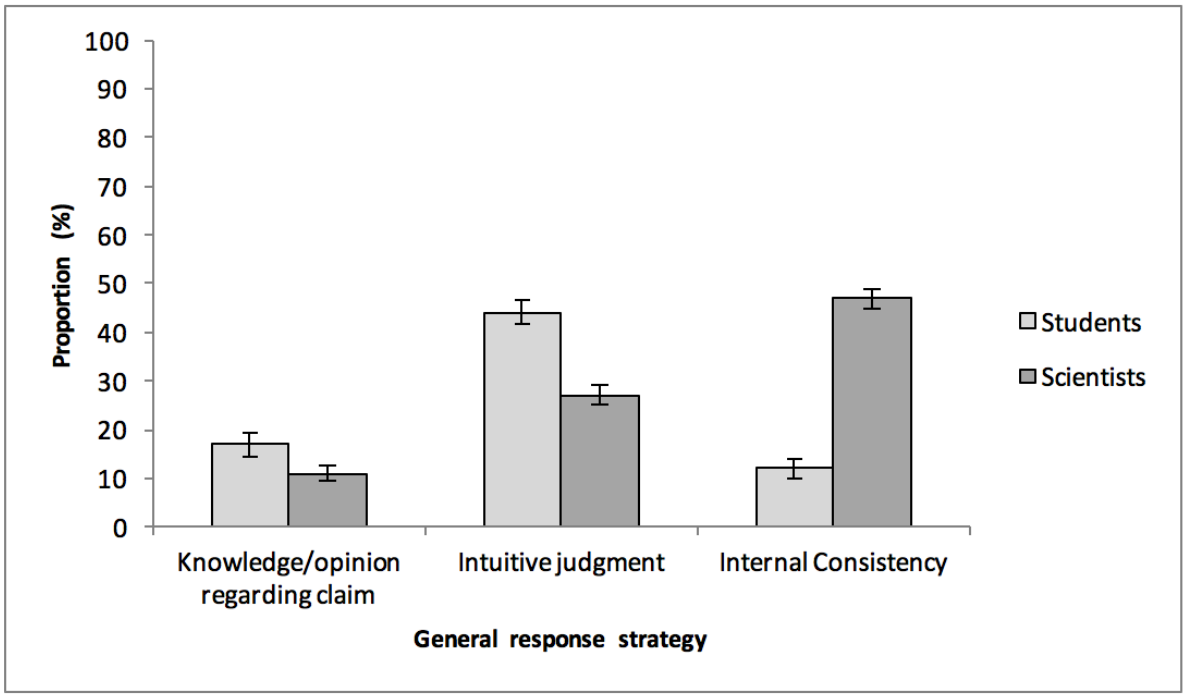

Figure 2. Mean proportions of general response strategies (with standard error of the mean) applied by scientists vs. students.

Interview

The concept of plausibility. Upon inquiry, the task requirement was clear to all but one participant from the student group who reported difficulties in understanding the concept of plausibility entirely. However, this concept seemed to be clear to all remaining participants and to all scientists. When asked to provide a definition of plausibility, both groups were able to give accurate answers (100% of scientists and 95% of students), although scientists

often provided more elaborative definitions. Accuracy of answers was determined by the definition of the word *plausibility* as defined in the German dictionary Duden 05 (2010).

Perceived task difficulty. Holm-Bonferroni corrections resulted in a Type I error probability of .017 for the familiarity with fallacies as an explanation for perceived task difficulty, .025 for the perceived difficulty to judge the plausibility of the statements, and .05 for the perceived difficulty to allocate weak items to specific fallacies.

On a six-point scale, ranging from *1 = very easy* to *6 = very difficult*, students rated their judgements of the plausibility of statements as significantly more difficult than scientists (M = 3.55, SE = .25 vs. M = 3.00, SE = .20, respectively), $t(37) = 1.72$, $p < .025$, $d = 0.57$. When asked what made the task difficult, they frequently mentioned the complexity of the task (12%), the fact that they had to ignore their own opinion about the content of the text while judging the plausibility (12%), or that they had to rely predominantly on intuitive judgements because they could not think of an appropriate strategy to judge the plausibility (12%). Similarly, students rated the allocation of weak sentences to specific argumentation fallacies as more difficult than scientists (M = 4.55, SE = .29 vs. M = 3.63, SE = .34, respectively), $t(37) = 2.06$, $p < .05$, $d = 0.68$. The most commonly cited reason for this was the lack of familiarity with the argumentation fallacies provided in the task. Of those who perceived the task as difficult (> 3 on the 6-point scale), 56% indicated that they were not entirely familiar with one or more of the argumentation fallacies. Importantly, this explanation was more frequently mentioned by students (74%) than scientists (33%), χ^2 (1, 34) = 5.54, $p < .01$.

Response strategies. Holm-Bonferroni corrections resulted in a Type I error probability of .025 for reported approaches based on the internal consistency of the arguments and .05 for those based on intuition. Analogously to the data derived from the think-aloud protocols, students reported an intuitive approach to judge the plausibility of the items far more often

than scientists (45% vs. 5%, respectively), χ^2 (1, 39) = 8.07, $p < .01$. In contrast, the majority of scientists reported that they focused on the internal consistency of the arguments when forming their judgements (95% vs. 35%, respectively), χ^2 (1, 39) = 15.11, $p < .001$. In addition, some more general strategies in dealing with the task were named by both students and scientists. The strategy most often named by scientists was thorough reading of the statement in a serial fashion (58%); the second most commonly named strategy was repeated reading of the statement (21%), followed by reframing of the sentences (11%). As mentioned above, students mainly relied on an intuitive strategy, followed by serial reading (25%), and repeated reading (10%).

Strategy Use as a Mediator for the Effects of Discipline Expertise on Response Accuracy

We applied multi-level logistic mediation analysis (with test items nested within participants) to test whether the superior performance of scientists in judging the plausibility of arguments was due to their use of more sophisticated strategies. Technically, we estimated Generalized Linear Mixed Models (GLMMs) with a logit link function (Dixon, 2008) and random effects (random intercept) of participants. In contrast to a one-level logistic regression model, this model allows for examining the link between strategies and response accuracy on the item level instead of using aggregated data, which could lead to erroneous conclusions. All models were estimated and tested with the software package *lme4* (Bates, Maechler, Bolker, Walker, Christensen, & Sigmann, 2014). Significance tests were based on a Type I error probability of .05.

Data analysis strategy. For testing the hypothesis that scientists' use of strategies mediate their superior performance, we used both the traditional causal steps approach to mediation analysis (Baron & Kenny, 1986) and the modern approach of computing the indirect effect and testing it for significance with a Monte Carlo technique (MacKinnon, Lockwood, &

Williams, 2004). The causal steps approach tests a mediation of an effect of a distal predictor (e.g., scientists vs. students) on an outcome variable (e.g., response accuracy) by one or several mediators (e.g., strategy use) in three subsequent steps. In Step 1, the distal predictor needs to exert a direct effect on the dependent variable. Step 2 tests the effect of the potential mediator(s) on the dependent variable while the effect of the distal predictor is controlled for. Step 3 re-examines the model estimated in Step 2 with regard to how the direct effect of the distal predictor has changed by including the mediator(s) in the model. A direct effect that is reduced to essentially zero and is no longer statistically significant indicates full mediation, whereas a direct effect that is reduced compared to the original effect but remains significant indicates partial mediation, implying that the mediator(s) do not fully account for the effect of the distal predictor on the dependent variable. Finally, Step 4 establishes that an effect of the distal predictor on the potential mediator(s) exists. Note that in logistic regression, the metric of the variables involved changes when additional predictors are included in the model. MacKinnon and Dwyer (1993) proposed standardizing of the coefficients as a solution to make the scale equivalent across models. We followed this approach here, and only standardized coefficients are reported.

The causal steps approach is useful, because it provides estimates for the paths in the mediational model (the direct effect of the distal predictor and the two paths forming its indirect, mediated effect). However, showing that the distal predictor has an effect on the mediator and that the mediator has an effect on the outcome variable is not sufficient for asserting that an indirect effect exists. In addition, the indirect effect needs to be estimated and tested. The indirect effect itself can be computed as the product of the two effects that it is based on (i.e. the effect of the distal predictor on the mediator and the effect of the mediator on the outcome variable), but several alternative methods have been proposed to test the significance of the effect (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The

Monte Carlo method proposed by MacKinnon, Lockwood, and Williams (2004) was used to test the indirect effect in the present study, because it is particularly suitable for small samples. In the following, we fill first report the results for the causal steps approach (with standardized coefficients, MacKinnon & Dwyer, 1993) and then the results for estimating and testing the indirect effect.

Results of the causal steps approach. In Step 1 of the causal steps approach, we estimated a model with discipline expertise (dummy-coded: 1 = scientists, 0 = students) and plausibility (contrast-coded: 1 = strong arguments, 0 = weak arguments) as predictors and response accuracy as dependent variable. In this model, discipline expertise had a significant positive effect ($\beta = 0.37$, $SE = 0.08$, $p < .001$, one-tailed), reflecting that scientists excelled students in their accuracy of judging the plausibility of arguments (Figure 3; see also Figure 1). Plausibility also had a significant positive effect ($\beta = 0.25$, $SE = 0.07$, $p < .001$, one-tailed), indicating that strong arguments were detected more accurately than weak arguments.

In Step 2, we additionally included the three main response strategies identified in the think-aloud protocols as item-level predictors: Judgements based on intuition, judgements based on knowledge or opinion (content-based judgements), and judgements based on the internal consistency of arguments. All three strategies were included as dummy-coded predictors (1=strategy was applied, 0=strategy was not applied). Of the three types of strategies, only judgements based on the internal consistency of arguments exerted a significant effect ($\beta = 0.23$, $SE = 0.10$, $p < .01$, one-tailed) on response accuracy (Figure 4). More specifically, when participants used this strategy for judging an argument, the plausibility of the argument was judged more accurately (90% accuracy responses) compared to when the strategy was not applied (75% accuracy responses). In Step 3, we examined the change of the effect of discipline expertise induced by including the three strategies as predictors in the model. In the model including strategies, the effect of discipline expertise

was still significant (β = 0.28, *SE* = 0.08, $p < .01$, one-tailed) but weaker than in the model without strategies estimated in Step 1.

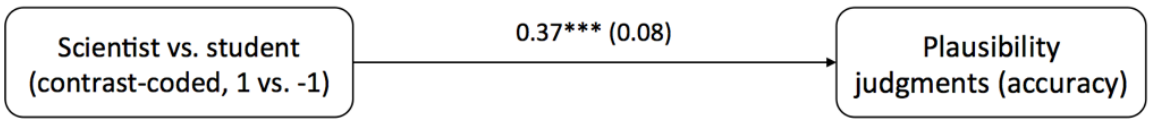

Figure 3. Unmediated effects of discipline expertise on response accuracy in plausibility judgements of arguments. Parameter estimates based on multilevel logistic regression (test items nested within participants) with fixed effects of predictors and random intercept. Plausibility was controlled for as a covariate.

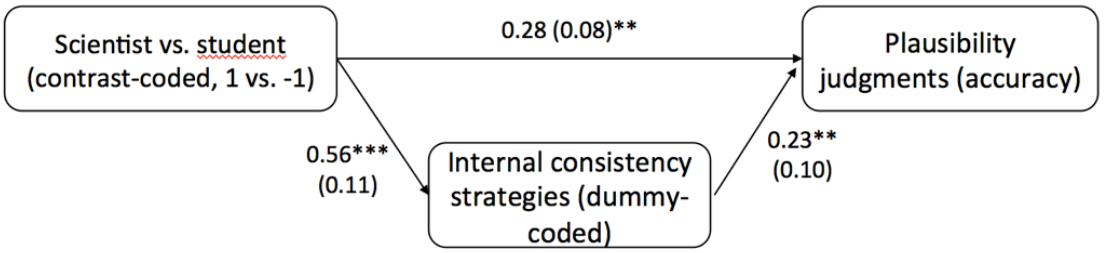

Figure 4. Mediation model for the effect of discipline expertise on response accuracy in plausibility judgements of arguments, with internal consistency strategies as mediator (standardized coefficients, MacKinnon & Dwyer, 2004). Parameter estimates based on multilevel logistic regression (test items nested within participants) with fixed effects of predictors and random intercept. Plausibility and intuition- and content-based strategies were controlled for as covariates.

The pattern of effects obtained in the first two steps suggests that the strategy of judging the internal consistency of arguments might have partially mediated the effects of discipline

expertise on response accuracy. To follow up on this possibility, we performed Step 4 of the causal steps approach and estimated an additional GLMM to obtain the effect of discipline expertise on the strategy of judging the internal consistency. Plausibility was also included as predictor in this model to control for the performance difference between strong and weak arguments. Discipline expertise had a significant positive effect on the use of internal consistency strategies ($\beta = 0.56$, $SE = 0.11$, $p < .001$, one-tailed), reflecting the fact that scientists were far more likely to use these strategies than students (see also Figure 2). Moreover, plausibility had a significant negative effect on strategy use ($\beta = -0.35$, $SE = 0.06$, $p < .001$), indicating that the strategy of judging the internal consistency was used more often in weak compared to strong arguments.

Estimating and testing the indirect effect. Finally, to achieve a complete test of the mediational hypothesis, we computed the indirect effect of discipline expertise on response accuracy (e.g., MacKinnon et al., 2002). A confidence interval for the indirect effect was obtained with the Monte Carlo method (MacKinnon et al., 2004), as implemented in the tool by Selig and Preacher (2008). The indirect effect was estimated as 0.13, with the lower limit of the 95%-confidence interval at 0.02 and the upper limit at 0.26. Thus, the indirect effect was significant, suggesting that the use of internal consistency strategies partially mediated the superior performance of scientists compared to students in judging the plausibility of arguments.

Discussion

The present study examined how scientists and students evaluate the plausibility of arguments embedded in expository texts from the domain of psychology. Results show that students of psychology, compared to scientists working in psychology, have deficits in accurately judging the plausibility of arguments and recognising common argumentation fallacies. Despite the fact that scientists usually receive no systematic training on how to

evaluate arguments, they likely acquire this skill within the process of academic socialisation (Britt et al., 2014). More importantly, the superior performance of scientists was partially mediated by strategy use. Whereas scientists predominantly evaluated the internal consistency of the arguments, students often relied on their intuition or opinion regarding the acceptability of the claim alone. Consistent with these results, students rated all tasks as more difficult and were often not familiar with the argumentation fallacies presented. Finally, the scientists in our study took more time to make their responses, indicating a more systematic approach. Our results are consistent with those found by Shaw (1996) who showed that college students often form assertion-based rather than argument-based judgements.

The present findings can be explained in terms of the mental model theory. When forming assertion-based judgements, constructing a model according to the truth-value of the information presented in the statement is sufficient for the reader. Students' evaluations are likely to involve simple consistency checking between the content of the text and their prior knowledge (Richter, Schroeder, & Wöhrmann, 2009). Our results suggest that their evaluations are predominantly intuitive, or are based on prior attitudes and beliefs rather than a systematic epistemic approach (Britt et al., 2014). Information that is inconsistent with prior attitudes and beliefs is rejected as implausible. In contrast, when forming argument-based judgements, readers need to represent premises, conclusion, and additional information (e.g., alternative explanations) simultaneously, evaluate the degree to which the premises support the conclusion, and qualify or disregard unsupported information. Evaluating the internal consistency of arguments is challenging, because it requires cognitive effort to represent all elements of information simultaneously (Shaw, 1996). Deliberate evaluation also requires structural knowledge, which again requires the activation of an appropriate argument schema prior to and during reading (Britt et al., 2014) to which the

information presented in the statement can be compared. However, lay readers often do not represent different syntactical argument components separately but instead hold a unified mental model of the information presented in the text (Shaw, 1996). When confronted with full-fledged arguments, the students in our study often struggled with the correct allocation of argument components. They experienced the greatest difficulty identifying warrants—probably owing to the fact that relations between reasons and conclusions are often not explicitly stated in commonplace arguments outside scientific discourse. In contrast, scientists were mostly familiar with the structure of arguments. Our results further show that the accuracy of judgements in the plausibility and the argument structure task were positively correlated. Thus, not only might knowledge about the structure of arguments be important for the competent evaluation of arguments, but the competences involved in successful argument evaluation and the systematic approach of identifying argument components might also be part of a common construct of scientific literacy (Britt et al. 2014).

One limitation of the study is that our sample was relatively small. Future studies should provide more stringent tests of the dimensional structure of epistemic competences using item-response models and a larger number of participants. Secondly, the extent that scientists' reasoning skills are general or rather specific to their domain of discipline expertise is not fully clear. More research is needed to examine such transfer effects. Adding materials that are not specific to the participants' domain of discipline expertise might be a useful means to identify possible transfer effects in the improved ability to evaluate arguments. Furthermore, the influence of text characteristics on evaluation processes has been well established (cf. Dole, Duffy, Roehler, & Pearson, 1991). The present study used expository texts from one domain that addressed one topic. The effects of pre-existing opinions on smoking on strategy use cannot be ruled out. With a different topic, they might

have applied different strategies. However, the similar results found by Shaw (1996) might be an indication that effects are not specific to the material used in our study. Future research should also assess how evaluation processes are affected by differences in reasoning ability. Individual differences in reasoning about informal arguments might depend on a broader ability of rational thinking (which seems to be independent of intelligence, Stanovich, 1999, 2012). For example, Chambliss (1995) asked high school students to construct the gist of elaborated arguments and found that successful students were able to identify reasons, conclusions, and even warrants to some extent when texts provided a concluding summary of the argument and contained no misleading information. The scientists in our study were a selected group of former students, and variables such as cognitive ability and motivation likely played an important role in acquiring their expertise. Future research should further examine the precise conditions under which lay readers evaluate arguments, and how the evaluation process is influenced by individual differences in cognitive ability and text materials. Moreover, longitudinal research is needed to reveal the mechanisms through which readers acquire a reasoning schema, and of how argumentation skills develop over the course of education. Our study also provided cross-sectional and correlational data for scientists and students from only one domain. Strictly speaking, causal inferences cannot be drawn from such data. Finally, the study should be replicated in other countries, because culture-specific effects cannot be ruled out in the current findings (cf. Hornikx & Hoeken, 2007; Hornikx & ter Haar, 2013).

Nevertheless, the findings from our study indicate that academic trainings might be helpful to improve the accuracy and the strategies employed in their plausibility judgements. Teaching common argumentation fallacies and some general strategies of argument evaluation to students could be useful means to improve their epistemic competences. Our results suggest that such trainings would likely profit from the inclusion of full-fledged

arguments—including warrants—to focalise relations between premises and conclusions (Toulmin, 1958). Some evidence has shown that even short-term interventions can be effective in improving argument evaluation skills in students (e.g., Hefter et al., 2014, 2015). For example, Larson et al. (2009) found that a tutorial explaining general skills associated with successful argument evaluation to high school and college students led to an increased performance when immediate feedback was provided during training. Including explicit refutations in scientific text seems to be a promising approach to change inaccurate prior misconceptions in university students (Braasch & Wiley, 2013). From an integrated perspective on reasoning and argumentation (Hornikx & Hahn, 2012), the ability to generate argument-based evaluations might depend partly on the ability to generate (counter-)arguments and that both might be instances of a broader ability of argumentative or rational thinking (e.g., Stanovich, 2012). Based on this perspective, fostering argumentation skills could also be a means to (indirectly) improve argument evaluation skills. Generally, focussing on epistemic rather than receptive reading goals helps readers to create a more elaborated and balanced mental model of multiple texts (e.g., Maier & Richter, 2013; Wiley & Voss, 1999). Research by Wiley and Voss (1999) has shown that students who wrote arguments rather than narratives gained a deeper conceptual and causal understanding of the subject matter. Armbruster, Anderson, and Meyer (1991) showed that teaching less competent readers to find the structure in a document improved their general comprehension. Our findings and the reviewed studies suggest the need for systematic training in fostering epistemic reasoning skills.

In sum, the findings from the present study indicate that, although introductory students do validate scientific information against their knowledge and beliefs, their judgements are often erroneous, in part because their strategies are immature. These deficits highlight the usefulness of systematic training of epistemic competences, and a change in the culture of

instruction towards a more argumentative dialogue in classrooms (Kuhn, 1992). Creating an environment that values controversy as a means of fostering understanding seems inevitable to accomplish scientific literacy.

References

Armbruster, B. B., Anderson, T. H., & Meyer, J. L. (1991). Improving content area reading using instructional graphics. *Reading Research Quarterly, 26*, 393–416.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Sigmann, H. (2014). *Lme4: Linear mixed-effects models using Eigen and S4* [Software]. R-package version 1.1-6. Retrieved May 1, 2014 from: http://cran.r-project.org/package=lme4

Blair, J. A., & Johnson, R. H. (1987). Argumentation as dialectical. *Argumentation, 1*, 41–56.

Britt, M. A., & Larson, A. (2003). Construction of argument representations during on-line reading. *Journal of Memory and Language, 48*, 749–810.

Britt, M. A., Richter, T., & Rouet, J.-F. (2014). Scientific Literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104–122.

Chambliss, M. J. (1995). Text cues and strategies successful readers use to construct the gist of lengthy written arguments. *Reading Research Quarterly, 30*, 778–807.

Chi, M. T. H., Bassok, M., Lewis, M.W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Dauer, F. W. (1989). *Critical thinking: An introduction to reasoning*. New York: Oxford University Press.

Dixon, P. (2008). Models of accuracy in repeated measures designs. *Journal of Memory and Language*, *59*, 447–456.

Dole, J. A., Duffy, G.G., Roehler, L. R., & Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research, 61*, 239–264.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review, 100*, 363–406.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fuchs, R., & Schwarzer, R. (1997). Tabakkonsum: Erklärungsmodelle und Interventionsansätze [Tobacco consumption: Explanatory models and approaches for intervention]. In R. Schwarzer (Ed.), *Gesundheitspsychologie: Ein Lehrbuch* (pp. 209–244). Göttingen, Germany: Hogrefe.

Green, D. W. (1994). Induction: Representation, strategy and argument. *International Studies in the Philosophy of Science, 8*, 45–50.

Hefter, M. H., Berthold, K., Renkl, A., Rieß, W., Schmid, S., & Fries, S. (2014). Effects of a training intervention to foster argumentation skills while processing conflicting scientific positions. *Instructional Science, 42*, 929–947.

Hefter, M. H., Renkl, A., Rieß, W., Schmid, S., Fries, S., & Berthold, K. (2015). Effects of a training intervention to foster precursors of evaluativist epistemological understanding and intellectual values. *Learning and Instruction, 38*, 11–22.

Hoeken, H., Timmers, R., & Schellens, P. J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking and Reasoning, 18*, 394–416.

Hoeken, H., & van Vugt (2014). Het bevooroordeelde gebruik van argumentatieschemaspecifieke criteria: Hangt argumentkwaliteit af van het standpunt van de gebruiker? [The biased use of evaluation criteria in argumentation: Does argument quality depend on the user's attitude?]. *Tijdschrift voor taalbeheersing, 36*, 87–105.

Hornikx, J., & Hahn, U. (2012). Reasoning and argumentation: Towards an integrated psychology of argumentation. *Thinking and Reasoning, 18*, 225–243.

Hornikx, J., & Hoeken, H. (2007). Cultural differences in the persuasiveness of evidence types and evidence quality. *Communication Monographs, 74*, 443–463.

Hornikx, J., & ter Haar, M. (2013). Evidence quality and persuasiveness: Germans are not sensitive to the quality of statistical evidence. *Journal of Cognition and Culture*, *13*, 483–501.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

HyperCam (Version 2.28.01) [Computer software]. Retrieved from http://de.hyperionics.com/

Inquisit (Version 3.0.6.0) [Computer software]. Retrieved from http://www.millisecond.com/

Johnson-Laird, P. N., & Byrne, R. M. J. (1992). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Johnson, B. T., Smith-McLallen, A., Killeya, L. A., & Levin, K. D. (2004). Truth or consequences: Overcoming resistance with positive thinking. In E. S. Knowles & J. A. Linn (Eds.), Resistance and persuasion (pp. 215–233) Mahwah, NJ: Erlbaum.

Kuhn (1992). Thinking as argument. *Harvard Educational Review, 62*, 155–178.

Larson, A. A., Britt, M. A., & Kurby, C. (2009). Improving students' evaluation of informal arguments. *Journal of Experimental Education, 77*, 339–365.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*, 99–128.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition & Instruction, 3*, 251–283.

Maier, J., & Richter, T. (2013). Understanding conflicting information on social science issues. *Journal of Media Psychology, 25*, 14–26.

Mayer, R. (1989). Models for understanding. *Review of Educational Research, 59*, 43–64.

MacKinnon, D.P., & Dwyer, J.H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review, 17*, 144–158.

O 'Brien, E. J., & Myers, J. L. (1999). Text comprehension: A view from the bottom up. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative Comprehension, Causality, and Coherence* (pp. 35–54), Mahwah, NJ: Lawrence Erlbaum Associates.

Perkins, D. N. (1986). *Knowledge as design*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 83–106). Hillsdale, NJ: Erlbaum.

Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology, 96*, 538–558.

Rouet, J. -F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction*, *15*, 85–106.

Schroeder, S., Richter, T., & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language, 59*, 237–255.

Shaw, F. W. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning, 2*, 51–80.

Selig, J. P., & Preacher, K. J. (2008, June). *Monte Carlo method for assessing mediation: An interactive tool for creating confidence intervals for indirect effects* [Computer software]. Available from http://quantpsy.org/.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 343–365). New York: Oxford University Press.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, MA: Cambridge University Press.

Van den Broek, P. W., Risden, K., & Husebye-Hartman, E. (1995). The role of readers' standards for coherence in the generation of inferences during reading. In R. F. Lorch, Jr., & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 353–373). Hillsdale, NJ: Erlbaum.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Van Eemeren, F. H., Garssen B., & Meuffels, B. (2009). *Fallacies and judgements of reasonableness: Empirical research concerning the pragma-dialectical discussion rules*. Dordrecht: Springer.

Van Eemeren, F. H., Garssen, B., & Meuffels, B. (2012). The disguised abusive ad hominem empirically investigated: Strategic manoeuvring with direct personal attacks. *Thinking and Reasoning, 2012, 18* (3), 344–364.

van Eemeren, F. H., Grootendorst, R., Henkemans,, F. S., Blair, J. A., Johnson, R. H., Krabbe, E. C. W., & Zarefsky, D. (1996). *Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments*, Mahwah, NJ: Lawrence Erlbaum Associates.

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*, 337–350.

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). The use of source-related strategies in evaluating multiple psychology texts: A student–scientist comparison. *Reading and Writing*.

Weinstock, M. P., Neumann, Y., & Glassner, A. (2006). Identification of informal reasoning fallacies as a function of epistemological level, grade level, and cognitive ability. *Journal of Educational Psychology, 89*, 327–341.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301–311.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the Myside Bias in written argumentation. *Written Communication, 26*, 183–209.

Wu, Y. -T., & Tsai, C. -C. (2007). High school students' informal reasoning on a socio-scientific issue: Qualitative and quantitative analyses. *International Journal of Science Education, 29,* 1163–1187.

Wyatt, D., Pressley, M., El-Dinary, P. B., Stein, S., Evans, P., & Brown, R. (1993). Comprehension strategies, worth and credibility monitoring, and evaluations: Cold and hot cognition when scientists read professional articles that are important to them. *Learning and Individual Differences, 5*, 49–72.

Table A1

Examples of Plausible and Implausible Items

Sentence type	Sentence
Plausible	The concept of nicotine sensitivity is used to explain why some people do not become addicted even though they have already smoked a considerable number of cigarettes.
Implausible	
Contradiction (false conclusion)	The proportion of teenagers smoking occasionally or regularly decreases towards the end of adolescence from 80% to 50%, which implies that there is an increasing interest in smoking over the course of youth.
Wrong example	Negative attitudes towards smoking are often reinforced in children, for example when the father displays how much he enjoys his cigarette after dinner.
False dichotomy	The expectation that smoking alleviates stressful situations leads more likely to smoking behaviour in these situations than the expectation that smoking is helpful in situations of high strain.
Circular reasoning	The concept of inherited nicotine sensitivity relates to the fact that some people react more sensitively to nicotine, because they are more susceptible to nicotine.
Overgeneralisation	It was found that students of a catholic boarding school started smoking when they did not make friends, thus, the (missing) influence of the peer group is generally an important factor to start smoking.

Table A2

Probing techniques applied in the interview

Probing technique	Question(s)
General probing	Was the instruction clear to you? / On a scale ranging from 1 = *not difficult at all* to 6 = *very difficult*, how difficult was it for you to judge the plausibility of the arguments?/ … to allocate weak arguments to a specific argumentation fallacy? / Do you have any further comments on this task?
Category selection probing	If it [the instruction] was not [clear], what was unclear? Why did you find it very easy / easy / rather easy / rather difficult / difficult / very difficult to judge the plausibility of the arguments / … to allocate weak arguments to a specific argumentation fallacy?
Special comprehension probing	What constitutes a plausible argument? Explain in your own words.
Information retrieval probing	How did you proceed when judging the plausibility of the arguments? / How did you proceed when allocating weak arguments to a specific argumentation fallacy?

Chapter VI

Study 2

The Use of Source-Related Strategies in Evaluating Multiple Psychology Texts: A Student-Scientist Comparison

A version of this chapter was published in:

von der Mühlen, S., Richter, T., Schmid, S., Berthold, K., & Schmidt, E.M. (2016). The use of source-related strategies in evaluating multiple psychology texts: A student-scientist comparison. *Reading and Writing, 8*, 1677-1698.

Abstract

Multiple text comprehension can greatly benefit from paying attention to sources and from using this information for evaluating text information. Previous research based on texts from the domain of history suggests that source-related strategies are acquired as part of the discipline expertise as opposed to the spontaneous use of these strategies by students just entering a field. In the present study, we compared the performance of students and scientists in the domain of psychology with regard to (a) their knowledge of publication types, (b) relevant source characteristics, (c) their use of sources for evaluating the credibility of multiple texts, and (d) their ability to judge the plausibility of argumentative statements in psychological texts. Participants worked on a battery of newly developed computerised tests with a think-aloud instruction to uncover strategies that scientists and students used when reading a text. Results showed that scientists scored higher in all of the assessed abilities and that these abilities were positively correlated with each other. Importantly, the superior performance of scientists in evaluating the credibility of multiple texts was mediated by their use of source information. Implications are discussed in terms of discipline expertise.

A common way of learning about a scientific topic is to read various texts. Students entering university are confronted with a variety of documents that present different, at times, conflicting theories backed up by more or less compelling evidence. In addition, the rise of the Internet has exposed students to a variety of information sources, and they seem to use the Internet increasingly for educational purposes (Flanagin & Metzger, 2001). Thus, the question arises of whether university students are able to process this information properly.

To successfully comprehend and evaluate scientific information, readers must be able to understand, integrate, and critically evaluate complex information presented in different texts (Britt & Rouet, 2012). Most searches, however, are time limited, and it is not possible to consider all aspects of the documents that are encountered. Imagine the typical situation of a student faced with the decision of selecting a rather small number of relevant documents from a very large number of possible sources for a class assignment. The student cannot read every document in detail but opts instead to inspect the documents more globally. Evaluating the credibility of a text not only by its content but also by characteristics of its source (e.g., text genre, publication outlet, or date of publication) is crucial in this regard, because it enables the reader to distinguish reliable sources from those possibly conveying inaccurate or biased information. Many documents, particularly those found on the Internet, contain interpretively misleading or incorrect information (Chung, Oden, Joyner, Sims, & Moon, 2012), and enquiring information on the source can help to form a more critical attitude towards the content of such information.

The strategies required to successfully evaluate scientific literature are usually not formally taught to students, whereas scientists are expected to possess these strategies. How scientists acquire these skills, however, is not fully understood. They work in an environment that values controversies as a means of fostering and advancing understanding (Britt, Richter, & Rouet, 2014). Reading and evaluating a broad range of scientific texts is

part of their daily activities, and they are usually familiar with multiple types of documents (Rouet, Favart, Britt, & Perfetti, 1997). It seems likely that they acquire these strategies implicitly in the course of their academic socialisation. In contrast, having relied mainly on textbooks during high school (Luke, de Castell, & Luke, 1989; Paxton, 1997) with little familiarity with other types of scientific texts and little experience on how to interpret and evaluate information from different texts (Britt et al., 2014), introductory university students will likely find the sudden exposure to multiple scientific documents challenging, and as a result, struggle with their evaluation.

In this article, we examine first-year university students' abilities to evaluate multiple documents, and in particular their use of source information. Our objective was to identify successful strategies for the evaluation of multiple documents by comparing their performance with the performance of scientists with several years of academic training (advanced doctoral students and beyond). Evidence will be presented comparing students' versus scientists' strategies to evaluate arguments and to compare their use of source information when evaluating text information. The concept of discipline expertise will be used as a framework to explain differences between scientists and students, and the concept will serve as a basis to make assumptions about the present research.

Assessing Reading Literacy

The PISA 2015 reading literacy framework defines reading literacy as "understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2013, p. 9). When reading a text, people construct a mental representation of the situation described in the text and integrate the text's meaning with prior knowledge (*situation model*, Kintsch, 1988). Across multiple documents, representation usually also includes an integrated understanding about source features (e.g., author, document type). The *documents*

model framework was developed to elucidate the cognitive processes involved in multiple text comprehension (Britt, Perfetti, Sandak, & Rouet, 1999). The rise of the internet, the rapid expansion of access to various information sources, and changing learning environments have yielded major challenges in research on multiple-source reading comprehension (Goldman & Oostendorp, 1999), which have continued to the present day (Bromme & Goldman, 2014). In the U.S., the Common Core State Standards (CCSS) alert readers and educational institutions to the need for integrating and evaluating information from multiple sources (Blanchard & Samuels, 2015). Two assessment groups have begun to develop test items for CCSS, the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC), encouraging students to understand, integrate, and critically evaluate information across documents. For example, in the PARCC, high school students need to integrate information from various sources into a coherent understanding of a situation. Similarly, in the SBAC, 11th graders need to judge the credibility, relevance, and accuracy of information across various sources.

Despite these efforts to understand, integrate, and evaluate information, previous research on learning and teaching reading literacy has predominantly focused on text comprehension, for example, inference-making and elaboration on the content of information stated in a text (e.g., Alexander & Fox, 2011; Brooks, 2011; Duke & Carlisle, 2011; OECD, 2013) rather than the evaluation of texts. Thus, more research is necessary to meet the requirements of a changing learning environment. Epistemic processing goals (i.e., goals that involve forming a valid understanding of the situation, require evaluating text information) and *epistemic strategies* are relevant for this purpose (Richter, 2003; Richter & Schmid, 2010). We propose that for scientific learning to be successful, a critical evaluation of how facts have been established is essential, because such an evaluation not only promotes the construction of a more elaborated situation model, it serves as the basis for

taking a reflective stance toward the issue at hand (Mayer, 1989). Moreover, previous research on reading literacy focuses on the evaluation of content rather than using source information as a criterion for evaluation (Blanchard & Samuels, 2015; Kamil, Pearson, Moje, & Afflerbach, 2011; Brooks, 2011; OECD, 2013), although paying attention to source information is commonly known to be important for selecting relevant and reliable information about a topic (e.g., Korpan, Bisanz, Bisanz, & Henderson, 1997). Finally, existing research on students' use of source information usually focuses on systematic rather than heuristic processing strategies (e.g., Bazerman, 1985; Lundeberg, 1987; Moje, Stockdill, Kim, & Kim, 2011; OECD, 2013; Wineburg, 1991). For example, in the PISA 2015 reading literacy framework, the processes of reflection and evaluation involve the ability to use general and specific knowledge to evaluate the content of the text, evaluate the strengths of arguments, and apply knowledge of the structure (OECD, 2013). Similarly, programs such as the PARCC and SBAC focus on the careful inspection of content across texts.

In contrast to the majority of extant research, the present study focuses on the use of epistemic strategies used to evaluate scientific texts. Readers possess a broad number of systematic and heuristic processing strategies (Wyatt et al., 1993), which they use interchangeably depending on the processing goal (Pressley, 2000; Rouet & Britt, 2011). The present study will focus on heuristic strategies associated with a successful document evaluation.

Systematic and Heuristic Epistemic Strategies when Processing Scientific Texts

To achieve an epistemic processing goal, both systematic and heuristic strategies can be applied. *Epistemic-systematic strategies* are used to carefully and strategically evaluate the plausibility of information presented in the text. In psychology, plausibility can be defined as the "acceptability or likelihood of a situation or a sentence describing it" (Matsuki

et al., 2011, p. 926). When presented with an argument, readers holding an epistemic-systematic processing goal might activate prior knowledge of the structure of an argument and evaluate its strength by evaluating the relevance of the premises for a conclusion (Richter and Schmid 2010). In contrast, *epistemic-heuristic strategies* are applied to form a quick and effortless (preliminary) judgement about the credibility of information presented in a text. These strategies are particularly important when systematic processes cannot be applied, for example, when relevant domain-specific prior knowledge is lacking or when motivational and cognitive resources are limited (Petty & Wegener, 1999; Richter, Schroeder, & Wöhrmann, 2009; Schroeder, Richter, & Hoever, 2008). Paying attention to source information is particularly relevant in this regard (e.g., Korpan, Bisanz, Bisanz, & Henderson, 1997).

Students' Evaluation of Informal Arguments

The epistemic-systematic evaluation of a text requires readers to evaluate the validity of arguments (Britt et al. 2014; Richter and Schmid 2010). Arguments differ from simple explanations in that they provide support for a theoretical assumption, an observation, or factual statement (Britt et al., 2014; Voss & Means, 1991). According to Toulmin's (1958) model of argumentation, a typical argument provides at least a *claim* and one or more theoretical, factual, or empirical reasons (*data*) to support the claim. Although not always explicitly stated in everyday reasoning, complete arguments also contain a proposition or justification of why the data presented are relevant for the claim (*warrant*), additional support for the warrant (*backing*), and one or more counterarguments or limitations (*rebuttals*). Consider the following example (Toulmin, 1958, p. 94):

> Harry was born in Bermuda. A man born in Bermuda will generally be a British subject, on account of the British National Acts. Therefore, Harry is presumably a British subject, unless he has become a naturalised American.

The claim that Harry is a British subject is supported by the datum that Harry was born in Bermuda. The datum lends support to the claim only on account of the warrant that a man born in Bermuda will generally be a British subject. Backing evidence for the warrant is stated by referring to the British National Acts. However, the argument is only conclusive if Harry has not changed his nationality since birth. This sentence constitutes the rebuttal. Scientific documents usually contain all of the elements described above. Experienced readers are able to build a correct structural representation of arguments, and to readily access relevant information when they evaluate the validity of arguments (Britt & Larson, 2003; Wolfe, Britt, & Butler, 2009).

Although students engage in such evaluations to some extent, extant research shows quite unequivocally that the evaluation often fails to meet normative standards. For example, university students often exhibit a certainty bias regarding truth status and confuse cause and correlation (Norris, Phillips, & Korpan 2003).

Students' Use of Source Information

When systematic processes cannot be applied, using heuristics, such as paying attention to characteristics of the source, is particularly important (Petty & Wegener, 1999). Sourcing skills are also crucial for the evaluation of multiple documents (Bråten, Strømsø, & Britt, 2009; Bråten, Strømsø, & Salmerón, 2011; Goldman, Braasch, Wiley, Graesser, & Brodowinska, 2012). Learning to successfully manage multiple information sources is vital for selecting reliable and accurate documents and for reducing information overload (Strømsø, Bråten, & Britt, 2010).

Prior research, mostly from the domain of history, suggests that scientists make extensive use of source-related strategies and use relevant criteria to judge the credibility of different documents (Bazerman, 1985; Lundeberg, 1987; Wineburg, 1991; Wyatt et al., 1993). For example, a historian will probably prefer to consult an article published in an

academic journal rather than a popular science magazine when enquiring information about a historical topic. The article published in the academic journal has been through the peer-reviewed process by other scientists from the same domain and will usually provide a greater number of (alternative) explanations and more elaborated descriptions and discussions of the methodology applied, allowing the reader to reconstruct the conclusions drawn from results. In addition, the historian may also enquire about other aspects of the document source, such as the quality of the journal, author competence and potential biases, or the topicality of the publication, and use this information to make further inferences about its content (Wineburg, 1991). Thus, based on expectations associated with the documents, documents activate a set of textual schemata (Anderson, 1977), enabling scientists to form a first impression about the credibility of the document. Wineburg (1991) refers to this process as *sourcing heuristic*.

Wineburg (1991) reported that historians regard the source of each document as key information in determining its credibility and use this information to make inferences about its content. He asked historians and above-average high school students to rate historical documents while thinking aloud and found that almost all historians paid attention to the source (e.g., text genre, author, date of publication) before reading the content of the text, and they used this information to draw inferences about the content. Similarly, Lundeberg (1987) instructed law professors to think-aloud while reading legal cases. She found that the scientists in this study not only reported information about the source (e.g., the name of the judge, the type of court, the nature of the parties [individual or company]) but used this information to make evaluative judgements about the credibility of the text.

In contrast, high school students reading the same texts in Wineburg's (1991) study and undergraduates in Lunderberg's (1987) study often neglected information about the sources of the documents. Only 31% of students in Wineburg's (1991) study looked to attributions of the source of the document before they read historical documents compared to

98% of historians. In addition, the students in Wineburg's (1991) study rated textbooks to be more credible than primary sources (cp. Paxton, 1997). Similarly, several studies have shown that high school and college students infrequently attend to source information spontaneously (e.g., Britt & Aglinskas, 2002; Korpan, Bisanz, Bisanz, & Henderson, 1997; Metzger, Flanagin, & Zwarun, 2003, Rouet, Britt, Mason, & Perfetti, 1996; Rouet et al., 1997; Wiley et al., 2009). For example, Britt and Aglinskas (2002) showed that 11th graders and undergraduates failed to identify or evaluate source information when reading multiple historical texts, although undergraduates used slightly more source information.

Rouet et al. (1996) observed that undergraduate students who were able to read primary documents in addition to other document types, rated primary documents just as credible as textbooks and more credible than other document types (e.g., historian essays), whereas students who were not given primary documents rated the textbooks as the most credible source. However, undergraduate students in the study justified their credibility rankings most often by the content of the documents. In contrast, more experienced graduate students based their ratings more often on document type. In addition, graduate students rated primary sources as the most credible source in contrast to undergraduates' strong trust in textbooks. These results intuitively suggest that an increase in domain knowledge and familiarity with different document types may lead to increased source awareness. The results from Rouet et al.'s (1997) study suggest that this effect occurs irrespective of the domain type. Graduate psychology and graduate history students (i.e. discipline specialists) who rated the usefulness and credibility of the same historical documents applied similar processing strategies. Thus, the strategies associated with multiple document evaluation are likely to develop as a result of a more general form of discipline expertise. The studies by Rouet et al. (1996, 1997) differ somewhat from previous studies, because they defined a rather short time limit for the evaluation of the documents (15 minutes for the evaluation of

seven short texts), which might have induced a heuristic processing goal, leading more experienced students to focus on aspects of the source rather than content. Therefore, the use of source information is likely to be particularly relevant in the context of a heuristic evaluation of documents.

In sum, previous research based mainly on texts from the domain of history suggests that strategies involved in the evaluation of multiple texts are acquired as part of the discipline expertise rather than being used spontaneously by students just entering a field. Moreover, multiple text comprehension seems to benefit not only from the systematic evaluation of arguments but also by paying attention to sources and from using this information for evaluating text information, which might be especially important when the processing goal is heuristic.

The Present Study

Against this background, the present study examined psychology students' ability to evaluate the credibility of multiple science-related texts in the domain of psychology under a heuristic processing goal and the extent to which they used source-related information to form their judgements. To examine which strategies are particularly relevant for successful document evaluation, students' performance on a computerised test was compared with the performance of scientists working in psychology. The present study extends prior research by assessing university students' use of source information in the domain of psychology, focusing on text evaluation rather than comprehension and by examining separately the use of systematic versus heuristic processing strategies.

Employing a think-aloud procedure, we aimed at assessing whether specific strategies mediate the performance of scientist and student readers. Scientists and students were also tested regarding their knowledge of publication types. Finally, systematic competences involved in the evaluation of informal arguments were investigated by asking

scientists and students to judge the plausibility of argumentative statements and to identify common argumentation errors. In this way, we were able to examine the relationship between systematic and heuristic competences.

Drawing on the assumption that scientists acquire both epistemic systematic and heuristic strategies as part of their academic socialisation (Britt et al., 2014), we expected scientists to score higher than students in all of the assessed abilities. We also proposed that scientists' use of relevant source characteristics would explain their superior performance in making credibility judgements. That is, we expected scientists to be more accurate in identifying text genres, as familiarity with different text genres is a major and generic source characteristic for judging the credibility of science-related texts (Rouet et al. 1997). In addition, we expected the performance in this task to be correlated with the accuracy of credibility judgements. Finally, based on the assumption that the assessed abilities form a unitary construct of discipline expertise (Rouet et al., 1997), we expected the performances in heuristic and systematic tasks to be positively correlated with each other.

Method

Participants

Twenty first-year psychology students and 20 scientists (8 postdocs and 12 advanced doctoral students who were at least in their third year of their doctoral studies in psychology and were close to graduation) participated in the study (students: 80% female; scientists: 74% female). The average age of students was 21.7 years (SD = 4.18) and the average age of scientists was 30.8 years (SD = 5.08). Participants provided informed consent at the beginning of the experiment and were reimbursed with course credits or financial remuneration (25 Euros per hour for scientists and 8 Euros per hour for students) after its completion.

Materials

Scientists' and students' performances were compared using a battery of computerised heuristic and systematic epistemic tests with each test consisting of two parallel versions. Different documents were used in both tasks, because independent tests were necessary to examine heuristic (credibility judgements) and systematic (plausibility judgements) competences separately.

Credibility judgements. The documents used for the credibility judgements were 14 excerpts from authentic scientific texts on the topic of partnership and depression (seven in each parallel version of the test; cf. Rouet et al., 1996, 1997). Documents were selected to ensure that there was a noticeable range of various source features. For example, the documents belonged to different genres and were selected to represent different types of source materials that university students typically encounter. Two texts were taken from each of the following genres: peer-reviewed original empirical articles, review-articles, textbooks, monographies, popular science books, and popular science articles. Two documents (edited books, text books) were written in a more neutral format; the remaining texts were more argumentative in nature (monographies, original empirical articles, review articles, popular sciences texts), although the writing style in the popular science texts was more strongly subjective. Although each document contained partly overlapping information with another document (e.g. "the divorce rate has increased over the last decades" or "symptoms of depression include sadness, joylessness and loss of interest"), most of the information presented in the texts was unique (i.e., different aspects of partnership or depression). Given the aims of the study, the objective of the task was to elicit heuristic rather than systematic processing strategies and to shift attention towards features of the source rather than content. Forming links between documents based on content was not central for making accurate credibility judgements. Author information was presented at the

beginning of the excerpts. References were included in all but the popular science articles. The covers of the books and journals were not included to ensure structural comparability of the texts and to prevent too much attention to layout. The text excerpts were comparable in length (range of 3-4 pages, average length 1,454 words excluding references).

Credibility of the documents was rated by three scientists (full professors), ICC(2,k) = .84. The texts were sorted in a normative rank-order, allowing quasi-paired comparisons (cf. WLST 7-12, Schlagmüller & Schneider, 2007). Each possible pair of texts was scored either 0 (reverse order compared to the normative rank order), 1 (both texts are assigned equal rank), or 2 (the order matches the normative rank order), resulting in the following rank order: 1 = peer-reviewed original empirical article; 2 = review article; 3 = edited book; 4 = text book/monography; 5 = popular science book; 6 = popular science article.

Credibility of the texts under consideration was rated by the participants after presentation of each text on a six-point scale ranging from 1 = *not credible at all* to 6 = *very credible* (Flanagin & Metzger, 2000). In addition, participants were given the opportunity to note down the criteria they had used to make their judgement in a text box. Finally, the importance of a criteria selection for credibility (i.e., line of argumentation, writing style, author information, text genre, structure, own opinion, title, presence of figures, presence of tables, layout, objectivity, quality and quantity of references) was rated on a six-point scale ranging from 1 = *not important at all* to 6 = *very important* (cp. Britt & Aglinskas, 2002; Goldman & Bisanz, 2002; Rouet et al., 1996). A time limit of one minute was set for the credibility ratings of each text. The reason for this rather short time limit was to reveal those strategies associated with a heuristic processing goal.

Genre identification. Knowledge of publication types was measured by asking participants to allocate each document to a genre. For each document, they were given a selection of ten possible genres (textbook, popular science article, popular science book,

edited book, peer-reviewed empirical journal article, peer-reviewed review article, monography, news release, encyclopaedia, and book of abstracts), including three additional items which were not used in the credibility task (i.e. news release, encyclopaedia, and book of abstracts). For each text, a time limit of 30 seconds was set to determine its genre.

Plausibility judgements. Two expository texts were provided for the plausibility judgements (one in each version of the test, see below), the content of which were theories on smoking behaviour (371 words in Text 1, 394 in Text 2; adapted from Fuchs & Schwarzer, 1997, and Schroeder et al., 2008). The texts contained 22 items including plausible and implausible statements. Five sentences in each version were implausible, that is, they contained an argumentation error. Implausible sentences were created by weakening the justification of an argumentative statement and inserting one of five common argumentation errors (Dauer, 1989). The following argumentation errors were included: circular reasoning, false conclusions, overgeneralisations, false dichotomies, and incorrect examples. Plausible and implausible sentences were selected by a quasi-random procedure. Both types of sentences were comparable in features such as length or semantic complexity. Plausible and implausible sentences had a mean length of 3.4 clauses (Text 1) and 3.5 clauses (Text 2). Moreover, they had similar readability scores (32 for the plausible sentences vs. 38 for the implausible sentences as indexed by the German adaptation of Flesch's Reading Ease Index, Amstad, 1978; Flesch, 1948).

Pilot-testing of text materials and items. All texts and items were pretested with 101 introductory psychology students in a pilot study in which response accuracies in the different tasks were the dependent variables. The pilot study served as a basis to select suitable items for the final test battery. Reliability was calculated separately for the two parallel versions of the tests. For the credibility task, Cronbach's α was .81 in version 1 and .75 in version 2. The correlation between the two test versions was $r = .48$, $p < .05$. For the

plausibility judgements, Cronbach's α was .64 in version 1 and .94 in version 2. The correlation between the two test versions was r =.49, p < .01. In addition, the texts used for the plausibility judgements had been normed and validated in a study by Schroeder et al. (2008) who found that plausible and implausible sentences differed in their plausibility. The rating scale for the credibility judgements was also tested in the pilot study and achieved a high reliability score, Cronbach's α = .96. Likewise, the rating scale for the importance of the criteria for credibility achieved high reliability, Cronbach's α = .95.

In addition, both the time limit for the credibility ratings and the time allowed to identify the text genre were pretested in the pilot study using retrospective cognitive interviews. Originally, the time limit for the credibility ratings was set for 3 min. However, the majority of participants (84%) in the pilot study reported that this was enough time to engage in systematic processing. In addition, 10 doctorate students completed the tasks and came to a similar result (80% of doctorate students indicated that the time allowed systematic processing). Reducing the time limit, therefore, seemed plausible to change their processing strategy to a more heuristic one. We determined the appropriate time limit for the text genre identification to be 30 s. The majority (81%) of participants in the pilot study reported that this was enough time to identify the genres of the texts, regardless of whether or not they were able to identify them correctly. Similarly, all but one doctoral student stated that the time limit was adequate.

Software. The software used to display the texts and collect data was a combination of the programs Inquisit 3.0.6.0 and Simple Learning Environment Developer (SLED). The software package HyperCam 2.28.01 was used to record screen activity (e.g., mouse movement, clicking). Recording screen activity was important to relate utterances in the think-aloud protocols to the part of the text on which participants were working.

Think-aloud protocols

Think-aloud protocols were obtained during task completion in one version of the test. All participants worked on one version of the test in silence and on the other while thinking aloud. In the think-aloud version, participants were instructed to say "everything that comes to mind" aloud while they were working on the tests. In particular, they were asked to think aloud persistently throughout the inspection of the texts and while they made their judgements so that online measures could be recorded and also to prevent disruption of a continuous flow of thought. Half of the participants received the silent version first, the other half the think-aloud version. Participants were audiotaped while wearing a headset. Because of text length restrictions and because the systematic task was not part of the central question addressed in the current paper, only the results of the credibility judgement protocols are reported in the present article.

Transcription. The audio-recordings were transcribed verbatim. From the total number of transcriptions, 30% were cross-checked for accuracy, indicating 100% accuracy. The wordings were entered into different cells of a Microsoft Excel sheet. Each cell contained a think-aloud protocol from one text. For every text, the credibility criteria mentioned by the participant during text inspection (e.g., different source characteristics, argumentation, methods) were noted and coded for analysis.

Coding. Based on theoretical approaches on scientific genres (Goldman & Bisanz, 2002) and research on multiple document evaluation (Bråten et al., 2009; Bråten et al., 2011; Britt & Aglinskas, 2002; Rouet et al., 1996), a coding scheme was derived from the think-aloud protocols regarding strategy use. The coding scheme for the analysis of the think aloud protocols in the credibility task was developed based on criteria which are commonly used to evaluate the credibility of multiple texts. In addition, an inductive approach was applied based on any additional criteria that participants mentioned during task completion. The

coding scheme comprised seven main categories, each based on different credibility criteria: Source characteristics, content, argumentation, writing style, method, structure, and other criteria that could not be classified in one of the above categories. Each main category consisted of several sub-categories, and sum scores of the sub-categories were calculated for each main category. The category *Source information* included use of publication outlet (e.g., quality of journal, conflicts of interest), publication date, text genre (e.g., peer-review), original language of the document (i.e., English or German), and author information (e.g., author expertise, conflicts of interest) as criteria for evaluation. The category *Content* included attendance to the topic of the text, its relevance, the theoretical foundation and the sophistication of information stated in the text, information stated in the abstract, clarity and complexity of title information, and the reader's opinion regarding the topic of the text. The category *Argumentation* included general line of argumentation, theoretical reasons, empirical evidence, and references provided for the claims. Use of *Writing style* comprised the comprehensibility, clarity, and coherence of the way the document was written. The category *Method* included attendance to research methods and statistics. Use of *Structure* included attendance to the topical structure and the general layout of a document. The category *Other* contained all criteria used by readers that could not be classified in one of the above categories. These included the presence of figures and tables in the document and formal aspects such as use of APA norms. The categories were not mutually exclusive, that is, participants could make use of more than one criterion. Therefore, inter-rater reliability was calculated separately for each category (Cohen's Kappa, based on ten randomly selected protocols coded by two coders). Inter-rater reliability ranged from 83. - .95 with an average value of .87, indicating that there was high agreement.

Procedure

Participants were tested individually in a computer lab equipped with notebooks

(15”) and headsets. The text excerpts for the credibility judgements appeared in their original form (see Appendix A for an example). All other texts (including task instructions and test items) were presented in Calibri black 12-point font against a white background, with a visual angle of 1.4 degrees. Two exceptions were the reminders for the keys representing *plausible* and *implausible* response options and the sentences participants had marked implausible that appeared in red font. Each participant completed two parallel versions of the tests; one task was completed in silence and the other while thinking aloud.

Credibility judgements. The texts were presented one at a time on a computer screen in randomised order. Participants were allowed to scroll through the texts. To be able to set a heuristic processing goal, participants were asked to rate the credibility of the texts as quickly as possible, and they were given a time limit for each text. After one minute, the text disappeared automatically, and participants were asked to rate its credibility. Subsequently, participants were requested to indicate the criteria they had used to make their judgement in a text box before the next text was presented. After task completion, participants rated the importance of a number of criteria for their credibility judgements (e.g., quality of argumentation, layout, structure, title information, research methods, style of writing, source information). The completion of both tests took approximately 1 hour.

Genre identification. After completion of the credibility judgements, participants allocated each of the documents presented in the previous task to a specific genre. Again, they were asked to form their decision as quickly as possible, and they were given a time limit of 30 s for each text.

Plausibility judgements. Subsequently, participants were asked to judge the plausibility of different statements in two argumentative texts. They were instructed to read the texts thoroughly. The texts were presented sentence by sentence on a computer screen, and participants were asked to read each statement in a self-paced fashion. Participants

judged the plausibility of each statement by pressing a key for *plausible* or another key for *implausible.* They were instructed to judge the internal consistency and quality of the arguments and not to base their judgements on their opinion or knowledge about the content of the text. Furthermore, they were told that global errors, that is, inconsistencies of a statement with other passages mentioned earlier in the text, were not included. After participants rated the plausibility of all text items, they were instructed to allocate all sentences they had marked implausible to specific argumentation errors, which were explained briefly.

Completion of all tests took approximately 1 hour. At the end of the sessions, all participants were interviewed (~15 min). The interview included questions on task difficulty and general difficulties with primary documents. However, the data derived from the interview were not central to our main hypothesis and are not reported in the results.

Design

The study was based on a design with pre-existing groups defined by different degrees of expertise (scientists vs. students). The test battery included two parallel versions of the test with seven texts in each version in the credibility task and one text in each version in the plausibility task. All participants read half of the texts while thinking aloud and the other half in silence to control for potential effects of thinking aloud. The order of the two versions as well as the assignment of test versions to the silent and the think-aloud conditions were counterbalanced across participants. The order in which the texts appeared was also counter-balanced across participants. Accuracy of answers, responses, and response strategies derived from the think-aloud protocols were used as dependent variables.

Results

Type-I-error probability was set at .05 for all hypothesis tests. One-tailed tests were applied for testing univariate hypotheses that predicted higher values of scientists compared

to students. The hypotheses were tested in a series of analyses. First, we conducted a Multivariate Analysis of Variance (MANOVA) for the response accuracies to avoid the increased likelihood of error rates that could result from multiple comparisons between scientists and students. Univariate follow-up tests were performed to interpret group differences. Second, we used univariate comparisons to test for differences in the use of specific strategies as revealed by the think-aloud data and the retrospective interview. Third, a mediation analysis was conducted to examine the extent that strategy use explains the expected superior performance of scientists as reflected in the response accuracies.

Response Accuracy

The MANOVA revealed a significant main effect of expertise regarding accuracy of answers when credibility judgements, plausibility judgements, and judgements about text genre were taken into account. As expected, students' answers were more often erroneous than those given by scientists, $F(6, 31) = 8\ .19$, $p < .001$, $\eta^2 = .61$. Univariate follow-up analyses were performed to interpret the multivariate group difference.

Credibility judgements. Scientists ($M = 1.69$, $SE = 0.03$), compared to students ($M = 1.52$, $SE = 0.06$), provided more accurate credibility judgements across both text versions, $t(37) = -2.52$, $p < .01$, $d = 0.83$.

Genre identification. Scientists ($M = .80$, $SE = .04$) were much better than students ($M = .46$, $SE = .03$) in identifying the genres of the texts, $t(36) = -6.80$, $p < .001$, $d = 2.26$ (except for text books, $p = .09$). Better performance on this task was positively correlated with the accuracy of the credibility judgements, $r = .44$, $p < .01$.

Plausibility judgements. Scientists ($M = .84$, $SE = .03$) outperformed students ($M = .69$, $SE = .03$) when judging the plausibility of the text items, $t(37) = -3.35$, $p = .001$, $d = 1.28$. In particular, they more often identified implausible sentences correctly ($M = .80$, $SE = .05$ vs. $M = .58$, $SE = .04$, respectively), $t(37) = -3.67$, $p < .001$, $d = 1.46$. There was a

tendency for scientists to identify more plausible sentences correctly (M = .86, SE = .02 vs. M = .81, SE = .03, respectively). However, this difference did not reach significance, $t(37) = -1.11$, $p > .05$, $d = 0.44$. Finally, scientists (M = .60, SE = .04) performed significantly better than students (M = .39, SE = .04) in allocating implausible items to specific types of argumentation errors, $t(37) = -4.03$, $p < .001$, $d = 1.27$.

Response accuracies in this task were positively correlated with the accuracy of the credibility judgements, $r = .58$, $p < .001$, and the accuracy to identify the genres of the texts, $r = .57$, $p < .001$.

Credibility Judgements by Text Genre

Empirical journal articles received the highest credibility ratings by both scientists (M = 5.19, SE= 0.18) and students (M = 5.00, SE = 0.25), whereas popular science articles received the lowest ratings (M = 2.36, SE = 0.13 by scientists vs. M = 2.80, SE = 0.20 by students), $t(36) = 1.79$, $p < .05$, $d = 0.60$. However, scientists provided lower ratings than students for review articles (M = 4.44, SE = 0.16 vs. M = 5.03, SE = 0.17), $t(36) = 2.45$, $p < .01$, $d = 0.83$, and popular science articles (M = 2.36, SE = 0.13 vs. M = 2.80, SE = 0.20), $t(32.4) = 1.83$, $p < .05$, $d = 0.60$. No differences were found between groups on rating of the remaining publication types. Results are displayed in Figure 1.

Response Strategies

The think-aloud protocols revealed significant differences between scientists and students with regard to the strategies they applied to judge the credibility of the texts (see Figure 2). Scientists used source-related criteria more often than students (M = .31, SE = .03 vs. M = .14, SE = .03), $t(34) = -3.98$, $p < .001$, $d = 1.37$. In addition, scientists more often considered argumentation (M = .19, SE = .03 vs. M = .09, SE = 0.2, $t(26.1) = -2.98$, $p < .01$, $d = 1.02$), structure (M = .17, SE = .02 vs. M = .09, SE = .02, $t(34) = -2.99$, $p < .001$, $d = 1.03$), and methods (M = .15, SE = .02 vs. M = .04, SE = .02, $t(34) = -3.61$, $p < .001$, $d = 1.23$) as

criteria in their credibility ratings.

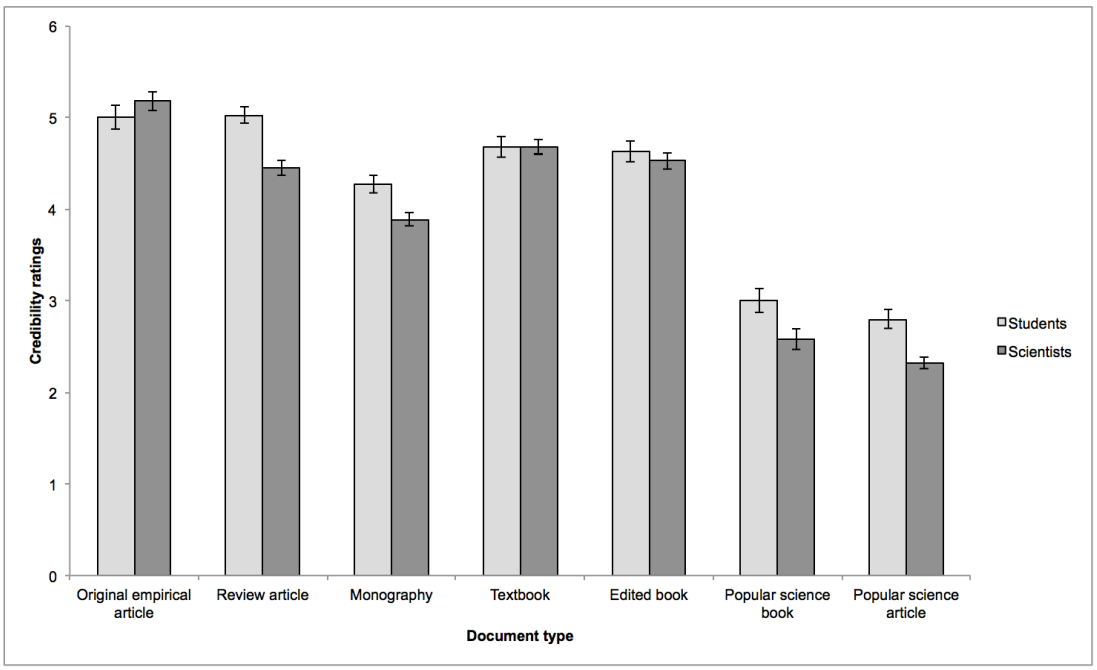

Figure.1 Credibility ratings (with SE_M) applied by scientists and students.

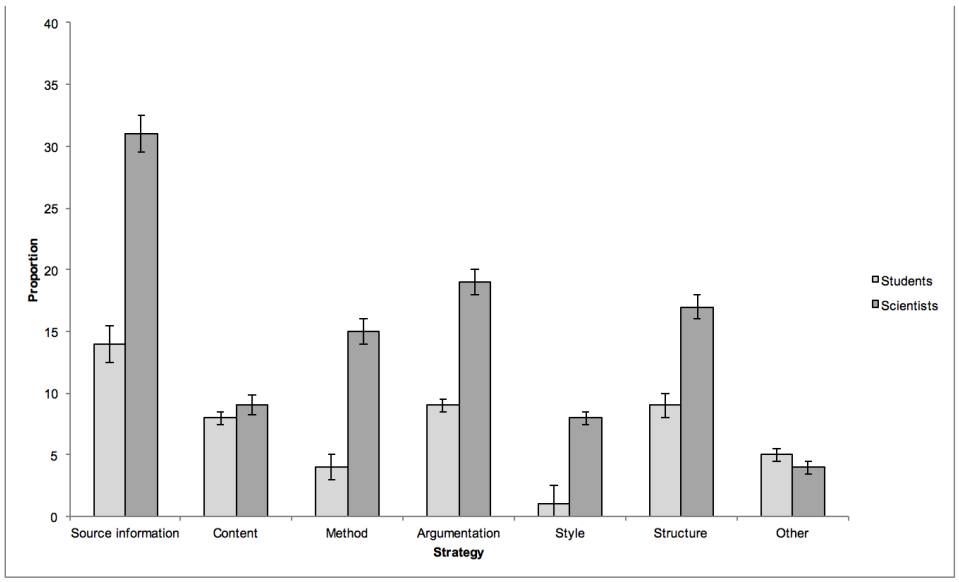

Figure.2 Mean proportions of strategies (with SE_M) for each category applied by scientists and students (strategies are non-exclusive).

The self-reported criteria for credibility were similar to the results derived from the

think-aloud protocols. Scientists gave higher ratings for the importance of source-related criteria (M = 4.56, SE = 0.19 vs. M = 3.71, SE = 0.26), $t(36)$= -2.54, p <.05, d = 0.85) – especially for the text genre (M = 5.14, SE = 0.22 vs. M = 4.13, SE = 0.32), $t(36)$= -2.56, p <.01, d = 0.86) and argumentation (M = 4.17, SE = 0.22 vs. M = 3.21, SE = 0.19), $t(36)$ = -3.35, p <.01, d = 1.12). In particular, ratings differed for quality of references (M = 4.28, SE = 0.33 vs. M = 3.10, SE = 0.30), $t(36)$ = -2.65, p <.01, d = 0.89) and quantity of references (M = 3.78, SE = 0.22 vs. M = 2.93 , SE = 0.27), $t(36)$ = -2.41, p <.05, d = 0.81). In addition, scientists used title information more often than students (M = 4.33, SE = 0.29 vs. M = 3.55, SE = 0.32), $t(36)$ = -1.82, p <.05, d = 0.61. None of the other criteria differed significantly between groups.

Response strategies were positively correlated with the accuracy of the credibility judgements (r = .51, p < .01), the plausibility judgements (r = .54, p < .01), and the accuracy to assign the texts to a genre (r = .48, p < .01). In particular, use of source information was correlated with a higher accuracy regarding credibility judgements (r = .46, p < .01), plausibility judgements (r = .54, p < .01), and the identification of text genres (r = .44, p < .01). It is important to note that use of argumentation and accuracy of answers were also positively correlated (r = .45, p < .01 for the credibility judgements; r = .44, p < .01 for the plausibility judgements; and r = .52, p < .01 for the identification of text genres). Likewise, there was a positive correlation between use of structure as a criterion and accuracy regarding credibility judgements (r = .37, p < .05), plausibility judgements (r = .46, p < .01), and the identification of text genres (r = .48, p < .01).

Strategy Use as a Mediator for the Effects of Expertise on Response Accuracy

In a stepwise regression model with expertise (contrast-coded, -1 = students, 1 = scientists) and the different criteria (based on the think-aloud protocols) as potential mediators, the use of source-related criteria was included first as it explained most of the

variance on the accuracy of the credibility judgements (ΔR^2 = .18). No other strategies were significant predictors (all *p*s > .17). The direct effect of expertise on the accuracy of the credibility judgements that was significant in a model without any strategies as predictors (*B* = 0.09, *SE* = 0.04, $p < .05$, ΔR^2 = .16), was weaker and no longer significant after including the use of source-related criteria in the model (*B* = 0.03, *SE* = 0.04, *p* = .50). Moreover, the indirect effect of expertise via source-related criteria on the accuracy of the credibility judgements was significant (Sobel test: *z* = 2.33, $p < .05$). This pattern of effects indicates that the use of source-related criteria fully mediated the superior performance of scientists in the credibility judgement task. Figure 3 displays the results of the mediation analysis.

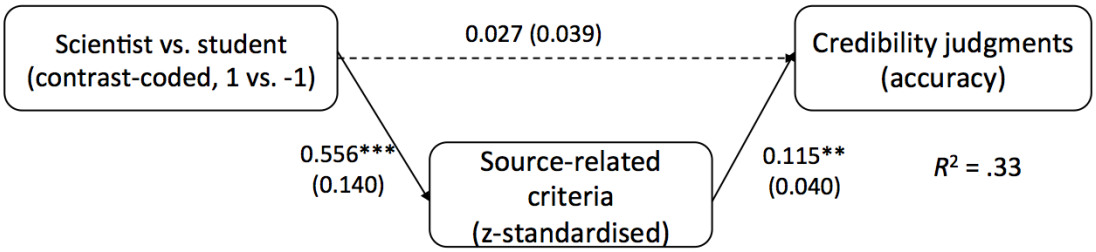

Figure.3 Mediation model for the effect of expertise (contrast-coded, -1 = students, 1 = scientists) on response accuracy in credibility judgements of science-related texts (range 0-2) with source-related criteria as mediator. (Unstandardized coefficients.)

Discussion

The main purpose of this study was to examine psychology students' abilities to evaluate multiple documents and to identify successful evaluation strategies by comparing university students' performance with those of scientists in the domain of psychology. The

results indicate that students had difficulties evaluating the credibility of multiple psychological texts and the plausibility of argumentative statements, and they often failed to attend to source information. In contrast, scientists were able to evaluate the documents in a flexible manner, depending on whether the processing goal was heuristic or systematic. Importantly, the superior performance of scientists in this study was mediated by their strategic use of source information. These findings are in line with prior research on multiple-text comprehension and evaluation (e.g., Britt & Aglinskas, 2002; Lunderberg, 1987; Metzger et al., 2003, Rouet et al., 1996; Rouet et al., 1997; Wiley et al., 2009; Wineburg, 1991) and extend those findings to the domain of psychology. In particular, they provide further evidence that sourcing skills are particularly important for evaluating multiple documents in a situation in which time pressure only allows the use of heuristic evaluation strategies. The fact that both historians and psychologists seem to use similar strategies when they evaluate historic or psychological texts suggests that the competences involved in successful document evaluation may in fact be a relatively generic form of discipline expertise (cp. Rouet et al., 1997). Our results also show that the accuracy of judgements in the systematic plausibility and the heuristic credibility task were positively correlated, suggesting that heuristic and systematic strategies may be regarded as facets of a common construct of scientific literacy (Britt et al., 2014). The finding that systematic and heuristic strategies correlate in addition to the superior performance of scientists in both systematic and heuristic tasks casts doubt on the applicability of two-process models, such as the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986) or the Heuristic-Systematic Model (HSM; Chen & Chaiken, 1999), to the processing of scientific information. According to these models, strategy use is largely dependent on motivation and ability (general cognitive ability and prior knowledge), and heuristic processes are applied primarily when recipients lack the motivation or the ability to process the persuasive

message systematically (i.e., via the central route). Thus, given that scientists possess the relevant abilities, they should mainly use systematic strategies, whereas students should apply mainly heuristic strategies. In contrast to this view, our results clearly show that scientists use systematic and heuristic strategies in a flexible manner depending on the processing goal (Pressley, 2000; Rouet & Britt, 2011). Finally, sourcing strategies were positively correlated with the ability to assign the documents to a genre. Thus, familiarity with different document types may contribute to the extent to which source information is used.

Contrary to prior research, which found that students rated information from textbooks as more credible than other types of documents (e.g., Bråten et al., 2011), students and scientists in our study rated the primary documents as the most credible source. This inconsistency of findings might be explained by the fact that the majority of students in the present study had already worked with primary literature in the context of their psychology education and were therefore already familiar with some aspects of the literature. However, our findings are in line with the results reported by Rouet and colleagues (1997) who found that primary documents were rated just as credible as secondary documents. Nevertheless, the students in our study showed deficits in judging the credibility of the documents correctly, and their lack of sourcing skills seemed to be the crucial factor.

It should be noted that in addition to source-related criteria line of argumentation, document structure, and the methodology applied in reported studies were also regarded as important factors for credibility, and these strategies were also mainly used by scientists. In fact, use of these strategies was – to a somewhat lesser degree than the use of source information – positively correlated with the accuracy of the credibility judgements, of assigning the texts to a specific genre, and of plausibility judgements of argumentative statements. A similar pattern of results was found in the interview. Thus, our results indicate

that scientists considered characteristics of the source *and* the content as vital factors for determining the credibility of scientific documents. However, criteria such as the quality of argumentation, structure, or methodology are probably more relevant under a systematic processing goal. The time allowed to inspect the documents in the credibility task was very limited. Other criteria are likely to be more pronounced when subjects are given more time. The results from the plausibility task demonstrate that scientists were more accurate in determining the validity of arguments in the systematic task, indicating that they were able to select relevant strategies in a flexible manner, depending on the processing goal. Viewing these results from the perspective of the documents model framework (Britt et al., 1999), it seems likely that scientists are able to construct more elaborate representations of content and source, enabling them to select appropriate strategies in a flexible manner and draw inferences about the credibility of the text (cp. Bråten et al., 2009; Wineburg, 1991).

Heuristics related to the source may not always lead to accurate judgements of the credibility of a document. Naturally, a complete evaluation process requires subsequent systematic document inspection. However, our findings suggest that under conditions in which systematic processes cannot be applied, attention to the source seems to be an efficient strategy to determine the credibility of multiple documents and to select more reliable documents in the first place.

Furthermore, the extent to which different strategies are applied to evaluate the credibility of multiple documents varies greatly depending on the texts provided and on how the materials are presented (Britt & Rouet, 2011). For example, all texts presented in the credibility task of the present study were authentic, and none of the arguments provided in those texts were entirely implausible. In addition, the information presented in the documents was supplementary rather than contradictory. We cannot rule out the possibility that line of argumentation would have been a more important criterion in determining the

credibility of the documents if the arguments provided in the texts had been less plausible or if more conflicting evidence had been provided. Future research is needed to explore this possibility. In another example, given that other genres (e.g., narratives) involve different kinds of processing (e.g., Zwaan 1994), results cannot be generalised to these genres. Moreover, only one topic was used for the systematic task. Future research should include other topics and also provide more stringent tests of the dimensional structure of epistemic strategies using item-response models and a larger number of participants, and researchers should consider additional factors associated with successful evaluation skills such as individual differences in cognitive ability. For example, Goldman el al. (2012) showed that better learners showed better evaluation processes than poor learners for reliable documents. Similarly, Bråten et al. (2011) showed that readers low in topic knowledge were more likely to trust less credible sources. Finally, our study provides cross-sectional and correlational data for scientists and students from only one domain, and strictly speaking, causal inferences cannot be drawn from such data.

Nevertheless, the present study clearly indicates that heuristic processing strategies can be as vital as systematic processing strategies and that source information is an important criterion used by scientists when the goal is to evaluate the credibility of multiple documents under time constraints. Scientific literacy may require some general competences which can be used in a flexible manner (cf. Rouet et al., 1997). Results from this and other studies (e.g., Wineburg, 1991; Britt & Aglinskas, 2002) indicate that students at this level often neglect to attend to source information, leading them to possibly select inaccurate information. Thus, instruction should focus on designing appropriate curricula or intervention strategies that would raise students' source awareness. Results from studies including high school and college students have indicated that computer-based training environments may facilitate sourcing skills (e.g., Britt & Aglinskas, 2002). Other evidence (e.g., Nokes, Dole & Hacker,

2007) indicates that reading multiple texts improved high school students' attention to sources. Thus, using multiple texts in higher education, rather than relying exclusively on textbooks (Paxton, 1997), may help to promote students' competences in dealing with scientific literature. Our findings suggest that both heuristic and systematic processes should be considered to be able to create a more complete view of the strategies involved in the evaluation of multiple documents.

References

Alexander, P. A., & Fox, E. (2011). Adolescents as readers. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 157–176). New York: Taylor & Francis.

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen? [How understandable are our newspapers?]*. Unpublished Doctoral Dissertation, University of Zürich, Switzerland.

Anderson, R. C. (1978). Schema-directed processes in language comprehension. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser (Eds.), *Cognitive psychology and instruction* (pp. 67–82). New York: Plenum.

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication, 2*, 3–23.

Blanchard, J. S., & Samuels, S. J. (2015). Common core state standards and multiple-source reading comprehension. In P. D. Pearson, & E. H. Hiebert (Eds.), *Research-based practices for teaching common core literacy* (pp. 93–105). New York: Teachers College Press.

Brante, E. W., & Strømsø, H. I. (2017). Sourcing in text comprehension: A review of interventions targeting sourcing skills. *Educational Psychological Review, 26*, 1–27.

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*, 6–28.

Bråten, I., Strømsø, H. I., & Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learning and Instruction, 21*, 180–192.

Britt, M. A., & Larson, A. (2003). Construction of argument representations during on-line reading. *Journal of Memory and Language, 48*, 749–810.

Britt, M. A., Richter, T. & Rouet, J.-F. (2014). Scientific Literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104–122.

Britt, M. A., & Rouet, J- F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276–314). New York: Cambridge University Press.

Britt, M. A., & Aglinskas, C. (2002). Improving student's ability to use source information. *Cognition and Instruction, 20*, 485–522.

Britt, M. A., Perfetti, C. A., Sandak, R., & Rouet, J.-F. (1999). Content integration and source separation in learning from multiple texts. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 209–233). Mahwah, NJ: Erlbaum.

Bromme, R., & Goldman, S. (2014). The public's bounded understanding of science. *Educational Psychologist, 49*, 59–69.

Brooks, G. (2011). Adult literacy. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol 4, pp. 177–196). New York: Taylor & Francis.

Chen, S., & Chaiken, S. (1999). The heuristic systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social and cognitive psychology* (pp. 73–96). New York: Guilford.

Chung, M., Oden, R. P., Joyner, B. L., Sims, A., & Moon, R. Y. (2012). Safe infant sleep recommendations on the Internet: Let's Google It. *The Journal of Pediatrics*, *161*, 1080–1084.

Dauer, F. W. (1989). *Critical thinking: An introduction to reasoning*. New York: Oxford

University Press.

Duke, N. K., & Carlisle, J. (2011). The development of comprehension. In M. L. Kamil et al. (Eds.), *Handbook of reading research* (Vol. 4, pp. 199–228). New York: Taylor & Francis.

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of Internet information credibility. *Journalism and Mass Communication Quarterly, 77*, 515–540.

Flanagin, A. J., & Metzger, M. J. (2001). Internet use in the contemporary media environment. *Human Communication Research, 27*, 153–181.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–233. doi: 10.1037/h0057532.

Fuchs, R., & Schwarzer, R. (1997). Tabakkonsum: Erklärungsmodelle und Interventionsansätze [Tobacco consumption: Explanations and interventions]. In R. Schwarzer (Ed.), *Gesundheitspsychologie: Ein Lehrbuch* (pp. 209–244). Göttingen, Germany: Hogrefe.

Goldman, S. R., & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. A. León & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 417–436). Mahwah, NJ: Erlbaum.

Goldman, S., Braasch, J. L., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from Internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly, 47*, 356–381.

Goldman, S., & van Oostendorp, H. (1999). Conclusions, conundrums, and challenges for the future. In H. van Oostendorp, & S. Goldman (Eds.), *The construction of mental representations during reading* (pp. 323–330). Mahwah, NJ: Erlbaum.

Goldman, S. R., & Rakestraw, J. A. (2000). Structural aspects of constructing meaning from

text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research, Vol. III* (pp. 311–335). Mahwah, NJ: Erlbaum.

Inquisit (Version 3.0.6.0) [Computer software]. Retrieved from http://de.hyperionics.com/

Kamil, M. L., Pearson, P. D., Moje, E. B., & Afflerbach, P. P. (2011). *Handbook of reading research (Vol. 4)*. New York: Taylor & Francis.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49,* 294–303.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163–182.

Luke, C., de Castell, S. C., & Luke, A. (1989). Beyond criticism: the authority of the school textbook. In S. C. de Castell, A. Luke, & C. Luke (Eds.), *Language, authority, and criticism* (pp. 245–260). London, Falmer.

Lundeberg, M. A. (1987). Metacognitive aspects of reading comprehension: studying understanding in legal case analysis. *Reading Research Quarterly, 22*, 407–432.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology, 37*, 913–934.

Mayer, R. (1989). Models for understanding. *Review of Educational Research, 59*, 43–64.

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003). College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education, 41*, 271–290.

Moje, B. M., Stockdill, D., Kim, K., & Kim, H. (2011). The role of text in disciplinary learning. In M. L. Kamil et al. (Eds.), *Handbook of reading research* (Vol. 4, pp. 453–481). New York: Taylor & Francis.

Norris, S. P., & Phillips, L.M. (2003). How literacy in its fundamental sense is central to

scientific literacy. *Science Education, 87*, 224–240.

OECD (2013). *PISA 2015 draft reading framework*. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.html.

Paxton, R. J. (1997). "Someone with like a life wrote it": The effects of a visible author on high school history students. *Journal of Educational Psychology, 89*, 235–250.

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp, & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mahwah, NJ: Erlbaum.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer.

Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 41–72). New York: Guilford Press.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of Reading Research* (Vol. 3, pp. 545–562). Mahwah, NJ: Erlbaum.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–532.

Richter, T. (2003). *Epistemologische Einschätzungen beim Textverstehen* [Epistemic validation in text comprehension]. Lengerich, Germany: Pabst Science Publishers.

Richter, T., & Schmid, S. (2010). Epistemological beliefs and epistemic strategies in self-regulated learning. *Metacognition and Learning, 5*, 47–65.

Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology, 96*, 538–558.

Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Greenwich, CT: Information Age Publishing.

Rouet, J.-F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88*, 478–493.

Rouet, J. -F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction, 15*, 85–106.

Schlagmüller, M. & Schneider, W. (2007). *Würzburger Lesestrategie-Wissenstest für die Klassen 7-12*. Göttingen: Hogrefe.

Schroeder, S., Richter, T., & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language, 59*, 237–255.

Shaw, F. W. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning, 2*, 51–80.

Strømsø, H. I., Bråten, I., & Britt, M. A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning and Instruction, 20*, 192–204.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, MA: Cambridge University Press.

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning & Instruction, 1*, 337–350.

Wiley, J., Goldman, S. R., Graesser, A.C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal, 46*, 1060–1106.

Wineburg, S. (1991). Historical problem solving: a study of the cognitive processes used in

the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*, 73–87.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication, 26*, 183–209.

Wyatt, D., Pressley, M., El-Dinary, P. B., Stein, S., Evans, P., & Brown, R. (1993). Comprehension strategies, worth and credibility monitoring, and evaluations: Cold and hot cognition when scientists read professional articles that are important to them. *Learning and Individual Differences, 5*, 49–72.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology, 20*, 920–933.

Chapter VII

Study 3

How to Improve Argument Comprehension in University Students: Experimental Tests of Two Training Approaches

A version of this chapter was submitted as:

von der Mühlen, S., Richter, T., Schmid, S., & Berthold, K. (submitted). How to Improve Argumentation Comprehension in University Students: Experimental Tests of Two Training Approaches. *Manuscript submitted for publication.*

Abstract

The ability to comprehend and evaluate informal arguments is essential for scientific literacy but students often lack structural knowledge about these arguments, especially when the arguments are more complex, and struggle with the reasoning processes involved in a successful evaluation. This study used a pre-post-test design with a follow-up four weeks later to investigate whether a computerised training in identifying structural components of informal arguments could improve students' competences to understand complex arguments (Experiment 1), and whether teaching normative aspects of argument evaluation and fallacies could foster students' ability to evaluate the plausibility of arguments and identify common fallacies (Experiment 2). The trainings were embedded in a constructionist learning environment. Results indicate that training in argument structure based on the Toulmin-model of argumentation was particularly helpful for identifying more complex arguments and relational aspects between key components. High achieving students profited the most from this intervention. Experiment 2 showed that training in plausibility and normative aspects of argument evaluation improved students' evaluation skills and, in addition, their competences to recognise structural components. Our results suggest that interventions to foster argumentation skills should be included into the curriculum and these interventions should be designed to match learners' ability level.

Keywords: argument comprehension, epistemic competences, informal arguments, training

Argument comprehension and evaluation skills are essential for learning and decision-making across the lifespan. Lay people interested in socio-scientific issues such as risks of cell phones, media, vaccinations, or genetically modified food (cf. Sadler, 2004) are confronted with an overwhelming number of different, and often conflicting, arguments. Similarly, when university students learn about a scientific topic, they are required to read a variety of documents, many of which contain opposing evidence for different theoretical claims. Being able to comprehend and critically evaluate the claims and arguments presented in different texts comprises an essential aspect of scientific literacy (Britt, Richter, & Rouet, 2014).

Nevertheless, a considerable number of students possess insufficient skills to comprehend and evaluate arguments (e.g., NAEP, 1996, 1998; OECD, 2011, 2014). For example, results from the Programme for International Student Assessment (PISA) for reading and scientific literacy revealed that the majority of high school students were able to use basic scientific knowledge to identify a valid conclusion or scientific evidence for a claim, but only a minority of them were able to identify more complex arguments, to use evidence for evaluating the quality of arguments, link different knowledge, or apply relevant knowledge to unfamiliar or real-life situations (OECD, 2014). Similarly, only a small number of students were able to discriminate between relevant and irrelevant information. Although German students performed slightly higher than the OECD average for scientific literacy, these students faced similar problems.

The present research investigated the effects of a training intervention designed to improve students' competences to comprehend and evaluate complex scientific arguments that students typically encounter in the course of their studies. We begin with an analysis of the skills required to understand and evaluate such arguments. In this context, we outline the Toulmin model of argumentation (Toulmin, 1958) to describe the typical structure of an

argument and an analysis of the skills required to understand and evaluate these arguments successfully (Shaw, 1996; Voss & Means, 1991). Following this we discuss frequent challenges that students face when trying to comprehend and evaluate informal arguments and the conditions under which training in argumentation might be effective for overcoming these challenges (e.g., Larson, Britt, & Larson, 2004; Hefter et al., 2014, 2015). We then present results from two training experiments based on Jonassen's (1999) constructivist learning environment approach. Experiment 1 aimed at improving students' familiarity with the structure of informal arguments by teaching them how to identify different argument components and their relations. In Experiment 2, a number of useful processing strategies were conveyed – with an emphasis on attending to the internal consistency of arguments – along with conceptual knowledge about common argumentation fallacies, to help students improve their abilities to judge the plausibility of arguments and identify fallacies in argumentation.

Understanding and Evaluating Informal Arguments

When people comprehend and evaluate textual information, they construct a *mental model* of the situation described in the text from their general prior knowledge, i.e. a referential representation of the arguments' content (Johnson-Laird, 1983). These models are used for the evaluation of the content presented in a text, including informal arguments (Galotti, 1989; Perkins, 1986; Shaw, 1996). Scientific texts are often structured like arguments, stating different (usually empirical) evidence for theoretical claims, including counter-arguments and limitations of the evidence (Suppe, 1998). An argument is an attempt to convince the reader to accept a proposition, or claim (Galotti, 1989). Arguments found in scientific documents are informal, rather than formal arguments and their quality cannot be determined by formal, deductive logic (Galotti, 1989; Toulmin, 1958). Instead, in a strong, informal argument, the conclusion *probably* follows from the stated evidence (Voss et al.,

1991). Although, similar to formal arguments, informal arguments consist of a claim and one or more reasons, they may contain additional components (Toulmin, 1958). According to Toulmin, full-fledged arguments contain a number of functional key components: a claim, reason(s) (or datum /data), a warrant, backing evidence, and a rebuttal (Toulmin, 1958). The *claim* is the main statement being argued for. Claims are, by definition, controversial, and need to be supported with theoretical or empirical evidence which is referred to as *datum* (or *data)*. Claims and data are connected by the *warrant*. The warrant determines the strength of the evidence for the main claim, or, in other words, indicates whether the conclusion can be justified given the data. Another component, called *backing evidence,* provides (empirical or theoretical) support for the warrant. The warrant and its corresponding backing evidence are often not explicitly stated in everyday arguments, but need to be inferred by the reader (e.g., Chambliss, 1995). However, in the scientific texts, it is crucial to explicitly state why a particular conclusion is drawn from the results. Thus, warrants are particularly important in the scientific domain. Finally, *rebuttals* contain counter-arguments or indicate circumstances in which the argument does not hold true. Consider the following example (a brief summary of a study by Freeman et al., 2017):

> People should not eat eggs (*claim*), because eggs contain high amounts of cholesterol (*datum*). High amounts of cholesterol are unhealthy (*warrant*), because they may lead to coronary diseases (*backing*). However, individual factors play an important role and eggs may not increase the risk for coronary diseases in all people (*rebuttal*).

The claim that people should not eat eggs is supported by the datum that eggs contain high amounts of cholesterol. The datum lends support to the claim only on account of the warrant that high amounts of cholesterol are unhealthy. Backing for the warrant is stated by referring to the finding that high amounts of cholesterol may lead to coronary diseases.

However, the argument does not apply to all people, but individual factors play an important role. This last sentence constitutes the rebuttal.

Typically, the order in which the different components are presented is hierarchical, whereby the claim holds the top position because all other elements are presented to either support or oppose the main claim (*claim-first arguments*, Britt & Larson, 2003). However, arguments can also be stated in a less typical way. For example, they can begin with the datum, followed by the main claim (*reason-first arguments*), or with the rebuttal (e.g., Larson et al., 2004). Typical arguments are processed faster and more accurately than less typical arguments, because they are usually more congruent with the readers' current mental model (Britt & Larson, 2003). Most arguments contain linguistic markers or connectives like "therefore" or "because". These markers signal relations across the different components and help the reader to construct a coherent representation of the text. Britt and Larson (2003) found that arguments with markers are processed faster than arguments without these signals and that statements including modal verbs (e.g., *should*) and uncertainty markers (e.g., *probably*) signaled controversial statements requiring support.

Awareness of an (accurate) argument schema, including relevant markers (e.g., modals and qualifiers), can help the reader to identify the main claim, link the data to this claim, guide coherence inferences, activate possible alternative explanations, and form a corresponding mental model– cognitive processes that are not only relevant for comprehension, but also for evaluation (Britt & Larson, 2003; Shaw, 1996; Wolfe, Britt, & Butler, 2009).

Voss and Means (1991) name several criteria that are important for a successful evaluation of informal arguments. According to Voss and Means, readers need to look at the truth of data and claim, the quality of their relationship (i.e. the relevance of the data for the claim), and at the presence of relevant information for all aspects of the issue. Similarly,

Shaw (1996) describes three forms of judgments readers may engage in when encountering informal arguments (cf. Blair & Johnson, 1987). When readers evaluate the truth of the claim or the data (*accuracy*), they form *assertion-based* judgments. When they consider the internal consistency of an argument, i.e. the extent to which the data provide relevant support for the claim (*relevance*), they form *argument-based* judgments. Finally, when readers focus on whether all aspects of information relevant for the truth of the claim have been considered, they form *alternative-based* judgments (*sufficiency*). For the evaluation of informal arguments, argument-based or alternative-based judgments are especially important, because they consider the internal consistency of an argument, which, in turn, allows the reader to determine the *strength* of the argument (Osherson, Smith, & Shafirl, 1986). Although these argument-based and alternative-based judgments may be used by expert readers, the question arises whether lay readers, who have not received formal training in argumentation, are able to understand and evaluate arguments in this way.

The Challenges of Dealing with Informal Arguments among Lay Readers

A number of studies suggest that lay readers use epistemic processing skills to guide comprehension and evaluation of arguments to some extent (see Johnson, Smith-McLallen, Killela, & Levin, 2004, for a review), but that they are not always accurate in doing so. Even younger students seem to use argument schemes to guide comprehension if the structure of arguments is made explicit to them (e.g., Chambliss, 1995; Chambliss & Murphy, 2002). For example, Chambliss (1995) provided high school students (12^{th} graders) with clearly structured argumentative texts that included strong syntactical elements (signals) and introductory and concluding paragraphs that summarised the structure of the text. She found that students were able to recognise the argument structure and signalling text cues, and used them to guide comprehension and to construct accurate representations of the argument. However, Larson et al. (2004) noted that given their optimized structure, the arguments used

in Chambliss's (1995) study were rather atypical for informal arguments. In their study, Larson et al. (2004) used more authentic arguments that included arguments with a less typical structure and found that university students identified only 30% of their key components correctly. For example, the students in their experiment often misidentified uncontroversial and unsupported statements, data, and even counter-arguments (when the rebuttal was stated first) as the main claim. Similarly, von der Mühlen, Richter, Schmid, Schmidt, and Berthold (2016) used think-alouds to compare the performance of experts (advanced doctoral and post-doctoral students) with that of introductory university students and found that undergraduates struggled to identify key components of the Toulmin model, especially warrants.

Further evidence suggests that students' deficits in the ability to deal with the structure of arguments is related to difficulties to evaluate their quality. The participants in Larson et al.'s (2004) and those in von der Mühlen et al.'s (2016) study often failed to appropriately judge whether the stated data supported the stated claim, i.e. they neglected the internal consistency of the argument. In addition, the students in von der Mühlen et al.'s study struggled with the identification of common argumentation fallacies, such as circular arguments or overgeneralisations. In a similar vein, lay readers often do not integrate information from alternative positions in their mental models (Britt, Perfetti, Sandak, & Rouet, 1999). This, again, poses a problem for the evaluation of an argument based on its internal consistency rather than personal opinion. Several other studies suggest that students do not adequately represent relations between argument components (e.g., Britt & Kurby, 2005; Larson, Britt, & Kurby, 2009; Shaw, 1996). For example, Britt, Kurby, Dandotkar, and Wolfe (2008) found that students had difficulties to precisely recall the main predicate of a claim and this poor performance was linked to their ability to detect poorly formed arguments. Importantly, not all students showed these deficits, and the authors suggested that

familiarity with the argument scheme may have influenced their ability to represent the claims accurately. One explanation may be that evaluations of relational aspects between argument components are more effortful (Shaw, 1996). Readers need to access relevant prior knowledge from memory, activate alternative explanations, and keep this information activated in working memory.

Thus, lay readers often seem to struggle with the comprehension and normatively accurate evaluation of more complex arguments. They lack relevant structural knowledge and find it particularly difficult to attend to relations between argument components, such as warrants and the internal consistency of arguments.

Improving Lay Readers' Competences to Comprehend and Evaluate Informal Arguments

The difficulties among students to correctly comprehend and evaluate arguments highlight the need for explicit instruction and training of the strategies involved. Part of the problem may be that students have never received formal training in the skill of argumentation (Perkins, 1985). Although students entering university are expected to possess relevant argumentation skills, they usually have little experience with more complex arguments and relevant reasoning strategies. Textbooks, the dominant genre type used in high school classrooms, rarely contain complex arguments (Calfee & Chambliss, 1988; Paxton, 1997) and underlying relationships are often neglected (Beck, 1989). Past research indicates that students may require practice in understanding the connection between data and claim (e.g., Larson et al, 2004; Larson et al., 2009; Shaw, 1996; von der Mühlen et al., 2016). In a second experiment, Larson et al. attempted to address some of the issues by developing a short (10 minutes) tutorial in which they defined key components of arguments and named a number of steps for comprehending arguments (e.g., writing down the main claim and supporting data). Larson et al. showed that teaching the structure of an argument

helped students to shift attention to the internal consistency of an argument when immediate feedback was provided.

Thus, it appears that even short-term interventions, with a focus on providing knowledge about structural components of arguments and their relations, can be a promising approach to help students improve their argument comprehension and evaluation skills. In addition, the results from von der Mühlen et al. (2016) suggest that students might benefit from an intervention that provides specific knowledge about common fallacies when dealing with the plausibility of informal arguments.

Using Constructive Leaning Environments for Instruction

Constructivist learning environments (*CLE*, Jonassen, 1999), which are based on the assumption that knowledge cannot be transmitted but is individually constructed by the learner, have been shown to be effective for instruction in a number of interventions (e.g., Berthold & Renkl, 2010; Hefter et al, 2014; Hefter et al., 2015). The use of learning goals for real-world problems, feedback, and interactive environments, in which learners are allowed to correct their responses and in which information is easily accessible, are central elements of a CLE (Jonassen, 1999). In addition, varied examples or cases of a problem should be included to represent complexity and enable cognitive flexibility. Finally, instructors are encouraged to provide support when needed. For example, experts can serve as *cognitive models* who demonstrate different cases (examples) of the problem and relevant strategies required to solve the problem (Jonassen, 1999, Renkl, 2009). Such illustrations can reduce cognitive complexity and help the learner to deeply process information during the practice phase (Renkl, 2009). Video tutorials are particularly useful, because they stimulate both visual and auditory channels and thereby reduce cognitive complexity (Mousavi, Low, & Sweller, 1995). Finally, instructional prompts, in which learners are required to self-explain

stated information, have been shown particularly useful for the acquisition of knowledge, because they stimulate active, deep processing of information (Berthold & Renkl, 2010).

The Present Research

The present research was designed to investigate whether training in argument structure and normative aspects of plausibility could improve psychology students' competences to comprehend and evaluate informal arguments. It extends previous studies by designing an intervention for university students that aims at improving both argument comprehension and evaluation skills for more complex arguments typically found in scientific texts, including warrants (Larson et al., 2004; Jonassen, 1999).

Teaching students to attend to relational aspects of argument components was a major concern (Larson et al., 2004; Larson et al., 2009; von der Mühlen et al., 2016). In addition, relevant knowledge about argumentation fallacies was included in the intervention (von der Mühlen et al., 2016). The study also extends prior research by considering characteristics of the reader. We were particularly interested in a possible (moderating) influence of study performance on the effects of our training intervention. Assuming that students with better study performance are more likely to be familiar with a broad range of scientific texts, this might (implicitly) provide them with some relevant prior knowledge (i.e. *discipline expertise*, Rouet et al., 1997) about the structure of arguments. This structural prior knowledge, in turn, should allow them to more easily integrate and apply information from the training intervention. Furthermore, their familiarity with various scientific texts might generally foster their ability to understand and evaluate the soundness of informal arguments (Britt et al., 2014; Rouet et al., 1997).

As knowledge about the structure of arguments has been shown to be particularly important for comprehension and evaluation (Britt et al., 2014; Britt & Larson, 2003; Larson et al., 2003; Larson et al., 2004; Wolfe et al., 2009), Experiment 1 evaluated the effects of an

intervention designed to improve students' competences to recognise the structural components of informal arguments and their relations, including relevant markers. Experiment 2 evaluated the effects of a training designed to enhance their abilities to evaluate the plausibility of arguments, and to identify common argumentation fallacies. Students' performance on a computerised pretest was compared to the performance in a posttest and follow-up four weeks after the posttest, and to the performance of a control group who received a speed-reading training. Furthermore, it was investigated whether training of argument structure would be sufficient to improve performance in the plausibility task and/or if training in plausibility judgements would enhance performance in the argument structure task. Presuming that specific conceptual and procedural knowledge about the relevance of paying attention to the internal consistency of arguments is a prerequisite for their successful evaluation (Shaw, 1996), we were particularly interested in whether training in plausibility judgements would improve performance in the argument structure task. In addition, it was examined whether study performance would influence or moderate posttest and follow-up accuracies. Finally, we investigated as an exploratory research question whether pretest accuracies would predict or moderate performance in the posttest and follow-up.

Experiment 1

Experiment 1 examined an intervention that was designed to increase students' familiarity with the basic structure of informal arguments, and to improve their ability to recognise different components and their relations using the Toulmin (1958) model. The following hypotheses were formulated:

1. **Participants in the experimental condition will improve their comprehension of different components of the Toulmin model, as reflected in higher posttest and**

follow-up accuracy scores, compared to the control condition. This expectation was based on the assumption that argument comprehension requires abstract representations of the functional components of arguments and their interrelations (Britt et al., 2014; Britt & Larson, 2003; Wolfe et al., 2009).

2. **Less typical arguments and arguments with five components will be more challenging to identify and will receive lower pretest accuracy scores than arguments with a typical structure and arguments with three components.** Arguments with a typical structure are more likely to be congruent with the current state of the reader's mental model than less typical arguments (Schroeder, Richter, & Hoever, 2008). In addition, some components (i.e. warrants, backing evidence) are often not explicitly stated in a text and therefore more difficult to identify than other components, such as claims, reasons, or rebuttals (von der Mühlen et al., 2016).
3. **Participants in the experimental condition will improve their competence to identify arguments with an atypical structure and less typical components (i.e. warrants and backing evidence), compared to the control condition**. The training intervention was particularly designed to improve students' competences to handle more complex and less typical argument types as well.
4. **Participants in the experimental condition will improve their competence to judge the plausibility of arguments and identify common argumentation fallacies, whereas no such transfer effects will be found for the control condition**. This was expected because knowledge about the structure of arguments is a prerequisite for their successful evaluation (Britt et al., 2014).
5. **Students with higher average grades will particularly profit from the argument structure intervention, as reflected in higher posttest and follow-up accuracy scores, compared to the control condition**. We assume that these students will

benefit from their past experience with scientific literature and prior knowledge about the structure of arguments (Rouet et al., 1997).

The (moderating) influence of pretest accuracies was examined as an exploratory research question. Those students with better pretest performance might particularly profit from the training if the training required a certain level of understanding. Alternatively, those students with lower pretest accuracies might profit from the intervention, because there is more room for improvement.

Method

Participants. Fifty-three psychology students (10 males, 43 females) with an average age of 24 years (*SD* = 5.70) participated in the study. The majority of students (37) were undergraduates in their second semester, nine of them were in their fourth semester, and four students were in their sixth semester. Three participants had started a Master's programme. Participants provided informed consent at the beginning of the experiment and were reimbursed with course credits or financial remuneration (8 Euros per hour) after the completion of all sessions. In addition, they received an optional feedback with regard to their progress a few weeks later.

Text materials. All materials were presented in German. The examples stated in the present paper were translated into English.

Argument structure test. The text materials provided for the identification of different argument components were short argumentative texts with a mean length of 89 words in each argument. Three parallel versions were created based on von der Mühlen et al.'s (2016) study, and additional arguments were taken from their pilot study. The texts were slightly adapted to create a varied sample of arguments with a more or less typical structure. The texts were summaries of existing empirical articles, adapted to fit the structure

of Toulmin's (1958) model. Each of the versions contained four texts. Three of those texts were full-fledged arguments, including a claim, a datum, a warrant, backing evidence, and a rebuttal (Toulmin, 1958), and one of them contained only a claim, a datum, and a rebuttal. Two of the texts (including the argument consisting of three components) exhibited a typical structure (claim-first arguments, Britt & Larson, 2003). The two remaining texts were atypically structured (reason-first arguments, Britt & Larson, 2003). The texts had rather low readability scores that were representative of the literature students typically read (M = 17) as indexed by the German adaptation of Flesch's Reading Ease Index (Amstad, 1978).

Plausibility judgements test. The text materials for the plausibility judgements were three expository texts similar to the documents psychology undergraduates typically read. Two texts were taken from von der Mühlen et al. (2016). Both texts dealt with different aspects and theories of smoking behaviour (adapted from Fuchs & Schwarzer, 1997; Schroeder, Richter, & Hoever, 2008). The third text was derived from their pilot study and was about objective self-awareness (adapted from Thomas, 1991; Schroeder, Richter, & Hoever, 2008). All texts were of similar length, with 371 words in Text 1, 394 words in Text 2, and 403 words in Text 3. Five sentences in each text were implausible, and the remaining sentences were either plausible arguments by themselves or formed plausible arguments with the previous sentence. Implausible arguments were created by weakening the justification for the claim, and including one of five common argumentation fallacies (false conclusion, false dichotomy, wrong example, circular reasoning, overgeneralisation; Dauer, 1989). It is important to note, however, that all implausible sentences were semantically and syntactically correct, and both plausible and implausible sentences were coherent with previous discourse context. Both types of sentences were comparable in features such as length or semantic complexity. Plausible sentences had a mean length of 31 words (Text 1), 35 words (Text 2), and 27 words (Text 3). Implausible sentences had a mean length of 36

words (Text 1), 37 words (Text 2), and 27 words (Text 3). Moreover, they had similar readability scores (M = 16 for the plausible sentences vs. M = 17 for the implausible sentences) as indexed by the German adaptation of Flesch's Reading Ease Index (Amstad, 1978).

Argument structure training. The training intervention conveyed both conceptual and procedural knowledge in a constructive learning environment, using a cognitive modelling approach (Jonassen, 1999). A theoretical introduction provided appropriate background knowledge about the structure of full-fledged arguments (Toulmin, 1958). Learning goals and prompts were used to foster active and focussed processing of the instructions, the central concepts of the explanations, and the practice items (Berthold & Renkl, 2010). Based on Jonassen's (1999) cognitive modelling approach, two video tutorials were used to explain the strategies needed to correctly identify different argument components. In the practical part, participants worked on a number of argumentative texts. Feedback was provided for each task and participants were able to access relevant information (e.g., theoretical information, video tutorials, notes) when needed at all stages of the experiment (cf. Hefter et al., 2014, 2015).

Theoretical introduction. In the theoretical introduction, relevant knowledge about the structure, relevance, and purpose of informal arguments for scientific literacy (Britt, Richter, & Rouet, 2014; Britt & Larson, 2003; Wolfe et al., 2009) was provided. The Toulmin (1958) model was explained using a visual scheme. All theoretical input was explained with several examples to reduce cognitive complexity (Mousavi et al., 1995) and enable deep cognitive processing during the practical exercises (Renkl, 2009). These examples portrayed the problem (e.g., *Identify the claim of an argument*), pointed out different strategies to solve the problem (e.g., *Pay attention to markers*), and revealed the solution to the problem (e.g., *The first sentence is the claim*). Furthermore, attention to

markers of epistemic modality, such as *should*, and connectors such as *as a result,* or *therefore* (Britt & Larson, 2003), was introduced as a strategy to recognise different argument components and their relations. A number of learning goals were formulated to foster focussed and active processing of information (Berthold & Renkl, 2010). These included three questions: (a) *What does the basic structure of an argument look like?* (b) *Which components does an argument include?* and (c) *How can we identify different argument components?*. Participants were prompted to answer these questions at different stages of the experiment.

Explanation prompts. Specific prompts requested participants to reproduce conceptual information. The following prompts were integrated into the learning environment: (a) *Name each argument component and enter them in a text field*; (b) *Assign each argument component to its corresponding position within the scheme using a dropdown button*; (c) *Provide a written definition of each component*; and (d) *Name useful strategies for recognising the components in an argument and write them down in a text field*.

Video tutorials. Two video tutorials were developed to convey the strategies needed to identify the components of arguments. Both tutorials included one full-fledged argument including a claim, a datum, a warrant, backing evidence, and a rebuttal (Toulmin, 1958). Again, the arguments were summaries of existing empirical articles from different fields within the domain of psychology, adapted to fit Toulmin's (1958) model. The first tutorial (length: 03:41 minutes) described a typical argument (73 words), beginning with a claim and followed by the datum, the warrant, backing for this warrant, and a rebuttal. The second tutorial (length: 03:58 minutes) included an atypical argument (76 words) and began with the datum, followed by the warrant, backing for the warrant, the claim, and the rebuttal. A male model, who was portrayed as an expert in argumentation, read aloud both arguments and explained its structure in a stepwise fashion (cp. Jonassen, 1990; Renkl, 2009). Each

argument component was explained separately and elaborative information was provided with an example. Markers signalling relations between argument components were highlighted in each statement and explained by the model. The two arguments can be found in Appendix A. Again, readability of the arguments was rather difficult (M = 29 in Argument 1 vs. M = 19 in Argument 2), as indexed by the German adaptation of Flesch's Reading Ease Index (Amstad, 1978).

Practice texts. The practice texts included twelve arguments. As in the video tutorials, the texts were based on existing empirical articles from different fields within the domain of psychology, summarised to represent each component of the Toulmin (1958) model. Generally, the structure of the arguments resembled the texts used in the pretest, posttest, and follow-up. Furthermore, different types of arguments were included to increase complexity (Jonassen, 1993, Larson et al., 2004). The texts in the training included both full-fledged arguments (Toulmin, 1958) and arguments with only three components (claim, datum, rebuttal), and both typical (claim-first) and atypical (reason-first) arguments (Britt & Larson, 2003). The arguments had a mean length of 88 words. As in the tests and tutorials, the texts had rather low mean readability scores (M = 31), as indexed by the German adaptation of Flesch's Reading Ease Index (Amstad, 1978). However, readability was slightly higher due to the inclusion of simpler arguments with only three components.

Feedback. In every exercise, participants received immediate feedback on the correctness of their response, including the correct solution. In addition, a table showing general progress was provided. This table gave informative feedback on the number and types of argument components that had been assigned (in)correctly so that participants could repeat more difficult tasks.

Speed-reading training. For the control group, the application *Schneller Lesen* (reading faster, Heku-IT) was used to practice fast reading. The application consists of

several exercises, whereby each takes about 60 seconds. These exercises are embedded in several superordinate lessons, each containing eight exercises. The most important strategies used by the application to improve speed-reading competences are avoiding setbacks, not reading every single word of a text silently to oneself, and conceiving groups of words as an entity to derive meaning. The application provides feedback by granting points for successfully completed exercises.

Validation of text and item materials. The text materials for the pretest, posttest, and follow-up were normed and validated in a study by Schroeder et al. (2008) and in the pilot study preceding the study by von der Mühlen et al. (2016). For the argument structure training (i.e. the texts used in the tutorial and practice session), interrater reliability was determined by two doctoral candidates in the domain of psychology. There was high agreement among raters that all argument components in the training material were described and assigned correctly, Cohen's $\kappa = .95$. The speed-reading application has been tested and rated as "best product" by the leading German product testing group (Stiftung Warentest, 2015) for improving speed-reading competences up to 50% and remembrance of a text without any decline in understanding.

Software. The testing software used to display the tests and to record responses and response times was Inquisit 3.0.6.0. It was run on four identical HP notebooks with 15`` screens. For the speed-reading training, Android OS, v4.4.2 (KitKat) was used for each of five identical ASUS computers (10.1'') on which the application was installed.

Procedure. Participants were tested in groups of up to four people in a laboratory, and completed a total of four sessions, including a pretest, a training intervention, a posttest, and a follow-up. The interval between the pretest and the training intervention was one week, the posttest was conducted 15 minutes after the training session, and the follow-up was performed four weeks later. The pretest took about one hour, the combined training and

posttest session approximately 90 minutes (60 minutes for the training and 30 minutes for the posttest), and the follow-up about 40 minutes.

Pretest. Upon arrival, participants were welcomed, briefly informed about the procedure, and seated in front of a computer where they gave informed consent to participate in the experiment. Study performance was assessed with self-reported average grades in their present course of studies. The participants worked on the argument structure test, followed by the plausibility judgements test. In both tests, every individual completed one of the three parallel versions. The other parallel versions were carried out in the posttest and in the follow-up. The order in which these versions were presented was counter-balanced, and participants were randomly assigned to one of the versions.

Argument structure test. In the argument structure test participants were asked to identify the different components of four short arguments. The participants were asked to read the complete text first. In a second step, the text was presented again in fragments which consisted of several paragraphs, whereby each paragraph represented a different component of the argument, i.e. claim, datum, warrant, backing, and rebuttal. The paragraphs were numbered, and participants were instructed to assign each number to its corresponding argument component that had to be selected from a list appearing at the bottom of the screen. For each argument component, a short definition was provided.

Plausibility judgements test. In the plausibility judgements test, participants were asked to judge the plausibility of different statements in two argumentative texts. They were instructed to read the texts thoroughly, sentence by sentence on a computer screen in a self-paced fashion. Participants judged the plausibility of each statement by pressing a key for *plausible* or another key for *implausible.* They were asked to judge the internal consistency and quality of the arguments, and not to base their judgements on their own opinion or knowledge about the content of the text. Furthermore, they were told that

global fallacies, i.e. inconsistencies of a statement with other passages mentioned earlier in the text, were not included. After participants rated the plausibility of each item in the text, they were instructed to allocate the sentences they had marked implausible to specific argumentation fallacies that were explained briefly.

Training. One week after the pretest, participants returned to the lab for the training intervention. As in the pretest, they were welcomed, briefly informed about the procedure upon arrival, and seated in front of a computer. Subsequently, they were randomly assigned to either the argument structure training intervention or to a control group in which they worked on their speed-reading competences.

Argument structure training. Participants in the argument structure training were allowed as much time as they needed to complete the training. They were provided with a headset that they were instructed to use during the video tutorial.

Participants received theoretical input first. After a short explanation of the relevance and purpose of arguments, the Toulmin (1958) model was introduced in a stepwise fashion using several examples and the importance of markers and key words was highlighted. Subsequently, a number of learning goals were formulated, followed by two prompts in which participants were instructed to answer the questions formulated in the first and second learning goals. In the first exercise, they were asked to allocate each argument component to its corresponding position in the Toulmin (1958) model with the help of dropdown elements. Immediate feedback was provided and participants were allowed to correct their responses, if necessary, or to proceed with the next task. In the second exercise, the argument components had to be entered in an empty text field. Again, participants received feedback on the correctness of their response and participants could either correct their response or continue. In the next step, participants were instructed to put their headsets on and watch the two video tutorials in which strategies to identify the components of arguments were

demonstrated by a model. Following this, they were prompted to write down useful strategies to identify each argument component (third learning goal). They were allowed to access this information, along with the theoretical input and the tutorials, throughout the experiment by pressing a button at the bottom of each page (cp. Britt & Aglinskas, 2002). In addition, this page appeared after every feedback, and participants could decide whether they wanted to review particular information or proceed. In the practical phase, a number of different arguments were presented. These arguments were preceded by an example text. Participants were instructed to select the appropriate argument component for each paragraph of a text that was presented as a complete text first, and then in fragments. In addition, they were asked to find markers and write them down in an empty text field. A scheme displaying the Toulmin (1958) model appeared at the bottom of each practice text. As soon as each argument component was assigned a position in a text, participants received feedback on the correctness of their response, and the correct solution appeared both in the text and in the scheme. Again, they were given the opportunity to correct their responses, review certain information (e.g., theoretical input, video tutorials, notes), or continue with the following text. Finally, participants were once again prompted to provide an answer to the three learning goals that had been formulated at the beginning of the experiment, and to write down which parts of the training they found most helpful, before they were allowed a short break (15 minutes).

Speed-reading training. Participants in the control group were provided with tablets and worked on eight exercises, whereby each of them was limited to a processing time of 60 seconds. The exercises included an initial assessment of reading speed (1), tasks whereby a moving dot had to be tracked while different words were presented (2), particular letters had to be identified (3, 6), dissimilar word pairs identified (4, 7), particular words tracked while fixating a row (5), and a task wherein a dot had to be followed along

several rows of words. After the completion of all exercises, participants were shown how many points they had collected in each exercise and took a break of 25 minutes. After that, another eight, participants were instructed to keep practising with similar exercises until the timer reached 50 minutes. Finally, participants were allowed another 15-minute break.

Posttest. After the break following the training intervention, participants completed both the argument structure test and the plausibility judgements test again. They were randomly assigned to one of the parallel versions that they had not completed yet. At the end of the session, they were asked to indicate how confident they felt in dealing with the argument structure model on a Likert scale ranging from *1 = not confident at all* to *6 = very confident*. Finally, the students were thanked again for their participation and reminded of the upcoming follow-up session, before they were dismissed.

Follow-up. Both tests were completed a third time in the final session, whereby participants worked on the remaining parallel versions of the tests. They were once again asked about their confidence with regard to the argument structure model and its application at the end of the session. Finally, they were thanked for participation, reimbursed with course credits or financial remuneration, and dismissed. Participants were debriefed a few weeks later, and received an individual feedback about their training success upon request.

Design. The study comprised of a single factor (intervention: argumentation structure training vs. speed-reading training) between-subjects design. Accuracy of responses in the posttest served as the dependent variable. Participants were randomly assigned to one training condition. The test battery included three parallel versions of the test with four short argumentative texts for the argument structure test in each version and one relatively long text for the plausibility judgements test in each version. The order in which the versions were presented was counter-balanced across participants. Differences in pretest accuracies and

study performance were controlled for as covariates.

Results

Type-I-error probability was set at .05 for all hypothesis tests. One-tailed tests were used for directional hypotheses. The hypotheses were tested with posttest or follow-up accuracies as the outcome variable. We used linear models with categorical and continuous predictors and interaction terms (Cohen, Cohen, West, & Aiken, 2003, Chapter 9) for testing univariate hypotheses that predicted higher values in the experimental compared to the control condition, and to examine whether and to what extent pretest accuracies predicted or moderated the expected superior performance of participants in the experimental group as reflected in the response accuracies (Model 1). Additional analyses were run to examine the (moderating) influence of study performance on posttest and follow-up accuracies (Model 2). All continuous predictors were *z*-standardised and entered simultaneously into the models at each step. Training condition was included as contrast-coded predictor (1: argument structure training, -1: speed-reading control condition).

Response accuracy. In the pretest, both training groups achieved similar accuracy scores in the argument structure test, $p > .05$. The results for Model 1 showed that training in argument structure did not generally improve performance in the posttest or follow-up, $p > .05$, but a significant effect of pretest accuracies was found for the posttest ($B = 0.06$, $SE = 0.03$, $p < .01$, $\Delta R^2 = .10$), and for the follow-up ($B = 0.08$, $SE = 0.02$, $p < .001$, $\Delta R^2 = .23$), indicating that students with higher pretest accuracies scored higher after the intervention and four weeks later. Results of the moderated regression analyses for the posttest are displayed in Table 1.

However, when study performance and its interaction with training condition were added to the model (Model 2), study performance moderated the effect of training condition in the posttest, ($B = -0.05$, $SE = 0.02$, $p < .05$, $\Delta R^2 = .06$). To interpret the interaction, we

estimated and plotted the simple slopes of study performance in the argument structure training and the speed-reading training condition (Figure 1) and estimated the effect of training condition at a low level of study performance and at a high level of study performance (Cohen et al., 2003, Chapter 9). The negative slope of study performance was steeper in the argument structure training condition (B = -0.12, SE = 0.03, $p < .001$, one-tailed, $\Delta R^2 = .06$) than in the speed-reading condition where it was not significant (B = -0.03, SE = 0.03, $p = .17$, one-tailed). Note that the simple slopes are negative, because lower values represent better performance in the German grading system. At a low level of study performance (i.e., a mean grade of 1 *SD* above the sample mean), the two training conditions did not differ in posttest accuracy (argument structure training: $M = .61$, $SE = .05$; speed reading training: $M = .64$, $SE = .05$), t (43) = - 0.50, $p = .31$, one-tailed. At a mean level of study performance, the posttest scores were higher after the argument structure training (M = .73, $SE = .03$) compared to the speed-reading training ($M = .67$, $SE = .03$), but the effect missed the significance criterion by a narrow margin, t (43) = 1.65, $p = .05$, one-tailed. In contrast, at a high level of study performance (i.e., a mean grade of 1 *SD* below the sample mean), participants in the argument structure training clearly outperformed those in the speed-reading training (argument structure training: $M = .86$, $SE = .04$; speed-reading training: $M = .70$, $SE = .05$), t (43) = 2.43, $p < .01$, one-tailed. Thus, students with very good grade average, i.e. above-average study performance, benefitted from the argument structure training. In addition, as in Model 1, a significant main effect of pretest accuracies was found in the posttest (B = 0.04, SE = 0.02, $p < .05$, $\Delta R^2 = .05$) and at follow-up (B = 0.07, SE = 0.02, $p < .01$, $\Delta R^2 = .15$).

Whereas the initial variables could only explain a moderate amount of the variance in our model ($R^2 = .19$), more than 40% of the variance could be explained after the addition of study performance and its interaction with training condition ($R^2 = .41$). The interaction of

study performance with the training condition was not found in the follow-up, $p > .05$.

To better understand these global effects that were found for the posttest, a number of follow-up analyses were performed, whereby Model 2 served as the basis for analysis. First, we looked at possible effects in different argument components. Subsequently, accuracies in different argument types were examined.

Accuracy in different argument components. In the pretest, it was most difficult to recognise the warrant ($M = .35$, $SE = .04$), followed by the backing evidence ($M = .51$, $SE = .04$), the datum ($M = .62$, $SE = .04$), the claim ($M = .62$, $SE = .04$), and the rebuttal ($M = .89$, $SE = .02$). The warrant was significantly more difficult to identify than all the other components, $p < .001$, and the rebuttal was significantly less difficult to identify than all the other components, $p < .001$. In the posttest, however, Model 2 revealed a significant effect of training condition for the ability to identify warrants ($B = 0.11$, $SE = 0.04$, $p < .001$, one-tailed, $\Delta R^2 = .09$), with significantly improved accuracy values in the argument structure training group ($M = .64$, $SE = .06$), as compared to the speed-reading training group ($M = .41$, $SE = .06$).

Moreover, study performance moderated the effect of training condition for identifying backing evidence ($B = -0.09$, $SE = 0.05$, $p < .05$, $\Delta R^2 = .06$). Estimation of the simple slopes of study performance for each training condition showed that the negative slope of study performance was steeper in the argument structure training condition ($B = -0.16$, $SE = 0.06$, $p < .01$, one-tailed, $\Delta R^2 = .06$) compared to the speed-reading condition, where it was not significant ($B = .01$, $SE = .06$, $p = .44$, one-tailed). At a low level of study performance (i.e., a mean grade of 1 *SD* above the sample mean), the two training conditions did not differ in postttest accuracy (argument structure training: $M = .46$, $SE = .09$; speed reading training: $M = .40$, $SE = .09$), $t\ (43) = -0.38$, $p = .34$, one-tailed. Similarly, posttest accuracies did not differ significantly at a mean level of study performance between the

argument structure training (M = .54, SE = .06) and the speed-reading training (M = 45., SE = .06), t (43) = -0.95, p = .17, one-tailed. However, at a high level of study performance (i.e., a mean grade of 1 *SD* below the sample mean), participants in the argument structure training outperformed those in the speed-reading training (argument structure training: M = .67, SE = .09; speed-reading training: M = .45, SE = .09), t (43) = -1.69, $p < .05$, one-tailed. Thus, students with very good average grades, i.e. above-average study performance, particularly benefitted from the argument structure training with regard to their ability to identify backing evidence.

No significant differences were found for the ability to identify claims (B = 0.04, SE = 0.03, $p > .05$, one-tailed), reasons (B = 0.01, SE = 0.04, $p > .05$, one-tailed), or rebuttals (B = 0.00, SE = 0.02, $p > .05$, one-tailed).

Accuracy in different argument types. In the pretest, atypical, full-fledged arguments (M = .52, SE = .04) were more difficult to identify than typical, full-fledged arguments (M = .68, SE = .05), $p < .001$, which, again, were more challenging than arguments with only three components (M = .89, SE = .03), $p < .001$. In the posttest, Model 2 revealed a main effect of training condition for the identification of atypical, full-fledged arguments (B = 0.08, SE = 0.03, $p < .01$, ΔR^2 = .09, one-tailed), with participants in the experimental condition receiving higher posttest accuracies (M = .68, SE = .04), as compared to those in the control condition (M = .53, SE = .04). For typical full-fledged arguments, pretest scores moderated the effect of training condition (B = 0.09, SE = 0.04, $p < .05$, ΔR^2 = .06). Estimation of the simple slopes of pretest scores for each training condition showed that the negative slope of pretest scores was steeper in the argument structure training condition (B = 0.20, SE = 0.06, $p < 01$, one-tailed, ΔR^2 = .06) compared to the speed-reading condition, where it was not significant (B = 0.04, SE = 0.06, p = .27, one-tailed). At a low level of pretest scores (i.e., a mean grade of 1 *SD* above the sample mean), the two training

conditions did not differ in posttest accuracy (argument structure training: $M = .88$, $SE = .09$; speed reading training: $M = .79$, $SE = .09$), $t\ (43) = 0.77$, $p = .22$, one-tailed. Again, no significant differences between the argument structure training ($M = .74$, $SE = .06$) and the speed-reading training ($M = .68$, $SE = .06$) were found at a mean level of pretest scores, $t\ (43) = 0.17$, $p = .25$, one-tailed. In contrast, at a high level of pretest scores (i.e., a mean grade of 1 *SD* below the sample mean), participants in the argument structure training performed better than those in the speed-reading training (argument structure training: $M = .69$, $SE = .09$; speed-reading training: $M = .48$, $SE = .09$), $t\ (43) = 1.75$, $p < .05$, one-tailed. Thus, students with high pretest scores particularly benefitted from the argument structure training with regard to the ability to identify typical full-fledged arguments. No significant training effects were observed for arguments with three components ($B = .01$, $SE = .03$, $p > .05$).

Transfer effects. The argument structure training did not improve students' abilities to judge the plausibility of arguments and identify common argumentation fallacies in the posttest ($B = 0.01.$, $SE = 0.02$, $p > .05$), or at follow-up ($B = 0.02$, $SE = 0.03$, $p > .05$). No significant correlations were found between the accuracies of both tasks in the posttest and follow-up, $p > 0.5$. Thus, no transfer effects on the ability to judge the soundness of arguments were found.

General feedback. When asked which parts of the training experiment participants found most helpful for improving their competence to recognise different argument components, the video tutorials were named most often (15), followed by the practice phase and the feedback (both 9), the theoretical input (6), and the prompts (1).

Discussion

Results indicate that the argument structure training did not generally improve performance in the posttest for all students in this group. However, it did improve the ability

to recognise warrants and atypical arguments. As some components, such as rebuttals, and some types of arguments, such as arguments with three components, were generally very easy to identify, our results were likely affected by ceiling effects. Interestingly, high achieving students, with higher average grades in their present studies, profited the most from the training intervention, with improved accuracies in the posttest for those who participated in the argument structure training, when pretest accuracies were controlled for. These students particularly improved their competence to identify backing evidence. Moreover, students with high pretest scores from the experimental group received higher posttest scores, and those who participated in the argument structure training improved their competence to identify typical, full-fledged arguments. Thus, students who were already able to recognise more complex types of arguments could further improve this competence during the training intervention. Unfortunately, the results could not be replicated in the follow-up. We assume that a more extensive intervention with multiple training sessions would be necessary to produce long-term effects.

No transfer effects were found from the argument structure task to the plausibility task. We presume that specific conceptual knowledge of argumentation fallacies and the concept of plausibility, as well as both conceptual and procedural knowledge about the relevance of paying attention to the strength or internal consistency of arguments for their successful evaluation is necessary to improve performance in this task. Therefore, Experiment 2 tested whether such a training would improve students' competences to judge the plausibility of arguments and identify common argumentation fallacies.

Experiment 2

The second experiment was designed to improve students' competences to judge the plausibility of informal arguments and to identify common argumentation fallacies. We

again compared performance in the pretest to performance in the posttest and follow-up, and to a control group who received a speed-reading training. In Experiment 2, the following hypotheses were formulated:

1. **Participants in the experimental condition will improve their ability to judge the plausibility of arguments and identify common argumentation fallacies, as reflected in higher posttest and follow-up accuracy scores, compared to the control condition**. Participants in the experimental condition were expected to improve their understanding of the quality of arguments after learning about and becoming familiar with the concept of plausibility (Matzuki et al., 2011) and the importance of argument strength and internal consistency for the evaluation of arguments (Blair & Johnson, 1987; Shaw, 1996; Voss & Means, 1991). Knowledge about common argumentation fallacies was assumed to help students detect weak arguments (Blair & Johnson, 1987; Dauer, 1989).
2. **Participants in the experimental condition will improve their competence to identify different components of the Toulmin Model, whereas no such transfer effects will be found for the control condition**. This was expected because participants in the plausibility training were also encouraged to pay attention to different argument components and their relationship when evaluating informal arguments.
3. **Students with higher average grades will particularly profit from the argument structure intervention, as reflected in higher posttest and follow-up accuracy scores, compared to the control condition**. We assume that these students will benefit from their past experience with scientific literature and prior knowledge about the structure of arguments (Rouet et al., 1997).

Again, the (moderating) influence of pretest accuracies was examined as an exploratory research question. Those students with better pretest performance might particularly profit from the training if the training required a certain level of understanding. Alternatively, those students with lower pretest accuracies might profit from the intervention, because there is more room for improvement.

Method

Participants. Twenty-seven undergraduate psychology students (6 males, 21 females) with an average age of 24.23 years (*SD* = 5.57) participated in the study. The majority of students (19) were in their second semester, three undergraduates were in their fourth semester, three students were in their sixth semester, and two had started a Master's programme. Participants provided informed consent at the beginning of the experiment and were reimbursed with course credits or financial remuneration (8 Euros per hour) after its completion. As in Experiment 1, they were given the opportunity to receive individual feedback with regard to their training success after the completion of all tests.

Materials used for the tests. As in Experiment 1, all text materials were presented in German. The examples provided in the present paper were translated into English. The same texts were used for the pretest, posttest, and follow-up as in Experiment 1.

Materials used for the trainings.

Plausibility training. Analogously to Experiment 1, the plausibility training conveyed both conceptual and procedural knowledge in a constructive learning environment. Again, it included a theoretical introduction, explanation prompts, a video tutorial, and one practice text. Feedback was provided after each exercise, and participants had access to important information (e.g., theoretical input, video tutorials, notes) throughout the experiment.

Theoretical introduction. The first part of the training contained a brief introduction about the relevance and usefulness of informal arguments and their evaluation for understanding scientific texts (Britt et al., 2014, Britt & Larson, 2003; Wolfe et al., 2009). In addition, a definition of a plausible and a definition of an implausible informal argument was provided with an example. Argument strength (Voss & Means, 1991) and the importance of the argument's internal consistency, i.e. the relevance of the datum for the claim (Blair & Johnson, 1987; Shaw, 1996), were presented as the crucial determinants for determining plausibility. Five common argumentation fallacies were presented and defined, the *False Conclusion* fallacy, the *Wrong Example* fallacy, the *False Dichotomy* fallacy, the *Circular Reasoning fallacy*, and the *Overgeneralisation* fallacy (Dauer, 1989). The fallacies were illustrated with several examples. Moreover, several learning goals were formulated. These goals included questions with regard to understanding the concept of an argument, plausibility, the difference between plausible and implausible arguments, and the strategies needed to identify common fallacies. The questions were prompted at several stages of the experiment.

Explanation/definition prompts. The following prompts were used both at the beginning and at the end of the experiment: (a) *Define an argument*; (b) *Define a claim*; (c) *Define a reason*; (d) *Define a plausible argument*; (e) *Explain how plausible arguments can be discriminated from implausible arguments*; (f) *Assign each fallacy to a statement in the text*. An empty text field was provided to enter responses for prompts (a) to (e). Dropdown elements were provided for prompt (f).

Video tutorial. The video tutorial (length: 07:40 minutes) contained one example for each of the five argumentation fallacies (see Table A1 for an overview of fallacies used in the tutorial), read aloud and explained by the same model who had served as an expert in Experiment 1. They were constructed based on common

fallacies students typically encounter. The *False Conclusion* fallacy confused cause and correlation, a mistake that is commonly made by undergraduate psychology students. The *Wrong Example* fallacy cited a number of instances that were meant to support the credibility of a statement. The examples were, however, irrelevant for the statement. The *False Dichotomy* fallacy comprised a statement with two options that were pictured as two mutually exclusive alternatives, but in fact overlapped. The *Circular Reasoning* fallacy used synonyms to obscure the fact that the reason provided for the claim was no different from the claim. Finally, the *Overgeneralisation* fallacy falsely drew a conclusion from an observation to a broader population that was not representative of this population. For each of the fallacies, crucial words were highlighted in the video tutorial. For example, in the statement *Last year I spent two weeks in Manchester. During my holiday, it rained every day. England is really a very rainy country* the word "England" appeared in red font to demonstrate the inappropriateness to generalise from an observation in one city to a whole country. For each statement, a plausible alternative was provided. In the example described above, a plausible statement would be *Last year I spent two weeks in Manchester. During my holiday, it rained every day. Presumably, it rains quite a lot in Manchester.*

Practice text. In the practice part, participants were provided with a text that was similar to the texts used in the pretest, posttest, and follow-up. The practice text addressed interpersonal attraction (adapted from Thomas, 1991; Schroeder et al., 2008) and comprised a total of 36 statements which were all part of a coherent text. The text had a length of 1164 words and a rather high reading difficulty typical for scientific documents (19, as indexed by the German adaptation of Flesch's Reading Ease Index, Amstad, 1978). Ten statements in each text were implausible, whereas the remaining statements were either plausible arguments by themselves or formed plausible arguments with the previous sentence. Implausible arguments contained one of the five argumentation fallacies,

whereby each fallacy occurred twice in the text. For example, the sentence *According to the balance theory, interaction partners are more likely to feel attracted to each other if they agree in their opinions with regard to certain persons, issues, objects, and events, because interpersonal liking increases* includes a Circular Reasoning fallacy. Table 3 provides an overview of the fallacies that were used in the practice session with an example for each fallacy.

Feedback. As in Experiment 1, participants received immediate feedback after every exercise on the correctness of their response and were provided with the correct solution.

Speed-reading training. The same materials were used for the speed-reading training as in Experiment 1.

Validation of text and item materials. As in Experiment 1, the texts used in the tutorial and in the practice session of the plausibility training were validated by two doctoral candidates in the domain of psychology, resulting in full agreement among raters that plausible items were indeed plausible and implausible items were implausible, and that all argumentation fallacies were described and allocated correctly, Cohen's $\kappa = 1$. The validation of all other materials was described in Experiment 1.

Software. The same software (and hardware) was used as in Experiment 1.

Procedure. The procedure was equivalent to that of Experiment 1, except that different materials were provided in the plausibility training session. The following procedure was applied in the plausibility training.

Participants were randomly assigned to either the plausibility training or the speed-reading training. As in Experiment 1, participants in the plausibility training group were welcomed and briefly informed about the procedure, before they were seated in front of a computer and provided with a headset for the video tutorial. They were allowed as much

time as they needed to complete the training. In the theoretical introduction, the relevance and purpose of the training was explained, followed by definitions of plausible and implausible informal arguments. Each definition was accompanied by an example that appeared when a button was pressed. Several leaning goals were formulated and participants were prompted to answer the questions formulated in these learning goals. They were asked to define an argument, a claim and a reason (or datum), and a plausible argument. In addition, they were asked to explain how plausible arguments could be discriminated from implausible arguments. They were instructed to enter their responses in an open text field. Immediate feedback was provided after each response and participants were allowed to correct their responses, if necessary, or to proceed with the next task. Following this, the five argumentation fallacies were explained in a stepwise fashion with several examples, and participants were prompted again to assign each fallacy to a statement using dropdown lists. Again, they received immediate feedback for this task, and were allowed to correct their response. Subsequently, participants were instructed to pick up their headsets and view the video tutorial in which strategies for the identification of these fallacies were explained. When they were finished watching the video tutorial, participants were provided with the practice text. The text was preceded by an example statement that included a circular reasoning fallacy. As in the pretest, participants were instructed to read the texts thoroughly, sentence by sentence on a computer screen in a self-paced fashion. Plausibility of each statement was judged by pressing a key for *plausible* or another key for *implausible.* Participants were instructed to apply the strategies they had learned in the video tutorial. They were asked to pay attention to the relevance of the datum for the claim (i.e. internal consistency) and to prove whether the conclusion (probably) followed from the evidence. Moreover, they were instructed not to base their judgements on their own opinion or prior knowledge about the content of the text. Immediate feedback on the correctness of their

response was provided, and participants were allowed to correct their responses, if necessary, or continue. Before the next statement was presented, participants were reminded of the possibility to access relevant information (e.g., theoretical input, tutorials, notes) by pressing the corresponding button. If a statement was marked as implausible, they were prompted to explain in their own words why the statement was implausible and what it could look like if it had been plausible. They were instructed to enter their response in an empty text field. Afterwards, participants selected the appropriate fallacy from a number of alternatives. Again, they received immediate feedback, corrected their response, or continued with the next statement, until they every item in the text was completed. As in Experiment 1, they were allowed a 15-minute break before they continued with the posttest.

Design. As in Experiment 1, the study comprised of a single factor (intervention: plausibility training vs. speed-reading training) between-subjects design. Accuracy of responses in the posttest served as the dependent variable. Again, participants were randomly assigned to one training condition. The same test battery was used as in Experiment 1 for the pretest, posttest, and follow-up, whereby the order of the test versions was counter-balanced across participants. Pretest accuracies and study performance were included as covariates.

Results

The data analyses in Experiment 2 were performed in a manner equivalent to Experiment 1, i.e. hypotheses were tested in linear models with categorical and continuous predictors and interaction terms, with posttest or follow-up accuracies as the outcome variable. Model 1 examined possible effects of the plausibility training and whether pretest accuracies would predict or moderate performance in the postttest or follow-up. As in Experiment 1, additional analyses were run to examine the (moderating) influence of study performance on posttest and follow-up accuracies (Model 2). One-tailed tests were used for hypotheses that predicted higher values in the experimental than in the control condition and

for students with better study performance who participated in the plausibility structure training. All continuous predictors were *z*-standardised and entered simultaneously into the models at each step. Training condition was included as contrast-coded predictor (1: plausibility training, -1: speed-reading control condition).

Transfer effects between the plausibility and the argument structure task were examined using Pearson correlations between the accuracies of both tasks in the posttest and follow-up, and tested in a linear model that included training condition (contrast-coded, 1: plausibility training, -1: speed-reading control condition), pretest accuracies in the argument structure task, study performance (*z*-standardized), and the interactions of pretest accuracies and study performance with the training condition as predictors, and posttest accuracies in the argument structure task as the dependent variable.

Response accuracy. In the pretest, both training groups achieved similar accuracy scores, $p > .05$. However, Model 1 revealed a significant effect of training condition in the posttest ($B = 0.16$, $SE = 0.08$, $p < .05$, one-tailed, $\Delta R^2 = .21$). Furthermore, as in Experiment 1, pretest accuracies predicted accuracies in the posttest ($B = -0.19$, $SE = 0.08$, $p < .05$, $\Delta R^2 = .12$), but in the opposite direction, indicating that students with higher pretest accuracies scored lower after the intervention. Results of the moderated regression analyses for the posttest are displayed in Table 2. When study performance and its interaction with training condition were added to the model, study performance strongly predicted (but did not moderate) performance in the posttest ($B = -0.29$, $SE = 0.04$, $p < .001$, one-tailed, $\Delta R^2 = .39$). Note again that coefficients are negative, because lower values represent better performance in the German grading system. Thus, students with higher average grades could improve their performance in both groups. The effect of training condition and pretest scores on posttest accuracies were still significant in this model ($B = 0.13$, $SE = 0.04$, $p < .01$, one-tailed, $\Delta R^2 = .11$ vs. $B = -0.15$, $SE = 0.04$, $p < .01$, $\Delta R^2 = .12$, respectively). Whereas the

initial variables could already explain a relatively large amount of variance in our model (R^2 = .48), a very large amount of the variance could be explained after the addition of study performance and its interaction with training condition (R^2 = .90). No significant differences were found for plausible items in the posttest in Model 1 (B = -0.01, SE = 0.03, p > .05, one-tailed) or Model 2 (B = -0.02, SE = 0.03, p > .05, one-tailed) or at follow-up (B = -0.02, SE = 0.03, p > .05, one-tailed in Model 1 and Model 2), with very high posttest accuracy values in both groups (M = .88, SE = .04 in the experimental group vs. M = .87, SE = .04 in the control group). Furthermore, no significant interactions were found in this experiment, and, similarly to Experiment 1, our results could not be replicated in the follow-up, p > .05.

Transfer effects. Model 2 revealed a significant effect of training condition on posttest accuracies in the argument structure task (B = -0.04, SE = 0.02, p < .05, one-tailed, ΔR^2 = .06). In addition, a positive correlation was found for the posttest between the plausibility task and the argument structure task (r = .67, p < .05). No correlation between these tasks was found in the control group, p > .05. Thus, training in plausibility not only improved performance in the plausibility task, but also in the argument structure task, indicating that knowledge was transferred between these tasks in this group.

Discussion

Results from Experiment 2 suggest that training and instruction of relevant knowledge about normative criteria for the evaluation of informal arguments (e.g., argument strength and the internal consistency of arguments) and strategies for dealing with common argumentation fallacies could improve students' abilities to detect implausible information. Similarly to Experiment 1, study performance had an effect on performance in the posttest, although it did not moderate the group effect. Interestingly, there was a negative correlation between pretest accuracies and posttest performance in both groups, indicating that those with high pretest accuracies could not profit from the intervention. It might be that the

training could not further improve their competence to evaluate arguments, because these students were already at a very high level. With regard to plausible information, our results were likely affected by ceiling effects, as these statements showed very high accuracy values in both groups. As in Experiment 1, results could not be replicated in the follow-up, suggesting that long-term training effects would probably require more extensive and more frequent interventions. Importantly, training in plausibility improved performance in the argument structure task, indicating that teaching students to pay attention to different argument components, and, in particular, their relationship (i.e. internal consistency), while evaluating informal arguments, might be a helpful strategy to improve students' general epistemic competences.

General discussion

The present experiments investigated how training in the ability to recognise different structural components of an argument (Experiment 1) or to determine the plausibility of arguments (Experiment 2) could improve psychology students' competences to comprehend and evaluate informal arguments. Results from our experiments indicate that both trainings successfully improved these skills. Experiment 1 showed that familiarising students with the structure of arguments improved their ability to recognise warrants and more complex (full-fledged) or less typical arguments. Students with very good grades particularly profited from the training intervention, as reflected in significantly improved performances after the intervention. Moreover, students who were initially able to recognise more complex argument types could further improve this ability in the intervention. The results found in Experiment 2 reveal that the acquisition of knowledge about normative aspects of argument evaluation, such as argument strength and the internal consistency of arguments, along with conceptual knowledge about common argumentation fallacies, may have been helpful to improve students' abilities to evaluate the plausibility of arguments and recognise such

fallacies. Most importantly, our results suggest that shifting attention towards relational aspects between argument components (i.e. warrants in Experiment 1 and argument strength or internal consistency in Experiment 2) showed the greatest increment in students' posttest performance. Finally, training in the ability to distinguish plausible from implausible arguments improved performance in the argument structure task, indicating that acquisition of conceptual and procedural knowledge about informal arguments and their successful evaluation may have helped with the formation of accurate representations of key components of arguments, including warrants.

Importantly, the students who participated in the argument structure training were generally able to improve their performance to recognise less typical argument components, such as warrants, and more complex (full-fledged) arguments with a less typical structure. However, participants in both groups were already relatively accurate in their ability to recognise more typical components, such as rebuttals (89% accuracy), and, to a lesser degree, claims (62% accuracy), and data (62% accuracy) or less complex argument types, such as arguments with only three components (89% accuracy), prior to the intervention. These results indicate that the students seemed to possess some prior knowledge of the structure of (less complex) arguments. However, only a minority of the participants in our study were able to correctly identify warrants (35% accuracy). Accuracy values for the identification of warrants almost doubled after the intervention for those who participated in the argument structure training (64%), suggesting that the intervention especially improved awareness of relational aspects between argument components. These findings indicate that training may be especially useful for less typical components, such as warrants and backing evidence, and for more complex, full-fledged arguments with a less typical structure (e.g. reason-first arguments, Britt & Larson, 2003). Results from Experiment 2 showed that shifting awareness to the internal consistency of arguments was helpful for both

comprehension and evaluation of arguments. Our results are in line with previous research indicating that students tend to neglect the internal consistency of arguments (e.g., Britt & Kurby, 2005; Larson et al., 2004; Larson et al., 2009; Shaw, 1996; von der Mühlen et al., 2016), but that training in argument structure and normative criteria for the evaluation of arguments can be effective in overcoming these deficits (e.g., Larson et al, 2004; Hefter et al., 2014).

The results found in our study can be interpreted in the framework of the mental model theory (Johnson-Laird, 1983). We presume that training in the identification of structural components of arguments, along with normative aspects of argument evaluation, and a strong focus on shifting students' attention towards the internal consistency of arguments, allowed the construction of more accurate representations of arguments in memory and helped students to activate different argument components simultaneously when trying to understand and evaluate these arguments (Britt et al., 2014; Shaw, 1996).

Not everyone profited from the training intervention to the same extent. Students with a better study performance profited the most from training in argumentation. We assume that students who performed very well in their current education were more familiar with a broad range of scientific texts than the average student, which might have (implicitly) provided them with some relevant prior knowledge (i.e. *discipline expertise*, Rouet et al., 1997) about the structure of arguments, and allowed them to more easily comprehend, integrate and apply information from the training intervention. In addition, their familiarity with various scientific texts might have fostered their competences to understand and evaluate arguments (Britt et al., 2014; Rouet et al., 1997). Our results indicate that students with very high study performance especially improved their competence to identify backing evidence for warrants, indicating that these students paid particular attention to identifying evidence for less typical, relational aspects of argument components, possibly because they

were already partly familiar with the structure of arguments and were able to focus on learning and improving their competences to identify less typical components. Our results are also in line with the assumptions of aptitude-treatment-interaction (API) effect (Snow, 1989). Whereas CLEs that allow the learner to explore a problem in a self-determined fashion seem to be particularly useful for readers with high prior knowledge, more structured interventions might be beneficial for readers with less prior knowledge.

In our experiments, we also examined effects of pretest accuracies on posttest performance. In this regard, our results are somewhat conflicting in both experiments. In Experiment 1, students with very high initial accuracy scores scored significantly higher after the intervention, indicating that they could further improve their ability to identify key components of arguments, especially warrants, and more complex, full-fledged arguments. Similarly to students with high study achievement, students with high initial accuracy scores were likely to possess some relevant prior knowledge about arguments. This knowledge might have helped them to concentrate on acquiring further knowledge about less typical argument components and more complex arguments, with which they were less familiar. In Experiment 2, students with high pretest accuracies could not repeat their very good performance, presumably because there was little room for improvement. We assume that these students suffered from fatigue effects after the intervention, which was cognitively very demanding.

It is not fully understood why the students with very good study performance could profit most from our intervention. While we assume that these students have acquired more experience with the structure of scientific texts and arguments, other mechanisms, such as differences in cognitive ability, might be responsible for the observed interaction of training condition and study performance. Although Stanovich (2012) found that the skill of rational thinking seems to be independent of intelligence, future research should address this issue.

Furthermore, the present study does not flesh out the precise mechanisms under which students acquire a reasoning schema. Although students indicated that they perceived the video tutorials, the practice phase, and the presence of feedback as very helpful, manipulations that include or exclude different tools and measures tapping into cognitive processes during training would be necessary to achieve more objective insights. It remains unclear which part of the training was responsible for improving the skill to detect implausible information in arguments.

Despite these limitations, our results indicate that interventions focusing on the construction of conceptual and procedural knowledge about informal arguments in a constructivist setting can be effective for fostering students' competences to comprehend and evaluate these arguments. Knowledge about the structure of (complex and less typical arguments) may lead to more accurate and flexible representations in memory, allowing students to access this knowledge when encountering numerous claims and evidence in scientific texts. Moreover, encouraging students to pay attention to relational aspects of argument components seems to be a prerequisite for the successful evaluation of the quality of arguments. Unfortunately, the results we observed after the interventions could not be replicated four weeks later. Future research should develop more extensive interventions with several practice sessions to produce long term effects. Finally, our results raise the question of how much value we assign to the acquisition of epistemic competences in formal instruction and education. Assuming that lack of practice is one of the main reasons why students find it difficult to comprehend (more complex) arguments and form appropriate decisions about their quality (Perkins, 1985; Perkins, Farady, & Bushey, 1991), interventions that aim at fostering argumentation skills should be included into the curriculum to help students develop argument schemes that become activated when needed to guide comprehension. Moreover, such interventions should be designed to match the

characteristics of learners (Snow, 1989).

Requiring students to read various scientific documents on a regular basis may be a first step to allow the construction of some relevant structural knowledge, which, in turn, could help them to particularly profit from further training in the skill of argumentation.

Acknowledgements

We would like to thank Katherine Bruns and Elisabeth Marie Schmidt for helpful ideas and discussions and our students Anna Helfers and Panagiotis Karageorgos for their hard work and dedication in helping us with the development and implementation of the training interventions.

References

Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen? [How understandable are our newspapers?]* (Unpublished doctoral dissertation). University of Zürich, Switzerland.

Berthold, K., & Renkl, A. (2010). How to foster active processing of explanations in instructional communication. *Educational Psychology Review, 22,* 25–40.

Berthold, K., & Renkl, A. (2010). How to foster active processing of explanations in instructional communication. *Educational Psychology Review, 22,* 25–40.

Blair, J. A., & Johnson, R. H. (1987). Argumentation as dialectical. *Argumentation, 1*, 41–56.

Britt, M. A., & Kurby, C. A. (2005, July). *Detecting incoherent informal arguments*. Paper presented at the 15th annual meeting of the Society for Text and Discourse, Chicago, IL.

Britt, M. A., Kurby, C. A., Dandotkar, S., & Wolfe, C. R. (2008). I agreed with what? Memory for simple argument claims. *Discourse Processes, 45*, 52–84.

Britt, M. A. & Larson, A. (2003). Construction of argument representations during on-line reading. *Journal of Memory and Language, 48*, 749–810.

Britt, M. A., Perfetti, C. A., Sandak, R., & Rouet, J. F. (1999). Content integration and source separation in learning from multiple texts. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.). *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso (pp. 209–233)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Britt, M. A., Richter, T., & Rouet, J.-F. (2014). Scientific Literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104–122.

Calfee, R. & Chambliss, M. (1988). Beyond decoding: Pictures of expository prose. *Annals of Dyslexia, 38*, 243–257.

Chambliss, M. J. (1995). Text cues and strategies successful readers use to construct the gist

of lengthy written arguments. *Reading Research Quarterly, 30*, 778–807.

Chambliss, M. J., & Murphy, P. K. (2002). Fourth and fifth graders representing the argument structure in written texts. *Discourse Processes, 34*, 91–115.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/ correlation analysis for the behavioral sciences.* Mahwah, NJ: Erlbaum.

Dauer, F. W. (1989). *Critical thinking: An introduction to reasoning*. New York: Oxford University Press.

Freeman, A. M., Morris, P. B., Barnard, N., Esselstyn, C. B., Ros, E., Agatston, A., Devries, S., O'Kneefe, J., Miller, M., Ornish, D., Williams, K., & Kris-Etherton, P. (2017). Trending Cardiovascular Nutrition Controversies. *Journal of the American College of Cardiology, 69*, 1127–1187.

Fuchs, R., & Schwarzer, R. (1997). Tabakkonsum: Erklärungsmodelle und Interventionsansätze [Tobacco consumption: Explanations and interventions]. In R. Schwarzer (Ed.), *Gesundheitspsychologie: Ein Lehrbuch* (pp. 209–244). Göttingen, Germany: Hogrefe.

Galotti, K. M. (1989). Approaches to study formal and everyday reasoning. *Psychological Bulletin, 105*, 331–351.

Hefter, M. H., Berthold, K., Renkl, A., Rieß, W., Schmid, S., & Fries, S. (2014). Effects of a training intervention to foster argumentation skills while processing conflicting scientific positions. *Instructional Science, 42*(6), 929–947.

Hefter, M. H., Renkl, A., Riess, W., Schmid, S., Fries, S., & Berthold, K. (2015). Effects of a training intervention to foster precursors of evaluativist epistemological understanding and intellectual values. *Learning and Instruction*, *39*, 11–22.

Heku-IT [Computer software]. http://www.heku-it.com/schneller-lesen-app/

Inquisit (Version 3.0.6.0) [Computer software]. Retrieved from http://www.millisecond.com/

Johnson, B. T., Smith-McLallen, A., Killeya, L. A., & Levin, K. D. (2004). *Truth or consequences: Overcoming resistance with positive thinking*. In E. S. Knowles & J. A. Linn (Eds.), Resistance and persuasion (pp. 215–233). Mahwah, NJ: Erlbaum.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*, 189–209.

Jonassen, D. H. (1993). Cognitive flexibility theory and its implications for designing CBI. In S. Dijkstra, H. P. Krarnmer, & J. V. Merrienboer (Eds.), *Instructional models in computer based learning environments.* Heidelberg, Germany: Springer.

Jonassen, D. (1999). Designing constructivist learning environments. In C. M. Reigeluth (Ed.), *Instructional design theories and models* (pp. 215–239). Hillsdale, New Jersey: Lawrence Erlbaum.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149.

Larson, M., Britt, M. A., & Larson, A. (2004). Disfluencies in comprehending argumentative texts. *Reading Psychology, 25*, 205–224.

Larson, A. A., Britt, M. A., & Kurby, C. (2009). Improving students' evaluation of informal arguments. *Journal of Experimental Education, 77*, 339–365.

Mousavi, S. Y., Low, R., & Sweller , J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*, 319–334.

National Assessment of Educational Progress (1986). *The writing report card: Writing achievements in our schools*. Princeton, NJ: Educational Testing Service.

OECD (2011). *PISA 2009 Results: Students on Line: Digital Technologies and Performance (Volume VI)*. PISA, OECD Publishing. Retrieved from

http://dx.doi.org/10.1787/9789264112995-en.

OECD (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. PISA, OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264208780-en.

Osherson, D., Smith, E., & Shafir, E. (1986). Some origins of belief. *Cognition*, *24*, 197–224.

Paxton, R. J. (1997). "Someone with like a life wrote it": The effects of a visible author on high school history students. *Journal of Educational Psychology, 89*, 235–250.

Perkins, D. N. (1986). *Knowledge as design*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*, 562–571.

Perkins, D.N., Farady, M., Bushey, B. Everyday reasoning and the roots of intelligence. In: Voss, J.F., Perkins, DN., Segal, J.W., (Eds*.). Informal reasoning and education* (pp. 83-105). Hillsdale, NJ: Erlbaum, 1991.

Renkl, A. (2009). Wissenserwerb [Knowledge acquisition]. In E. Wild & J. Möller (Eds.). *Pädagogische Psychologie* (pp. 3–26). Berlin: Springer.

Rouet, J. - F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction, 15*, 85–106.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching, 41*, 513–536.

Schroeder, S., Richter, T., & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language, 59*, 237–255.

Shaw, F. W. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning, 2*, 51–80.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework of research in individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 13–59). New York, NY: Freeman.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 343–365). New York: Oxford University Press.

Stiftung Warentest (2015). Lesetrainings im Test: Wie Sie zum Schnellleser werden [Reading trainings tested: How to become a fast reader]. Retrieved from https://www.test.de/Lesetrainings-im-Test-Wie-Sie-zum-Schnellleser-werden-4817442-0/

Suppe, F. (1998). The structure of a scientific paper. *Philosophy of Science, 65*(3), 381–405.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, MA: Cambridge University Press.

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). Judging the plausibility of argumentative statements in scientific texts: An Expert – Novice Comparison. *Thinking and Reasoning, 22*, 221–246.

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*, 337–350.

Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 83–106). Hillsdale, NJ: Erlbaum.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the Myside Bias in written argumentation. *Written Communication, 26*, 183–209.

Table 1

Experiment 1. Summary of Nested Multiple Regression Analyses for Variables Predicting Posttest Performance after the Training Intervention.

Variable	*B*	SE_B	*t* (*df*)	*F* (df_h, df_e)	R^2	ΔR^2
Step 1				3.45 (3,45)*	.19	
Intercept	.71	.02	29.56***(46)			
Training condition (TC)	.04	.02	1.50(46)			.04
Pretest accuracy (PA)	.06	.03	2.50**(46)			.10
PA* TC	.03	.03	1.01(46)			.03
Step 2				6.09 (5,43)***	.41	
Intercept	.70	.02	33.93***(44)			
Training condition (TC)	.04	.02	1.65(44)			.04
Pretest accuracy (PA)	.04	.02	1.88*(44)			.05
PA* TC	-.01	.02	.39(44)			.00
Study performance (SP)	-.08	.02	-3.45**(44)			.17
SP* TC	-.05	.02	2.08*(44)			.06

Note. $N = 50$. Pretest accuracy and study performance were z-standardized. Training condition was included as contrast-coded predictor (1: argument structure training, -1: speed-reading control condition). Note that lower values represent better performance in the German grading system.

*$p < .05$. **$p < .01$ ***$p < .001$. One-tailed tests were used to test for effects of training condition and study performance, and for the interaction of study performance and training condition.

Table 2

Experiment 2. Summary of Nested Multiple Regression Analyses for Variables Predicting Posttest Performance after the Training Intervention.

Variable	*B*	SE_B	*t* (*df*)	*F* (df_h, df_e)	R^2	ΔR^2
Step 1				4.24 (3,14)*	.48	
Intercept	.63	.08	8.28 (18) ***			
Training condition (TC)	.16	.08	2.08 (18) *			.21
Pretest accuracy (PA)	-.19	.08	-2.46 (18) *			.12
PA* TC	-.02	.08	-.25 (18)			.02
Step 2				21.60(5,12)***	.90	
Intercept	.61	.04	16.91 (16) ***			
Training condition (TC)	.13	.04	3.60 (16) **			.11
Pretest accuracy (PA)	-.15	.04	-3.73 (16) **			.12
PA* TC	.07	.04	1.63 (16)			.02
Study performance (SP)	-.29	.04	-6.86 (16) ***			.39
SP* TC	-.04	.04	-1.01 (16)			.01

Note. $N = 22$. Pretest accuracy and study performance were z-standardized. Training condition was included as contrast-coded predictor (1: plausibility training, -1: speed-reading control condition). Note that lower values represent better performance in the German grading system.

*$p < .05$. **$p < .01$ ***$p < .001$. One-tailed tests were used to test for effects of training condition and study performance, and for the interaction of study performance and training condition.

Table 3

Examples of Plausible and Implausible Training Items in the Practice Session

Sentence type	Sentence
Plausible	It is likely that people who live in student homes find each other, because they share several social and personal attributes, such gender, age, socio-economic status, birth family, religion, or political attitudes, which may decrease physical and social distance.
Implausible	
False conclusion	Friends talk more frequently about their thoughts and feelings, and therefore show less emotional openness.
Wrong example	We are equipped with a complex and finely graduated repertoire of verbal forms of behaviours and responses (e.g., mimicry, different forms of eye contact, gestures) that are used to initiate, regulate, and control interpersonal contacts.
False dichotomy	Different theories emphasise adversative aspects of the interaction process (i.e., emotional versus sentimental factors, respectively) and therefore fail to deliver a satisfying explanation for the development of interpersonal attraction.
Circular reasoning	According to the balance theory, interaction partners are more likely to feel attracted to each other if they agree in their opinions with regard to certain persons, issues, objects, and events, because interpersonal liking increases.
Overgeneralisation	If motivation is high, friendships often lead to an increase in performance. If motivation is low, however, and if performance goals deviate, friendships always hinder group performance.

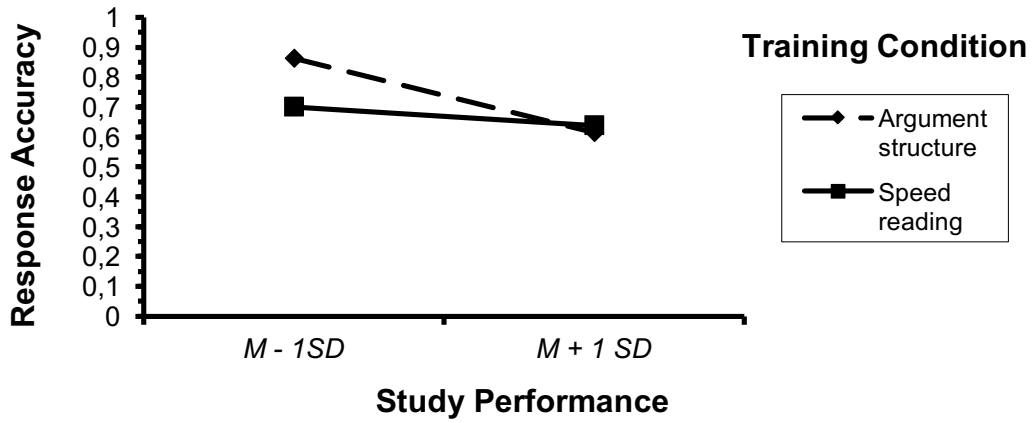

Figure 1. Estimates of the simple slopes (with standard errors) of the effect of study performance on posttest accuracies, after participating in the argument structure training or the control training (speed-reading training).

Appendix A. Sample item for a typical argument (Argument 1) and a less typical argument (Argument 2) used in the video tutorial (translated from German)

Argument 1

We can help children who suffer from nightmares with a simple method [*claim*]. In this regard, results from a study showed that repeated drawings of the threatening content in the nightmare (e.g., Dracula) and tearing them apart afterwards made the nightmare disappear [*datum*]. The procedure is simple, because it can be easily used by parents [*warrant*]. For example, parents can integrate it as part of their daily night-time routines [*backing*]. For the method to be successful, however, it is important that the child is ready to confront herself with her fears [*rebuttal*].

Argument 2

Results from a study indicate that women showing a confident, dominant appearance during the application procedure receive lower ratings for their social competence [*datum*]. Social competence is in great demand on the job market [*warrant*]. In a different study, it was found that social competence was an important criterion in 70% of application processes [*backing*]. Thus, high expectations for social competence may ironically create more discrimination in women [*claim*], provided they present themselves as confident career women [*rebuttal*].

Chapter VIII

General Discussion

Chapter VIII: General Discussion

Students entering university are required to deal competently with the various scientific claims and evidence they encounter in different scientific documents (Britt et al., 2014). However, as described in **Chapters 1** and **3**, the ability to competently evaluate scientific texts and arguments constitutes a great challenge for many students and most of them have never received formal training (Britt et al., 2014; Kuhn, 1991; Perkins et al., 1991). Furthermore, a substantial body of evidence shows that students' evaluations of scientific texts and arguments often fail to meet normative standards (e.g., Britt et al., 2008; Larson et al., 2009; Norris et al., 2003; Shaw, 1996; Wolfe & Kurby, 2017). In addition, students rarely spontaneously use source information for evaluation (e.g., Goldman et al., 2012; Wiley et al., 2009; Wineburg, 1991). However, prior research also shows that these deficits may be successfully addressed with appropriate interventions (e.g., Larson et al., 2004, Larson et al., 2009). The empirical studies presented in this dissertation further investigated the strategies that underlie a valid understanding of scientific texts (Richter, 2003, 2011; Richter & Schmid, 2010) in the domain of psychology. In Study 1 and Study 2, introductory psychology students' performance to systematically evaluate the plausibility of informal arguments (**Chapter 5**), as well as their ability to heuristically judge the credibility of multiple scientific texts (**Chapter 6**), was compared to the performance of scientists in the domain of psychology. Furthermore, some of the most prevalent challenges students demonstrated in these expert-novice comparisons were addressed in a training intervention (Study 3). In particular, students received training in argument structure based on Toulmin's (1958) argumentation model (Experiment 1, **Chapter 7**) or in normative aspects of argument evaluation and strategies for recognising common fallacies (cf. Blair & Johnson, 1987; Shaw, 1996; Experiment 2, **Chapter 7**). In the following sections, the main results of the present research are summarised and discussed against the background of existing theories and

literature. In addition, limitations and implications of the present research for education and practice are described.

Summary and Discussion of Expert-Novice Comparisons

The expert-novice comparisons investigated university students' competences to systematically evaluate the plausibility of informal arguments (**Chapter 5)** and their ability to heuristically judge the credibility of multiple scientific texts (**Chapter 6)**, whereby their performance was compared to the performance of scientists, using think-alouds. Results from Study 1 show that many students, compared to scientists, struggled to form normatively accurate evaluations. In particular, they often failed to evaluate the internal consistency of arguments and did not recognise common fallacies. Instead, they frequently relied purely on their intuition or opinion regarding the acceptability of a claim. Scientists, in contrast, predominantly judged the internal consistency of arguments. These differences in strategy use partly mediated the performance differences between both groups. In addition, results from Study 2 revealed that, although students did show some source awareness, as they rated primary sources as the most trustworthy documents, they insufficiently used source information as a criterion for evaluation, even when judgements had to be formed within a relatively short time limit. Scientists, on the other hand, used source information as a major criterion for evaluation when time constraints did not allow systematic evaluations. Although they also evaluated the content of the texts to some extent (e.g., relevance, choice of research method), their superior performance was fully mediated by their use of source information. Thus, scientists were able to apply different strategies in a flexible manner, depending on the processing goal: when systematic evaluations were possible, they used more sophisticated normative criteria for their evaluations and carefully scrutinised the arguments. When the conditions for evaluation did not allow systematic processing, they primarily used source characteristics for their evaluations.

Useful Strategies for the Evaluation of Informal Arguments and Scientific Texts

The results from the expert-novice comparisons indicate that a competent evaluation of scientific texts and arguments seems to involve both systematic and heuristic processing strategies, depending on processing goals and task conditions (Pressley, 2000; Rouet & Britt, 2011; Wyatt et al., 1993) and raise some doubt on traditional dual-process models, such as the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986) or the Heuristic-Systematic Model (HSM; Chen & Chaiken, 1999) that would assume scientists to apply only systematic strategies. Our findings indicate that scientists possess a large variety of strategies that are successfully used in different tasks. Although heuristics related to the source may not always lead to accurate judgements about a document's credibility, forming a quick preliminary judgement about a document's credibility based on characteristics of its source likely precedes a more elaborated analysis of a document and can help the reader to select more reliable sources in the first place (cf. Bråten et al., 2011; Strømsø et al., 2009). When systematic evaluations are possible, evaluations that consider not only the accuracy of arguments, but also aspects of their internal consistency, such as relevance and sufficiency (Blair & Johnson, 1987), seem to lead to more accurate plausibility judgements (cf. Shaw, 1996; Voss & Means, 1991). In addition, performances in both heuristic and systematic tasks were positively related, indicating that both competences might be part of a common construct of scientific literacy (Britt et al. 2014).

How scientists acquire their expertise is not entirely clear. It seems likely that domain experts, such as psychologists or historians, develop their expertise in the course of their academic socialisation and disciplinary practice (Britt et al., 2014). Presumably, their experience with various scientific documents allows them to form richer and more flexible mental representations that can be used for evaluation. In terms of the documents model framework (Britt et al., 1999), scientists are likely to construct more elaborate representations of content and source, allowing them to select appropriate strategies in a flexible manner and

draw inferences about the plausibility and credibility of a text (cp. Bråten et al., 2009; Wineburg, 1991). However, more research is needed to further examine the factors involved in the development of scientific literacy.

The Challenges among Students to Evaluate Scientific Texts and Arguments

The results from the expert-novice comparisons are in line with findings from previous research showing that students often base their evaluations mainly on prior attitudes and beliefs rather than a systematic epistemic approach and neglect relations between argument components (e.g., Britt & Kurby, 2005; Larson et al., 2004; Shaw, 1996; Wolfe et al., 2009; Wolfe & Kurby, 2017). Furthermore, our results are in accordance with other research showing that students rarely spontaneously attend to source information or use source information for evaluation without explicit instruction or training (e.g., Barzilai et al., 2015; Bråten et al., 2016; Britt & Aglinskas, 2002; Gerjets et al., 2011; Goldman et al., 2012; Korpan et al., 1997; Stadtler & Bromme, 2007; Stahl et al., 1996; Wiley et al., 2009; Wineburg, 1991). Failures to activate relevant argument schemes, a lack of knowledge about normative criteria for evaluation, and little familiarity with different publication types might be related to these difficulties. Our results showed that students struggled with the correct identification of different argument components, especially warrants. One explanation for this finding could be that, whereas most students become familiar with narratives early in life, knowledge about expository texts, such as arguments and scientific texts, is not learned in the same way (Britt et al., 2014). Arguments have a difficult structure, because they require the linking of claims to supporting reasons or evidence, and people do not spontaneously acquire knowledge about arguments. Importantly, the ability to identify structural components of arguments was positively correlated with the ability to judge the plausibility of arguments and to identify common fallacies. Assuming that competent readers use their prior knowledge about the structure of arguments to form accurate judgements about their quality (Britt et al., 2014; Britt & Larson, 2003; Halpern, 1998; Hummel and Holvoak, 1997; Suppe, 1998; Wolfe

et al., 2009), students might need to be equipped with relevant knowledge about the structure of arguments first, before they are able to initiate elaborative epistemic processes and form competent evaluations (Richter, 2011). Moreover, systematic evaluations are effortful (Shaw, 1996). When readers evaluate the internal consistency of an argument, they not only need to access and represent relevant structural knowledge about arguments from memory, but keep this information activated in working memory. This likely requires practice and relevant domain knowledge (Britt et al., 2014). Judgements that focus on accuracy only, rather than relevance or sufficiency (cf. Blair & Johnson, 1987), only require relatively simple consistency checking between the content of a text and prior knowledge (Richter, et al., 2009). These processes are memory-based (Gerrig & McKoon, 1998; O'Brien et al., 1998) and can occur without much cognitive effort (epistemic monitoring; Isberner & Richter, 2014; Richter & Maier, in press). Presumably, students predominantly engaged in such epistemic monitoring processes, rather than more elaborative evaluations, because they lacked relevant prior knowledge about normative criteria for evaluation and experience with different scientific texts and arguments that could have been used for evaluation (cf. Britt et al., 2014). Whereas students struggled with the identification of implausible information, they were much more accurate at recognising plausible information, possibly because such information was consistent with their current mental model (Johnston-Laird, 1994; Richter & Maier, in press).

The students in our study also had difficulty to allocate different documents to a genre, and the ability to assign different documents to a genre was positively related to the ability to determine the credibility of these documents. Thus, introductory students might not be very familiar (yet) with reading multiple scientific texts, making it difficult for them to use important source features, such as document type, for evaluation. Previous research indicates that reading multiple sources is positively related to sourcing activities (e.g., Britt & Aglinskas, 2002; Nokes et al., 2007). Similarly, Tabak (2016) noted that students' failures to

critically evaluate a document's source might be related to their exclusive experience with textbooks and their strong trust in these textbooks as an "authority" that can be taken at face value. In addition, some research indicates that information about sources may not always be easily found by students and that sourcing requires additional skills and practice, such as relevant searching techniques (Britt & Gabrys, 2002). Thus, relevant practice and training is likely to be necessary to foster students' sourcing skills.

In sum, the results from the expert-novice comparisons indicate that, although introductory students do validate scientific information against their knowledge and beliefs, their evaluations of scientific texts and arguments often do not meet normative standards, and students struggle with competent evaluations in different heuristic or systematic situations. Their judgements are inaccurate, in part, because their use of strategy is immature. They do not pay enough attention to the internal consistency of arguments and fail to use source information for evaluation. These difficulties might be related to their lack of knowledge about the structure of arguments, normative criteria for evaluation, common fallacies, and /or their inexperience with different publication types. Scientists, in contrast, seem to be able to form competent judgements, and our results indicate that scientific literacy may require some general competences that can be used in a flexible manner. Our results suggest that both systematic and heuristic strategies are important for evaluation and highlight the need for explicit training and instruction (cf. Kuhn, 1991; Norris et al., 2003; Perkins et al., 1991; Phillips & Norris, 1999). Our results also indicate that it might be useful to adopt the strategies used by domain experts for designing appropriate curricular or interventions for students, and teach relevant genre and domain knowledge that is involved in a successful evaluation of scientific literature.

Summary and Discussion of the Intervention Study

Recently, there has been an increasing interest in fostering students' sourcing skills (Bråten et al., in press), and a number of interventions, such as the Sourcer's Apprentice (Britt

& Aglinskas, 2002), or the SEEK web tutor (Graesser et al., 2007; Wiley et al., 2009) have been successfully applied to increase students' source awareness when dealing with multiple documents. In contrast, the number of interventions that have addressed students' evaluations of informal arguments is still rather limited (Larson et al., 2009). Therefore, acknowledging the results from the expert-novice comparisons, our intervention (**Study 3**) concentrated on teaching knowledge about the structure of informal arguments (Experiment 1) and normative criteria for evaluation in addition to common fallacies (Experiment 2).

Results from Experiment 1 show that teaching the structure of Toulmin's (1958) argumentation model was successful for improving the comprehension of complex and less typical arguments, and for understanding relations between key components (i.e. warrants), as compared to a speed-reading control group, suggesting that acquisition of conceptual and procedural knowledge about the structure of informal arguments may have helped with the formation of more accurate representations of key components of arguments, including warrants. In terms of the mental model theory (Johnson-Laird, 1983), we assume that knowledge about different argument components, including components signalling relations between key components (i.e. warrants), allowed students to construct more accurate representations of arguments and to simultaneously activate different argument components in memory (Britt et al., 2014; Shaw, 1996). In addition, the inclusion of explicit markers may have helped students to represent different components (Britt & Larson, 2003). Results from Experiment 1 also show that most students were relatively accurate at recognising some argument components, such as rebuttals, and less complex arguments that did not contain warrants, prior to the intervention, presumably because students were more familiar with these argument components. Whereas the intervention could not further improve the ability to recognise all parts of Toulmin's Model in all students, students with above average grades who took part in the argument structure training achieved significantly higher accuracy values after the intervention. Thus, better students profited the most from our training intervention.

Similarly, students who were initially able to recognise more complex argument types could further improve this ability in the intervention. Presumably, high achieving students were more familiar with a variety of scientific documents, which might have (implicitly) provided them with some relevant background knowledge (i.e. *discipline expertise*; Rouet et al., 1997) about the structure of arguments. This knowledge, in turn, might have facilitated comprehension, integration, and application of the training materials, and possibly improved their argumentation skills as well (Britt et al., 2014; Rouet et al., 1997). Research by Gil, Bråten, Vidal-Abarca, and Strømsø (2010) provides further support for the notion that individual differences in prior knowledge may be responsible for the observed interaction between training in argument structure and study performance. Gil et al. (2010) had undergraduate students work with multiple documents on a science topic and found that only those with high prior knowledge were able to take advantage of instructions to construct arguments during reading, whereas students with little prior knowledge were more hindered than helped by such task instructions. Other research indicates that good learners may particularly profit from CLEs, whereas guided and more structured interventions might be beneficial for struggling readers or readers with less prior knowledge (see Kalyuga, 2007 for a review). Thus, training in argument structure can be successful for improving students' understanding of arguments, but instructions should be designed to match learners' abilities.

Although training of argument structure improved students' understanding of informal arguments, it did not improve their skills to form normatively accurate evaluations about the plausibility of arguments. Therefore, Experiment 2 tested the effects of a training that conveyed knowledge about arguments and, additionally, knowledge about the normative standards required for a competent evaluation of informal arguments (cf. Blair & Johnson, 1997; Shaw, 1996; Voss & Means, 1991), along with conceptual knowledge about common fallacies (cf. Dauer, 1989; Johnson & Blair, 1977; Schroeder et al., 2008). The training could

improve students' performance to evaluate the plausibility of arguments and recognise common fallacies in the experimental group, compared to a speed-reading control group.

Some researchers have argued that teaching fallacies can bias students to see more fallacies than there are and, therefore, suggested to focus exclusively on teaching normative criteria for evaluating arguments (e.g., Hitchcock, 1995). However, in response to that view, Blair (1995) argued that, if explanations are provided for why fallacies are erroneous in combination with normative aspects for evaluation, they can be effective for fostering students' reasoning skills. Our results provide some support for Blair's view. Although it is not entirely clear which aspects of the training were responsible for the improved performance in this group, the intervention was successful and we did not observe any biases. Moreover, the fallacies were refuted and replaced with more plausible explanations that could be integrated into the reader's mental model, and such refutations have been shown to generally improve comprehension and even change relatively stable misconceptions (e.g., Kendeou et al., 2013). Thus, we assume that the inclusion of fallacies in the tutorial, in addition to conceptual knowledge about arguments and normative criteria for their evaluation, was a reasonable approach.

In line with previous research indicating that students tend to neglect relations between argument components (e.g., Britt & Kurby, 2005; Larson et al., 2004; Larson et al., 2009; Shaw, 1996, Wolfe & Kurby, 2017; Wolfe et al., 2009; Wu & Tsai, 2007), our results also indicate that shifting students' attention towards the internal consistency of arguments may be helpful for improving their reasoning skills. Both training in argument structure and teaching normative criteria for the evaluation of arguments seem to be effective in overcoming these deficits (cf. Hefter et al., 2014; Larson et al, 2004, Larson et al., 2009).

Importantly, in Experiment 2, transfer effects to the ability to correctly identify structural components of arguments were found, indicating that teaching normative criteria may also be a useful approach for the formation of more accurate representations of key

components of arguments, including warrants, and thus foster students' epistemic competences more generally. On the other hand, teaching structural components of arguments alone did not improve students' competences to evaluate the plausibility of arguments. Thus, both structural knowledge about arguments and knowledge about normative criteria for evaluating arguments may be required to improve students' understanding of scientific arguments.

CLEs that use manipulative environments and a combination of different tools, such as expert-based video tutorials, practical exercises, instructional prompts, and immediate feedback, may be useful for encouraging active construction of knowledge (cf. Chi et al., 1992; Jonassen, 1999; Renkl, 2009). These elements were perceived as very helpful by many participants. However, our results also indicate that better learners might profit more from such constructivist settings than less able learners or those with less prior knowledge (cf. Kalyuga, 2007; Snow, 1989). Although study performance did not moderate the training effect on posttest performance in Experiment 2, it did affect performance in both experiments. Results from recent PISA studies provide further support for the notion that instructions should be designed to match learners' abilities. For example, students who reported that their teachers adapted the lessons to their students' individual needs, achieved higher scores for scientific literacy (OECD, 2014).

In sum, our results indicate that interventions that focus on teaching argument structure, normative aspects of argument evaluation, and knowledge about common fallacies, can be useful means to improve university students' epistemic competences. Such trainings may be particularly useful for fostering knowledge about more complex and less typical arguments and for sensitising students to pay increased attention to relational aspects between key components (i.e. warrants in Experiment 1 and argument strength or internal consistency in Experiment 2). Better learners might particularly profit from interventions that are set in CLEs.

Limitations of Reported Studies and Directions for Future Research

The empirical work presented in this dissertation revealed important findings that contribute to our understanding of undergraduate students' epistemic competences. However, there are several limitations that should be considered when interpreting our results.

First, the reported studies were based on cross-sectional data and the possibility of uncontrolled group differences cannot be ruled out. Given that the results were consistent with our predictions and with similar research from other domains (e.g., Larson et al., 2009; Larson et al., 2004; Rouet et al., 1997; Shaw, 1996; Wineburg, 1991), it seems unlikely that the observed group differences were merely due to uncontrolled group differences. However, longitudinal studies are needed to explore the precise conditions under which epistemic competences develop. For example, some studies indicate that graduate students use more sophisticated strategies for evaluation, such as relevance (e.g., Britt & Kurby, 2005), and pay increased attention to sources (e.g., Rouet et al., 1996), compared to undergraduates. Korpan et al. (1997) found that the number of university courses completed correlated with students' ability to evaluate research methodologies. Thus, student's reasoning skills might improve as a result of engaging in different learning activities at university. Yet, it is reasonable to assume that stronger and much faster improvement can be achieved by targeted instruction based on explicit instruction and practice. As pointed out by several others (e.g., Kuhn, 1991; Osborne et al., 2004; Perkins et al., 1991), epistemic competences generally do not develop on their own, but need to be taught explicitly through suitable instruction. Our samples were also relatively small, especially in Experiment 2. Future studies should use a larger number of participants to increase the power of our results. They should also provide more stringent tests of the dimensional structure of epistemic strategies using item-response models. In addition, manipulations that include or exclude different tools and measures tapping into cognitive processes may be helpful to more closely examine the precise mechanisms under which readers develop a reasoning scheme.

Second, our studies examined epistemic competences in introductory university students (Study 1 and Study 2) and university undergraduates (Study 3). Whereas this may be a reasonable approach, given that this group of students are mainly confronted with the challenges described in **Chapter 1**, future studies should replicate our studies with younger and less advanced students (e.g, high school and college students), or with more advanced students (e.g., graduate students) to examine the extent to which our interventions would be effective for these students as well. In addition, our studies should be replicated in other countries, as it cannot be ruled out that the effects were to some extent culture-specific (cf. Hornikx & Hoeken, 2007; Hornikx & ter Haar, 2013).

Third, future research should further examine how evaluation processes are influenced by different individual and contextual variables. Although we assume that differences in prior knowledge are likely to be responsible for the superior performance of scientists in the expert-novice comparisons and for the observed interaction of training condition and study performance in the argument structure training (cf. Rouet et al., 1997), it cannot be ruled out that other factors, such as differences in cognitive ability, general thinking disposition, or motivation contributed to the superior performance of scientists and high achieving students. For example, scientists and better learners usually have an internal, enduring interest in reading (Alexander, 2016) and, although Stanovich (1999, 2012) found that the skill of rational thinking seems to be relatively independent of intelligence, some other research indicates that differences in cognitive ability might influence the ability to reason about informal arguments or to use source information for evaluation (cf. Bråten et al., 2011; Goldman et al., 2012; Weinstock et al., 2006). Moreover, West, Toplak, and Stanovich (2008) found that critical thinking skills may be related to more general thinking dispositions, such as active open-mindedness and need for cognition. In addition, future research should consider possible influences of epistemological beliefs on evaluation processes (cf. Ferguson, 2015). For example, students who believe that knowledge is uncertain and evolving may pay more

attention to alternative viewpoints (Barzilai & Eshet-Alkalai, 2015), and those who believe that knowledge should be justified by several sources may show a higher ability to deal with multiple documents (e.g., Bråten et al., 2014). In contrast, relying on one's personal opinions rather than those of experts may detract attention away from external sources and hinder critical evaluations of a document's source (Barzilai et al., 2015).

Strategy use and evaluation processes also depend on the materials provided (Britt & Rouet, 2011; Dole et al., 1991). For example, the documents presented in our study did not contain any conflicting information. Including conflicting information rather than supplementary texts might have facilitated understanding and source awareness (cf. Braasch et al., 2012; Kammerer et al., 2016; Strømsø & Bråten, 2014). Futhermore, Salmerón and Bråten (2018) recently examined the influence of reading real documents, such as books, rather than excerpts, and found that reading real documents increased students' sourcing activities. Thus, the results that were found in our studies and in previous studies might underestimate students' capacities to pay attention to sources when dealing with real texts, rather than excerpts, to some extent. Our results should also not be generalised to other genres, such as narratives, as other genres often involve different kinds of processing (e.g., Zwaan, 1994). Moreover, a limited number of scientific topics from the domain of psychology were used in our studies. Topic knowledge has been shown to influence evaluation processes (e.g., Bråten et al., 2011). Similarly, introductory students often hold strong prior beliefs about certain topics and favour one position of a scientific debate over the other (Richter, 2015), which can influence validation processes (e.g., Bråten et al., 2016; de Pereyra et al., 2014; Stanovich & West, 1997; Wolfe, 2012; Wolfe et al., 2009; Wolfe & Kurby, 2017). Therefore, future research should include a larger variety of topics and more closely examine students' topic beliefs. Additionally, integrating materials that do not deal with the participants' domain of discipline expertise may be useful for identifying possible transfer effects. The finding that the students and scientists in our study and those from other

countries and domains (cf. Rouet et al., 1997; Shaw, 1996; Wineburg, 1991) applied similar strategies provides some support for the notion that epistemic competences and scientific literacy may be acquired in a relatively fundamental, rather than derived sense that does not strongly depend on a specific content domain (Norris & Phillips, 2003; see also Dole et al., 1991; Halpern, 1998; Lehman & Nisbett, 1990; Pearson et al., 1992; Yore et al., 1998; Yore et al., 2007). Yet, our results also indicate that appropriate genre and domain knowledge, such as knowledge about the structure of arguments, normative criteria, and knowledge about different publication types, is likely to be necessary for forming competent judgements about the quality of scientific texts and arguments (Goldman & Bisanz, 2002; Yore et al., 2003). More research is needed to further examine how individual and contextual factors influence evaluation processes.

Fourth, our studies examined students' epistemic competences in very specific conditions. A complete account of students' evaluations should not only adapt tasks to different learning contexts and adapt materials to individual needs (Barzilai & Strømsø, 2018; Snow, 1989), but also aim at the development of a full documents model that represents the content of the multiple texts, information about the (credibility of the) source, and the argumentative relationships between different texts (Perfetti et al., 1999). Such an account should foster students' corroboration, integration, and search skills as well (Britt & Gabrys, 2002). Although epistemic strategies are particularly relevant for dealing with scientific texts, receptive strategies should compliment these strategies. For example, students need to know how to locate relevant source information quickly. Recently, Goldman et al. (2016) developed a conceptual framework for disciplinary literacy. According to Goldman et al. (2016), such a framework should consider five core constructs in science education: Epistemology (1), inquiry practices and reasoning strategies (2), overarching concepts and principles (3), forms of information representations/ types of texts (4), and discourse and language structures (5). Knowledge about how (scientific) knowledge is derived is considered central in their

approach and acknowledging all five core constructs might be necessary to achieve higher levels of scientific literacy among students.

Finally, it should be noted that the results observed in our intervention could not be replicated four weeks later, indicating that a single session may not be enough to produce long-term effects, or even allow transfer of skills to other tasks, such as constructing arguments or writing coherent research papers. Expertise takes time to develop (Britt et al., 2014; Ericsson et al., 1993). Therefore, future research should design more extensive interventions that include several practice sessions and integrate such interventions into the curriculum. For example, van Gelder, Bissett, and Cumming (2004) found that university students who actively practiced informal reasoning skills for 12 weeks significantly improved this ability. Multiple exposures that include several (longer) intervals seem to be ideal to foster learning (*distributed learning*; e.g., Glenberg, 1979). Encouraging students to consult multiple documents when reading about scientific issues might be a first step to help them deal more competently with scientific literature and to particularly profit from training in argumentation. Some research suggests that using adapted primary scientific literature in education may help students to better represent different argument components (Norris & Phillips, 2009). Although such interventions were mainly designed for high school students, they might be useful for introductory psychology students as well to familiarise them with the structure of scientific texts and arguments. Educators should also create structures that facilitate critical thinking around the campus, such as academic advising offices, and integrate relevant courses into the curriculum (King & Kirchener, 1994).

Conclusion

The aim of this dissertation was to examine the strategies involved in a successful evaluation of scientific texts and arguments and to design suitable interventions to foster epistemic competences in university students. As expected, students showed deficits in their ability to systematically evaluate the plausibility of informal arguments and insufficiently

used source information as a criterion for determining document credibility, even when the processing goal was heuristic. Our findings also indicate that both systematic and heuristic strategies seem to be involved in the development of scientific literacy and discipline expertise and that understanding the propositional structure of arguments and, in particular, relations between key components, seems to be an important prerequisite for their successful evaluation. Teaching knowledge about the structure of arguments improved students' understanding of more complex and less typical arguments and this training was most effective for high achieving students. In addition, conveying knowledge about normative evaluation criteria and common fallacies was helpful for improving students' abilities to evaluate the plausibility of arguments. This training also improved their understanding of the rhetorical structure of arguments, indicating that such knowledge may lead to more accurate and flexible representations in memory, allowing students to access this knowledge when dealing with the numerous claims and evidence in scientific texts and activate knowledge about relational aspects of argument components when needed.

Our results suggest that not much seems to have changed since Perkin's (1985) appeal to focus more strongly on developing students' epistemic competences in educational practice. To accomplish scientific literacy among students, a change of culture towards a more argumentative dialogue in classrooms that values controversy as a means for understanding is inevitable. Ultimately, a dynamic and changing society, in which biased and inaccurate information is readily accessible, in which non-experts determine political decisions, and in which companies are constantly trying to influence people to buy new products, needs people who are able to reflect on what they hear, see, and read, who take a critical stance in public debates, who can adapt flexibly to changing situations and make informed decisions, and who are able to challenge current views and conceptions, if necessary. I would like to end this chapter with a critical quote from an influential and inspiring teacher, Jiddu Krishnamurti (Krishnamurti, 1969, p.15), whose words reminded me not to forget the whole picture:

“You have to be your own teacher and your own disciple. You have to question everything that man has accepted as valuable, as necessary.”

References

Alexander, P. A. (2016). The arguments for and the reasoning about epistemic cognition: A response to the chapters on psychological perspectives. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Handbook of epistemic cognition* (pp. 100–110). New York: Routledge.

Alexander, P. A., & Fox, E. (2011). Adolescents as readers. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol. IV, pp. 157–176). New York: Taylor & Francis.

Allen, E. S., Burke, J. M., Welch, M. E., & Rieseberg, L. H. (1999). Reliability of science on the web. *Nature, 402*, 722.

Anderson, J. (1991). Cognitive Architectures in a rational analysis. In K. VanLehn (Ed.), *Architectures for Intelligence* (pp. 1–24). Hillsdale, NJ: Erlbaum.

Anderson, C. A., Lepper, M. R., & Ross, L. R. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*, 1037–1049.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 255–291). New York: Longman.

Anderson, R. C., Spiro, R., & Anderson, M. C. (1978). Schemata as scaffolding for the representation of information in connected discourse. *American Educational Research Journal, 15*, 433–440.

Baker, L. (1985). Differences in the standards used by college students to evaluate their comprehension of expository prose. *Reading Research Quarterly, 20*, 297–313.

Baron, J. (1991). Beliefs about thinking. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 169–186). Hillsdale, NJ: Erlbaum.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning, 1*, 221–

235.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.

Barzilai, S., & Eseth-Alkalai, Y. (2015). The role of epistemic perspectives in comprehension of multiple author viewpoints. *Learning and Instruction, 36*, 86–103.

Barzilai, S., Tzadok, E., Eshet-Alkalai, Y. (2015). Sourcing while reading divergent expert accounts: Pathways from views of knowing to written argumentation. *Instructional Science, 43*, 737–766.

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication, 2*, 3–23.

Berthold, K., & Renkl, A. (2010). How to foster active processing of explanations in instructional communication. *Educational Psychology Review, 22,* 25–40.

Blair, J. A. (1995). Premise adequacy. In F. H. van Eemeren, R. Grootendorst, J. A. Blair, & C.A. Willard (Eds.), *Perspectives and approaches. Proceedings of the third ISSA conference on argumentation* (Vol. II, pp. 191–202). Amsterdam: SieSat.

Blair, J. A., & Johnson, R. H. (1987). Argumentation as dialectical. *Argumentation, 1*, 41–56.

Blanchard, J. S., & Samuels, S. J. (2015). Common core state standards and multiple-source reading comprehension. In P. D. Pearson, & E. H. Hiebert (Eds.), *Research-based practices for teaching common core literacy* (pp. 93–105). New York: Teachers College Press.

Braasch, J. L. G., Bråten, I., Britt, M. A., Steffens, B., & Strømsø, H. I. (2014). Sensitivity to inaccurate argumentation in health news articles: Potential contributions of readers' topic and epistemic beliefs. In D. N. Rapp, & J.L.G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 117–137). Cambridge, MA: The MIT Press.

Braasch, J. L. G., Goldman, S. R., & Wiley, J. (2013). The influences of text and reader characteristics on learning from refutations in science texts. *Journal of Educational*

Psychology, 105, 561–578.

Braasch, J.L.G., Rouet, J-F., Vibert, N., & Britt, M.A. (2012). Readers' use of source information in text comprehension. *Memory & Cognition, 40*, 450–465.

Brand-Gruwel, S., & Wopereis, I. (2006). Integration of the information problem-solving skill in an educational programme: The effects of learning with authentic tasks. Technology, Instruction. *Cognition and Learning, 4,* 243–263.

Bråten, I., Britt, M.A., Strømsø, H. I., & Rouet, J. F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. *Educational Psychologist, 46*, 48–70.

Bråten, I., Ferguson, L.E., Strømsø, H.I., & Anmarkrud, Ø. (2014). Students working with multiple conflicting documents on a scientific issue: Relations between epistemic cognition while reading and sourcing and argumentation in essays. *British Journal of Educational Psychology, 84*, 58–85.

Bråten, I., Salmerón, L., & Strømsø, L. (2016). Who said that? Investigating the plausibility-induced source focusing assumption with Norwegian undergraduate readers. *Contemporary Educational Psychology, 46*, 253–262.

Bråten, I., Stadtler, M., & Salmerón, L. (in press). The role of sourcing in discourse comprehension. In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *Handbook of discourse processes* (Vol. II). New York: Routledge.

Bråten, I., Strømsø, H.I., & Andreassen, R. (2016). Sourcing in professional education: Do text factors make any difference? *Reading and Writing, 8,* 1599–1628.

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*, 6–28.

Bråten, I., Strømsø, H. I., & Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learning and Instruction, 21*, 180–192.

Brem, S.K., Russell, J., & Weems, L. (2001). Science on the Web: Student evaluations of scientific arguments. *Discourse Processes, 32*, 191–213.

Britt, M. A., & Aglinskas, C. (2002). Improving student's ability to use source information. *Cognition and Instruction, 20*, 485–522.

Britt, M. A., & Gabrys, G. (2002). Implications of document-level literacy skills for Web site design. *Behavior Research Methods, Instruments, & Computers, 34,* 170–176.

Britt, M. A., & Kurby, C. A. (2005, July). *Detecting incoherent informal arguments*. Paper presented at the 15th annual meeting of the Society for Text and Discourse, Chicago, IL.

Britt, M. A., Kurby, C. A., Dandotkar, S., & Wolfe, C. R. (2008). I agreed with what? Memory for simple argument claims. *Discourse Processes, 45*, 52–84.

Britt, M. A., & Larson, A. (2003). Construction of argument representations during on-line reading. *Journal of Memory and Language, 48*, 749–810.

Britt, M. A., Perfetti, C. A., Sandak, R., & Rouet, J. F. (1999). Content integration and source separation in learning from multiple texts. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.). *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 209–233). Mahwah, NJ: Erlbaum.

Britt, M. A., Richter, T., & Rouet, J.-F. (2014). Scientific Literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104–122.

Britt, M. A., & Rouet, J.- F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby, & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276–314). New York: Cambridge University Press.

Britt, M.A., Rouet, J. F., & Braasch, J. L. G. (2013). Documents experienced as entities: Extending the situation model theory of comprehension. In M.A. Britt, S.R. Goldman, & J.F. Rouet (Eds.), *Reading from words to multiple texts* (pp. 160–179). New York:

Routledge.

Bromme, R., & Goldman, S. (2014). The public's bounded understanding of science. *Educational Psychologist, 49*, 59–69.

Brooks, G. (2011). Adult literacy. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol IV, pp. 177–196). New York: Taylor & Francis.

Calkins, S., & Kelley, M. R. (2010). Evaluating internet and scholarly sources across the disciplines: Two case studies. *College Teaching, 55*, 151–156.

Chambliss, M. J. (1995). Text cues and strategies successful readers use to construct the gist of lengthy written arguments. *Reading Research Quarterly, 30*, 778–807.

Chambliss, M. J., & Murphy, P. K. (2002). Fourth and fifth graders representing the argument structure in written texts. *Discourse Processes, 34*, 91–115.

Chen, S., & Chaiken, S. (1999). The heuristic systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social and cognitive psychology* (pp. 73–96). New York: Guilford.

Chi, M. T. H., Bassok, M., Lewis, M.W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In: R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 17–76). Hillsdale, NJ: Erlbaum.

Chi M. T. H., de Leeuw, N., Chiu, M., La Vancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

Chinn, C., & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*, 1–49.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*, 175–218.

Chung, M., Oden, R. P., Joyner, B. L., Sims, A., & Moon, R. Y. (2012). Safe infant sleep recommendations on the Internet: Let's Google It. *The Journal of Pediatrics, 161*, 1080–1084.

Cobb, P. (1994). Constructivism in mathematics and science education. *Educational Researcher, 23*, 4.

Coté, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. van Oostendorp, & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 151– 174). Mahwah, NJ: Erlbaum.

Cronbach, L., & Snow, R. (1969). *Individual differences in learning ability as a function of instructional variables* (Final Report). Stanford, CA: School of Education, Stanford University.

Dauer, F. W. (1989). *Critical thinking: An introduction to reasoning*. New York: Oxford University Press.

De Pereyra, G., Britt, M. A., Braasch, J. L. G., & Rouet, J. F. (2014). Readers' memory for information sources in simple news stories: Effects of text and task features. *Journal of Cognitive Psychology*, *26*, 187–204.

De Vries, E., Lund, K., & Baker, M. (2002). Computer-mediated epistemic dialogue: Explanation and argumentation as vehicles for understanding scientific notions. *The Journal of the Learning Sciences, 11*, 63–103.

Deci E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*, 105–115.

Dewey, J. (1938). *Experience and education*. New York: Collier Books, MacMillan.

Dole, J. A., Duffy, G.G., Roehler, L. R., & Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research, 61*, 239–264.

Duke, N. K., & Carlisle, J. (2011). The development of comprehension. In M. L. Kamil et al. (Eds.), *Handbook of reading research* (Vol. IV, pp. 199–228). New York: Taylor & Francis.

Dunbar, K. (2000). How scientists think in the real world: Implications for science education. *Journal of Applied Developmental Psychology, 21*, 49–58.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*(1), 39–72.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth, TX: Harcourt, Brace, Jovanovich.

Ennis, R. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*, 45–48.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363–406.

Evans, J. S. B. T., & Thompson, V. A. (2004). Informal reasoning: Theory and method. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 58*, 69–74.

Exxon Mobile Corporation (2008). *2007 Worldwide contributions and community investments: Public information and policy research*. Retrieved February 8, 2018, from http://www.commoncause.org/issues/more-democracy-reforms/corporate-accountability/alec/whistleblower-complaint/supplemental-complaint-2016/Exhibit_68_2007-exxon-giving-report.pdf

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus*. Millbrae, CA: California Academic Press.

Fang, Z. (2008). Going beyond the fab five: Helping students cope with the unique linguistic challenges of expository reading in intermediate grades. *Journal of Adolescent & Adult Literacy, 51*, 476–487.

Feinstein, N. W. (2011). Salvaging science literacy. *Science Education, 95,* 168–185.

Feretti, T. R., Singer, M., & Patterson, C. (2008). Electrophysiological evidence for the time-course of verifying text ideas. *Cognition*, *108*, 881–888.

Ferguson, L. E. (2015). Epistemic beliefs and their relation to multiple-text comprehension: A Norwegian program of research. *Scandinavian Journal of Educational Research, 59*, 731–752.

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of Internet information credibility. *Journalism and Mass Communication Quarterly, 77*, 515–540.

Flanagin, A. J., & Metzger, M. J. (2001). Internet use in the contemporary media environment. *Human Communication Research, 27*, 153–181.

Flanagin, A. J., & Metzger, M. J. (2008). Digital media and youth: Unparalleled opportunity and unprecedented responsibility. In M. J. Metzger, & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 5–27). Cambridge, MA: The MIT Press.

Ford C. L., Yore L. D. (2012). Toward Convergence of Critical Thinking, Metacognition, and Reflection: Illustrations from Natural and Social Sciences, Teacher Education, and Classroom Practice. In A. Zohar, & Y. Dori (Eds.), *Metacognition in Science Education. Contemporary Trends and Issues in Science Education* (pp. 251–271). Dordrecht: Springer.

Galotti, K. M. (1989). Approaches to study formal and everyday reasoning. *Psychological Bulletin, 105*, 331–351.

Geddis, A. (1991). Improving the quality of classroom discourse on controversial issues. *Science Education, 75*, 169–183.

Gerjets, P., Kammerer, Y., & Werner, B. (2011). Measuring spontaneous and instructed

evaluation processes during web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. *Learning and Instruction, 21*, 220–231.

Gerrig, R. J., & McKoon, G. (1998). The readiness is all: The functionality of memory based text processing. *Discourse Processes, 26*, 67–86.

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. (2010). Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology, 35*, 157–173.

Glenberg A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*, 95–112.

Goldman, S. R., Britt, M. A., Brown, M., Greenleaf, C., & Lee, C. D. (2009). *Reading for understanding across grades 6 through 12: Evidence-based argumentation for disciplinary learning*. Funded July, 2010 by the Institute of Education Sciences, U.S. Department of Education, Grant # R305F100007.

Goldman, S. R., & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 417–436). Mahwah, NJ: Erlbaum.

Goldman, S., Braasch, J. L., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from Internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly, 47*, 356–381.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M. A., Greenleaf, C., Lee, C. D., Shanahan, C., & Project READI (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 0*, 1–28.

Goldman, S. R., Ozuru, Y., Braasch, J. L. G., Manning, F. H., Lawless, K. A., Gomez, K. W., & Slanovits, M. J. (2011). Literacies for learning: A multiple source comprehension

illustration. In N. L. Stein, & S. W. Raudenbush (Eds.), *Developmental cognitive science goes to school* (pp. 30–44). New York: Routledge.

Goldman, S., & van Oostendorp, H. (1999). Conclusions, conundrums, and challenges for the future. In H. van Oostendorp, & S. Goldman (Eds.), *The construction of mental representations during reading* (pp. 323–330). Mahwah, NJ: Erlbaum.

Gottlieb, E., & Wineburg, S. (2012). Between veritas and communitas: Epistemic switching in the reading of academic and sacred history. *The Journal of the Learning Sciences, 21*, 84–129.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Graesser, A. C., Wiley, J., Goldman, S. R., O'Reilly, T., Jeon, M., & McDaniels, B. (2007). SEEK Web Tutor: Fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition and Learning, 2*, 89–105.

Green, D. W. (1994). Induction: Representation, strategy and argument. *International Studies in the Philosophy of Science*, *8*, 45–50.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114*, 704–732.

Halpern, D. F. (1998). Teaching critical thinking across domains: dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449–455.

Hefter, M. H., Berthold, K., Renkl, A., Rieß, W., Schmid, S., & Fries, S. (2014). Effects of a training intervention to foster argumentation skills while processing conflicting scientific positions. *Instructional Science, 42*, 929–947.

Hefter, M. H., Renkl, A., Riess, W., Schmid, S., Fries, S., & Berthold, K. (2015). Effects of a training intervention to foster precursors of evaluativist epistemological understanding and intellectual values. *Learning and Instruction*, *39*, 11–22.

Hitchcock, D. (1995). Do fallacies have a place in the teaching of reasoning skills or critical

thinking? In H. V. Hansen, & R. Pinto (Eds.), *Fallacies: Classic and contemporary readings*. University Park: Penn State Press.

Hoeken, H., Timmers, R., & Schellens, P. J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking and Reasoning, 18*, 394–416.

Holtzman, N. A., Bernhardt, B. A., Mountcastle-Shah, E., Rodgers, J. E., Tambor, E., & Geller, G. (2005). The quality of media reports on discoveries related to human genetic diseases. *Public Health Genomics, 8*, 133–144.

Hornikx, J., & Hoeken, H. (2007). Cultural differences in the persuasiveness of evidence types and evidence quality. *Communication Monographs, 74*, 443–463.

Hornikx, J., & ter Haar, M. (2013). Evidence quality and persuasiveness: Germans are not sensitive to the quality of statistical evidence. *Journal of Cognition and Culture*, *13*, 483–501.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466.

Ikuenobe, P. (2004). On the theoretical unification and nature of the fallacies. *Argumentation*, *18*, 189–211.

Institute for the Future for University of Phoenix Research Institute (2011). *Future Work Skills 2020*. Retrieved February 2, 2018, from http://www.iftf.org/uploads/media/SR-1382A_UPRI_future_work_skills_sm.pdf

Isberner, M-B., & Richter, T. (2013). Can readers ignore implausibility? Evidence for nonstrategic monitoring of event-based plausibility in language comprehension. *Acta Psychologica, 142*, 15–22.

Isberner, M.-B., & Richter, T. (2014). Does validation during language comprehension depend on an evaluative mindset? *Discourse Processes, 51*, 7–25.

Jarman, R., & McClune, B. (2010). Developing students' ability to engage critically with science in the news: Identifying elements of the 'media awareness' dimension. *The*

Curriculum Journal, 21, 47–64.

Johns, C. L., Matsuki, K., & van Dyke, J. A. (2015). Poor readers' retrieval mechanism: Efficient access is not dependent on reading skill. *Frontiers in Psychology, 6*, 1–20.

Johnson, R., & Blair, A. (1977). *Logical self-defense*. Toronto: McGraw-Hill Ryerson.

Johnson, H. M., & Seifert, C. M. (1994). Sources of continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1420–1436.

Johnson, B. T., Smith-McLallen, A., Killeya, L. A., & Levin, K. D. (2004). Truth or consequences: Overcoming resistance with positive thinking. In E. S. Knowles & J. A. Linn (Eds.), *Resistance and persuasion* (pp. 215–233). Mahwah, NJ: Erlbaum.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*, 189–209.

Johnson-Laird, P. N., & Byrne, R. M. J. (1992). *Deduction*. Hillsdale, NJ: Erlbaum.

Jonassen, D. (1999). Designing constructivist learning environments. In C. M. Reigeluth (Ed.), *Instructional design theories and models* (pp. 215–239). Hillsdale, New Jersey: Erlbaum.

Kalyuga, S. (2008). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychological Review, 19*, 509–539.

Kamil, M. L., Pearson, P. D., Moje, E. B., & Afflerbach, P. P. (2011.). *Handbook of reading research* (Vol. IV). New York: Taylor & Francis.

Kammerer, Y., Kalbfell, E., & Gerjets, P. (2016). Is this information source commercially biased? How contradictions between web pages stimulate the consideration of source information. *Discourse Processes*, 53, 430–456.

Kelly, G. J., & Crawford, T. (1997). An ethnographic investigation of the discourse processes

of school science. *Science Education, 81*(5), 533–560.

Kelly, G.J., Druker, S., & Chen, C. (1998). Students' reasoning about electricity: combining performance assessments with argumentation analysis. *International Journal of Science Education, 20*, 849–871.

Kendeou, P., Smith, E. R., & O'Brien, E. J. (2013). Updating during reading comprehension: Why causality matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 854–865.

Kim, S. M., & Hannafin, M. J. (2016). Synergies: effects of source representation and goal instructions on evidence quality, reasoning, and conceptual integration during argumentation-driven inquiry. *Instructional Science, 44*, 441–476.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgement: Understanding and promoting intellectual growth and critical thinking in adolescents and adults*. San Francisco, CA: Jossey-Bass.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163–182.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Kintsch, W., & Welsch, D. M. (1991). The construction– integration model: A framework for studying memory for text. In W. E. Hockley, & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennett B. Murdock* (pp. 367–385). Hillsdale, NJ.: Erlbaum.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29*, 133–159.

Korpan, C. A., Bisanz, G. L., Bisanz, J., & Henderson, J. M. (1997). Assessing literacy in science: Evaluation of scientific news briefs. *Science Education, 81*, 515–532.

Krishnamurti, J. (1969). *Freedom from the known*. Bramdean, Hampshire: Random House.

Kuhn, D. (1991). *The skills of argument*. Cambridge, UK: Cambridge University Press.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74,* 1245–1260.

Larson, A. A., Britt, M. A., & Kurby, C. (2009). Improving students' evaluation of informal arguments. *Journal of Experimental Education, 77*, 339–365.

Larson, M., Britt, M. A., & Larson, A. (2004). Disfluencies in comprehending argumentative texts. *Reading Psychology, 25*, 205–224.

Le Bigot, L., & Rouet, J.-F. (2007). The impact of presentation format, task assignment, and prior knowledge on students' comprehension of multiple online documents. *Journal of Literacy Research, 39*, 445–470.

Lehman, D. R., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology, 26,* 431–442.

Leinhardt, G., & Young, K. M. (1996). Two texts, three readers: Distance and expertise in reading history. *Cognition and Instruction*, *14*, 441–486.

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology, 94*, 778–784.

Lipman, M. (1988). Critical thinking - what can it be? *Educational Leadership, 46*, 38–43.

Luke, C., de Castell, S. C., & Luke, A. (1989). Beyond criticism: the authority of the school textbook. In S. C. de Castell, A. Luke, & C. Luke (Eds.), *Language, authority, and criticism* (pp. 245–260). London: Falmer.

Lundeberg, M. A. (1987). Metacognitive aspects of reading comprehension: studying understanding in legal case analysis. *Reading Research Quarterly, 22*, 407–432.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition & Instruction, 3*, 251–283.

Maier, J., & Richter, T. (2013a). How nonexperts understand conflicting information on social science issues: The role of perceived plausibility and reading goals. *Journal of Media Psychology, 25*, 14–26.

Maier, J., & Richter, T. (2013b). Text-belief consistency effects in the comprehension of multiple texts with conflicting information. *Cognition and Instruction, 31*, 151–175.

Maier, J., & Richter, T. (2016). Effects of text-belief consistency and reading task on the strategic validation of multiple texts. *European Journal of the Psychology of Education, 31*, 479–497.

Majetic, C., & Pellegrino, C. (2014). When science and information literacy meet: An approach to exploring the sources of science news with non-science majors. *College Teaching, 62*, 107–112.

Manuel, K. (2002). How first-year college students read popular science: An experiment in teaching media literacy skills. *SIMILE: Studies in Media & Information Literacy Education, 2*, 1–12.

Marsh E. J., Edelman G., & Bower G. H. (2001). Demonstrations of a generation effect in context memory. *Psychonomic Society, 29*, 798–805.

Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and Instruction, 16*, 492–509.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology, 37*, 913–934.

Mayer, R. (1989). Models for understanding. *Review of Educational Research, 59*, 43–64.

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review, 19*, 113–139.

McCrudden, M. T., & Sparks, P. C. (2014). Exploring the effect of task instructions on topic beliefs and topic belief justifications: A mixed methods study. *Contemporary Educational Psychology, 39*, 1–11.

McCrudden, M. T., Stenseth, T., Bråten, I., & Strømsø, H.I. (2016). The effects of author expertise and content relevance on document selection: A mixed methods study. *Journal of Educational Psychology, 108*, 147–162.

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction, 14,* 139–178.

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003). College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education, 41*, 271–290.

Mountcastke-Shah, E., Tambor, E., Bernhardt, B. A., Geller, G., Karaliukas, R., Rodgers, J. E., & Holzman, N. A. (2003). Assessing mass media reporting of disease-related genetic discoveries: Development of an instrument and initial findings. *Science Communication, 24*, 458–478.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*, 319–334.

Myers, G. (1991). Lexical cohesion and specialized knowledge in science and popular science texts. *Discourse Processes, 14*, 1–26.

National Assessment of Educational Progress (1996). *NAEP 1994 U.S. History Report Card: Findings from the National Assessment of Educational Progress*. Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress (1998). *NAEP 1998 Writing Report Card: Findings from the National Assessment of Educational Progress*. Princeton, NJ: Educational Testing Service.

National Center for Education Statistics. (2010). *The nation's report card: Grade 12 reading and mathematics 2009 national and pilot state results* (pp. 2011–455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Nokes, J., Dole, J., Hacker, D. J. (2007). Teaching high school students to be critical and strategic readers of historical texts. *Journal of Educational Psychology, 99*, 492–504.

Noordman, L. G. M., Vonk, W., & Kempff, H. J. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language, 13*, 573–590.

Norris, S. P., & Phillips, L. M. (1994). Interpreting pragmatic meaning when reading popular reports of science. *Journal of Research in Science Teaching, 31*, 947–967.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*, 224–240.

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science, 12*, 123–145.

Nwogu (1991). Structure of science populizations: A genre-analysis approach to the schema of popularized medical texts. *English for Specific Purpose, 10*, 111–123.

O'Brien, E. J., Lorch, R. F., & Myers, J. L. (1998). Memory-based text processing. *Discourse Processes*, *26,* 2–3.

O'Brien, E. J., & Myers, J. L. (1987). The role of causal connections in the retrieval of text. *Memory & Cognition, 15,* 419–427.

OECD (2011). *PISA 2009 Results: Students on Line: Digital Technologies and Performance (Volume VI)*. Paris: PISA, OECD Publishing. Retrieved February 5, 2018 from http://dx.doi.org/10.1787/9789264112995-en.

OECD (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014)*. Paris:

PISA, OECD Publishing. Retrieved February 5, 2018 from http://dx.doi.org/10.1787/9789264208780-en.

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: PISA, OECD Publishing. Retrieved February 5, 2018 from http://dx.doi.org/10.1787/9789264266490-en

Osborne, J. F., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*, 994–1020.

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction?. *Science Education, 95,* 627–638.

Park, H. S., Levine, T. R., Kingsley Westerman, C. Y., Orfgen, T., & Foregger, S. (2007). The effect of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social judgement predictions. *Human Communication Research, 33*, 81–102.

Paxton, R. J. (1997). "Someone with like a life wrote it": The effects of a visible author on high school history students. *Journal of Educational Psychology, 89*, 235–250.

Pearson, P. D., Roehler, L. R., Dole, J. A. & Duffy, G. G. (1992). Developing expertise in reading comprehension. In S.J. Samuels & A.E. Farstrup (Eds.), *What research has to say about reading instruction* (pp. 145–199). Newark, DE: International Reading Association.

Penney, K., Norris, S. P., Phillips, L. M., & Clark, G. (2003). The anatomy of junior high school science textbooks: An analysis of textual characteristics and a comparison to media reports of science. *Canadian Journal of Science, Mathematics and Technology Education, 3*, 415–436.

Perfetti, C. A. (1985). *Reading Ability*. New York, NY: Oxford University Press.

Perfetti, C. A., Britt, M. A., & Georgi, M. C. (1995). *Text-based learning and reasoning: Studies in history*. Hillsdale, NJ: Erlbaum.

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of documents representa-

tion. In H. van Oostendorp, & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mawah, NJ: Erlbaum.

Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology, 77*, 562–571.

Perkins, D. N. (1986). *Knowledge as design*. Hillsdale, NJ: Erlbaum.

Perkins, D. N., Farady, M., Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In Voss, J. F., Perkins, DN., Segal, J. W., (Eds*.). Informal reasoning and education* (pp. 83–105). Hillsdale, NJ: Erlbaum.

Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength: Creating attitudes that are persisent, resistant, and predictive of behavior. In R. E. Petty, & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 93–130). Mawah, NJ: Erlbaum.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer.

Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken, & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 41–72). New York: Guilford Press.

Phillips, L. M., & Norris, S. P. (1999). Interpreting popular reports of science: What happens when the reader's world meets the world on paper? *International Journal of Science Education, 21*, 317–327.

Phillips, L. M., & Norris, S. P. (2009). Bridging the gap between the language of science and the language of school science through the use of adapted primary literature. *Research in Science Education, 39*, 313–319.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of Reading Research* (Vol. III, pp. 545–562). Mahwah, NJ: Erlbaum.

Prüfer, P., & Rexroth, M. (2000). *Zwei – Phasen Pretesting [Two phase pretesting]*. ZUMA-Arbeitsbericht 2000/08. Mannheim: ZUMA.

Ramasamy, S. (2011). *An analysis of informal reasoning fallacy and critical thinking dispositions among Malaysian undergraduates*. Retrieved February 24, 2018 from https://files.eric.ed.gov/fulltext/ED525513.pdf

Ransohoff, D. F., & Ransohoff, R. M. (2000). Sensationalism in the media: When scientists and journalists may be complicit collaborators. *Effective Clinical Practice, 4*, 185–188.

Renkl, A. (2009). Wissenserwerb [Knowledge acquisition]. In E. Wild, & J. Möller (Eds.). *Pädagogische Psychologie* (pp. 3–26). Berlin: Springer.

Ricco, R. B. (2007). Individual differences in the analysis of informal reasoning fallacies. *Contemporary Educational Psychology, 32,* 459–484.

Richter, T. (2003). *Epistemologische Einschätzungen beim Textverstehen* [Epistemic validation in text comprehension]. Lengerich, Germany: Pabst Science Publishers.

Richter, T. (2011). Cognitive flexibility and epistemic validation in learning from multiple texts. In J. Elen, E. Stahl, R. Bromme, & G. Clarebout (Eds.), *Links between beliefs and cognitive flexibility.* Berlin: Springer.

Richter, T. (2015). Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes, 52*, 337–352.

Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A Two-step Model of Validation. *Educational Psychologist, 52*, 148–166.

Richter, T., & Maier, J. (in press). The role of validation in multiple source use. In J. Braasch, I. Bråten, & M. McCrudden (Eds.), *Handbook of multiple source use*. Routledge.

Richter, T.*, & Schmid, S.* (2010). Epistemological beliefs and epistemic strategies in self-regulated learning. *Metacognition and Learning, 5*, 47–65 (Authors contributed equally to this article).

Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology, 96*, 538–558.

Richter, T., & Singer, M. (2017). Discourse updating: Acquiring and revising knowledge through discourse. In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *The Routledge handbook of discourse processes* (Vol. II, pp. 167–190*)*. New York: Routledge.

Richter, T., & van Holt, N. (2005). ELVES: Ein computergestütztes Diagnostikum zur Erfassung der Effizienz von Teilprozessen des Leseverstehens. *Diagnostica, 51*, 169– 182.

Rottenberg, A. (1988). *Elements of Argument: A Text and Reader*. New York: St. Martin's.

Rouet, J. -F. (2006). *The skills of document use: From text comprehension to Web-based learning*. Mahwah, NJ: Erlbaum.

Rouet, J. -F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Greenwich, CT: Information Age Publishing.

Rouet, J. -F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88*, 478–493.

Rouet, J. -F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction, 15*, 85–106.

Rouet, J. -F, Le Bigot, L., de Pereyra, G., & Britt, M.A. (in press). Whose story is this? Discrepancy triggers readers' attention to source information in short narratives. *Reading and Writing*.

Sanchez, C. A., Wiley, J., & Goldman, S. R. (2006). Teaching students to evaluate source reliability during internet research tasks. In *Proceedings of the 7th international conference of the learning sciences* (pp. 662–666). Bloomington, IN: ACM Digital Library.

Scharrer L., & Salmerón, L. (2016). Sourcing in the reading process. Introduction to the special issue. *Reading and Writing, 29*, 1539–1548.

Schroeder, S., Richter, T., & Hoever, I. (2008). Getting a picture that is both accurate and stable: Situation models and epistemic validation. *Journal of Memory and Language, 59*, 237–255.

Secko., D. M., Amend, E., & Friday, T. (2013). Four models of science journalism: A synthesis and practical assessment. *Journalism Practice, 7*, 62–80.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review*, *78*(1), 40–59.

Shanahan, C., Shanahan, T., & Misischia, C. (2011). Analysis of expert readers in three disciplines: History, mathematics, and chemistry. *Journal of Literacy Research, 43*, 393–429.

Shaw, F. W. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning, 2*, 51–80.

Siegel, H., & Biro, J. (1997). Epistemic normativity, argumentation, and fallacies. *Argumentation, 11, 277–*292.

Singer, M. (2006). Verification of text ideas during reading. *Journal of Memory and Language, 54*, 574–591.

Singer, M. (2013). Validation in reading comprehension. *Current Directions in Psychological Science, 22*, 361–366.

Slob W. H. (2002). What is Wrong with Fallacies?. In W. H. Slob (Ed.), *Dialogical Rhetoric: An essay on truth and normativity after postmodernism* (pp. 101–134). Dordrecht: Springer.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework of research in individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 13–59). New York, NY: Freeman.

Spencer, R. W. (2016). *A guide to understanding global temperature data*. Austin, Texas: Texas Public Policy Foundation. Retrieved February 5, 2018 from https://www.texaspolicy.com/library/doclib/FFP-Global-Temperature-booklet-July-2016-PDF.pdf

Spiro, R. J., & Jehng, J. (1990). Cognitive flexibility and hypertext: Theory and technology for the nonlinear and multidimensional traversal of complex subject matter. In D. Nix & R. Spiro (Eds.), *Cognition, education, and multimedia: Exploring ideas in high technology* (pp. 163–205). Hillsdale, NJ: Erlbaum.

Stadtler, M., & Bromme, R. (2007). Dealing with multiple documents on the WWW: The role of metacognition in the formation of documents models. *International Journal of Computer Supported Collaborative Learning, 2*, 191–210.

Stadtler, M., & Bromme, R. (2008). Effects of the metacognitive computer-tool met.a.ware on the web search of laypersons. *Computers in Human Behavior, 24*, 716–737.

Stadtler, M., Scharrer, L., Skodzik, T., & Bromme, R. (2014). Comprehending multiple documents on scientific controversies: Effects of reading goals and signaling rhetorical relationships. *Discourse Processes, 51*, 93–116.

Stahl, S.A., Hynd, C.R., Britton, B.K., Mc Nish, M.M., & Bosquet, D. (1996). What happens when students read multiple source documents in history? *Reading Research Quarterly, 31*, 430–456.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 343–365). New York: Oxford University Press.

Stanovich, K. E., & West, R. (1997). Reasoning independently of prior belief and individual

differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342–357.

Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., & Majewski, H. (2007). The time course of plausibility effects on eye movements in reading: Evidence from noun-noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 1162–1169.

Stanovich, K.E., & West, R.F. (1998). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342–357.

Strømsø, H. I., & Bråten, I. (2014) Students' sourcing while reading and writing from multiple web documents. *Nordic Journal of Digital Literacy, 9*, 92–111.

Strømsø, H. I., Bråten, I., & Britt, M. A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning and Instruction, 20*, 192–204.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Suppe, F. (1998). The structure of a scientific paper. *Philosophy of Science, 65*(3), 381–405.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. New York: Cambridge University Press.

Tabak, I. (2016). Functional scientific literacy: Seeing the science within the within the words and across the web. In L. Corno & E. A. Anderman (Eds), *Handbook of educational psychology* (pp. 269–280). New York: Routledge.

Tenopir, C., & King, D. W. (2004). *Communication patterns of engineers*. Hoboken, NY: Wiley.

Tittle, P. (2011). *Critical thinking: An appeal to reason*. New York: Routledge.

Thomas, A. (1991). *Grundriss Sozialpsychologie, Bn. 1 [Outline of social psychology, Vol. 1]*. Göttingen, Germany: Hogrefe.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, MA: Cambridge University Press.

Van Eemeren, F. H., Garssen B., & Meuffels, B. (2009). *Fallacies and judgments of reasonableness: Empirical research concerning the pragma-dialectical discussion rules*. Dordrecht: Springer.

Van Eemeren, F. H., Grootendorst, R., Henkemans,, F. S., Blair, J. A., Johnson, R. H., Krabbe, E. C. W., Plantin, C., Walton, D. N. , Willard, C. A. , Woods, J., & Zarefsky, D. (1996). *Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments*. Mahwah, NJ: Erlbaum.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

van Gelder, T., Bissett, M., & Cumming, G. (2004). Cultivating expertise in informal reasoning. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 58*, 142–152.

Van Strien, J. L. H., Brand-Gruwel, S., & Boshuizen, H. P. A. (2014). Dealing with conflicting information from multiple nonlinear texts: Effects of prior attitudes. *Computers in Human Behavior, 32*, 101–111.

Voss, J. F. (2005). Toulmin's model and the solving of ill-structured problems. *Argumentation, 19*, 321–329.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1989). Informal reasoning and subject matter knowledge in the solving of economics problems by naive and novice individuals. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 217–249). Hillsdale, NJ: Lawrence Erlbaum Associates.

Voss, J. F., Fincher-Kiefer, R., Wiley, J., & Silfies, L. N. (1993). On the processing of arguments. *Argumentation, 7*, 165–181.

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*, 337–350.

Voss, J. F., Perkins, D. N., & Segal, J. W. (1991). *Informal reasoning and education*. Hillsdale, NJ: Erlbaum.

Walton, D. (1995). *A Pragmatic Theory of Fallacy*. Tuscaloosa: The University of Alabama Press.

Weinstock, M. (2009). Relative expertise in an everyday reasoning task: Epistemic understanding, problem representation, and reasoning competence. *Learning and Individual Differences, 19*, 423–434.

Weinstock, M. P., Neumann, Y., & Glassner, A. (2006). Identification of informal reasoning fallacies as a function of epistemological level, grade level, and cognitive ability. *Journal of Educational Psychology, 89*, 327–341.

Weinstock, M., Neuman, Y., & Tabak, I. (2004). Missing the point or missing the norms? Epistemological norms as predictors of students' ability to identify fallacious arguments. *Contemporary Educational Psychology, 29*, 77–94.

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*, 930–941.

Wigfield, A., Gladstone, J. R., & Turci, L. (2016). Beyond cognition: Reading motivation and reading comprehension. *Child Development Perspectives, 10*, 190–195.

Wild, K. P. (2000). *Lernstrategien im Studium: Strukturen und Bedingungen [Learning strategies at university: structures and conditions]*. Münster: Waxmann.

Wiley, J. (2005). A fair and balanced look at the news: What affects memory for controversial arguments? *Journal of Memory and Language, 53*, 95–109.

Wiley, J., Goldman, S. R., Graesser, A.C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal, 46*, 1060–1106.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that

promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301–311.

Wineburg, S. (1991). Historical problem solving: a study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*, 73–87.

Winter, S., & Krämer, N. C. (2012). Selecting science information in web 2.0: How source cues, message sidedness, and need for cognition influence users' exposure to blog posts. *Journal of Computer-Mediated Communication, 18*, 80–96.

Wolfe, C. R. (2012). Individual differences in the "myside bias" in reasoning and written argumentation. *Written Communication, 29*, 477–501.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the Myside Bias in written argumentation. *Written Communication, 26*, 183–209.

Wolfe, M. B., & Kurby, C. A. (2017): Belief in the claim of an argument increases perceived argument soundness. *Discourse Processes, 54*, 599–617.

Wopereis, I., Brand-Gruwel, S., & Vermetten, Y. (2008). The effect of embedded instruction on solving information problems. *Computers in Human Behavior, 24*, 738–752.

Wu, Y. -T., & Tsai, C. -C. (2007). High school students' informal reasoning on a socio-scientific issue: Qualitative and quantitative analyses. *International Journal of Science Education, 29,* 1163–1187.

Wyatt, D., Pressley, M., El-Dinary, P. B., Stein, S., Evans, P., & Brown, R. (1993). Comprehension strategies, worth and credibility monitoring, and evaluations: Cold and hot cognition when scientists read professional articles that are important to them. *Learning and Individual Differences, 5*, 49–72.

Yore, L. D., Bisanz, G. L., & Hand, B. M. (2003), Examining the literacy component of scientific literacy: 25 years of language arts and science research, *International Journal of Science Education, 25*, 689–725.

Yore, L. D., Craig, M. T., Maguire, T. O. (1998). Index of science reading awareness: An interactive-constructive model, test verification, and grades 4–8 results. *Journal of Research in Science Teaching, 35*, 27–51.

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science and Mathematics Education, 5*, 559–589.

Zimmerman, C., Bisanz, G. L., Bisanz, J., Klein, J. S., & Klein, P. (2001). Science at the supermarket: A comparison of what appears in the popular press, expert's advice to readers, and what students want to know. *Public Understanding of Science, 10*, 37–58.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology, 20*, 920–933.

Appendix

Table of Contents

Appendix A

Stimulus Materials for Study 1

Test Items– Plausibility Task

No.	Version	Test Item	Plausibility	Argumentation Fallacy
1	1	Die Entwicklung zum Raucher und später ggf. auch zum Nicht-Mehr-Raucher basiert auf dem Zusammenwirken einer Vielzahl sozialer, psychologischer und biologischer Faktoren.	Plausible	
2	1	Eine zentrale Rolle scheint dabei das Konstrukt der ererbten Nikotinsensitivität zu spielen. Dieses Konstrukt bezieht sich auf die Tatsache, dass manche Menschen sensitiver auf Nikotin reagieren, weil sie sensibler auf Nikotin ansprechen.	Implausible	Circular Reasoning
3	1	Mit ihm soll erklärt werden, warum manche Menschen – obwohl sie schon eine relativ große Zahl von Zigaretten geraucht haben – nicht vom Nikotin abhängig werden und Gelegenheitsraucher bleiben, während andere eine hochgradige Nikotinabhängigkeit entwickeln.	Plausible	
4	1	Die Frage nach der genetischen Prädisponiertheit für das Rauchen bezieht sich aber nicht nur auf differentielle Unterschiede bei der Nikotinsensitivität, sondern darüber hinaus auch auf angeborene Persönlichkeitsunterschiede.	Implausible	False Dichotomy
5	1	Auf der Grundlage der Eysenckschen Drei-Faktoren-Theorie wurde insbesondere ein positiver Zusammenhang zwischen dem Rauchen und dem Persönlichkeitsmerkmal „Extraversion“ postuliert und auch in einer Vielzahl von Studien empirisch nachgewiesen (z.B. Lipkus, Barefoot, Williams & Siegler, 1994).	Plausible	
6	1	Gleiches gilt auch für den Zusammenhang zwischen Rauchen und dem von Zuckerman postulierten Persönlichkeitsmerkmal „Sensation	Implausible	Wrong Example

		Seeking“, der Suche nach immer neuen Reizen und Stimulationen. Menschen, die eher zurückhaltend wirken und ihre Aufmerksamkeit mehr auf ihr Innenleben richten und solche, die wenig Bedürfnis nach verschiedenen, neuen, intensiven und risikoreichen Sinneserfahrungen verspüren, haben somit ein erhöhtes Risiko, mit dem Rauchen anzufangen.		
7	1	Ob es in der biologischen Grundausstattung des Menschen wirklich Unterschiede innerhalb eines Individuums gibt, die mit der Wahrscheinlichkeit des Rauchens, seiner Initiierung, seiner Aufrechterhaltung und seines Beendens kovariieren, ist unklar.	Plausible	
8	1	Ist das Rauchverhalten erst einmal zur Gewohnheit geworden, ist es für die meisten jedenfalls schwer, wieder davon loszukommen, da es Unterschiede in der Nikotinsensitivität gibt.	Implausible	False Conclusion
9	1	Unklar ist, in welcher Weise solche differentiellen Faktoren Einfluss auf jene aktuell ablaufenden kognitiven und emotionalen Prozesse haben, die für die Herausbildung, die Beibehaltung oder den Abbruch des Rauchverhaltens in einer spezifischen Lebenslage unmittelbar verantwortlich sind.	Plausible	
10	1	Während in den Aneignungsphasen des Rauchverhaltens vor allem biopsychologische Einflussgrößen eine wichtige Rolle spielen, so scheint im Stadium der Aufrechterhaltung das Rauchverhalten in erster Linie eine Funktion interpersonaler Faktoren zu sein.	Plausible	
11	1	In einem Experiment wurden Gewohnheitsraucher über mehrere Wochen hinweg mit Zigaretten versorgt, die entweder stark oder schwach nikotinhaltig waren, ohne dass dies für die Probanden erkennbar war. Die Probanden rauchten im Durchschnitt 25% mehr von den leichten Zigaretten als von den starken. Dieses Ergebnis ist ein eindeutiger	Implausible	Overgeneralisation

		Beweis dafür, dass es Unterschiede in der Nikotinsensitivität gibt.		
1	2	Wenn Kinder bzw. Jugendliche anfangen, darüber nachzudenken, einmal eine Zigarette zu probieren, werden erstmals auf das Rauchen bezogene Vorstellungen und Erwartungen herausgebildet. Man bezeichnet dieses Stadium auch als die Phase der Vorbereitung.	Plausible	
2	2	So führt z.B. die Erwartung, dass einem das Rauchen Stresssituationen erleichtert, eher dazu, dass man in solchen Situationen raucht, als die Erwartung, dass das Rauchen in Belastungssituationen hilfreich ist.	Implausible	False Dichotomy
3	2	Da eigene Erfahrungen mit Zigaretten noch nicht vorliegen, basiert die Bildung solcher Erwartungen hauptsächlich auf der Beobachtung des Modellverhaltens der relevanten Bezugspersonen.	Plausible	
4	2	Lässt der Vater z.B. nach dem Essen erkennen, wie herrlich ihm jetzt die Verdauungszigarette schmeckt, wird bei den anwesenden Kindern eher eine negative Vorstellung des Zigarettenrauchens bekräftigt.	Implausible	Wrong Example
5	2	In einer Längsschnittuntersuchung von Dinh, Sarason, Peterson und Onstad (1995) ist gezeigt worden, dass solche positiven Vorstellungen von den Wirkungen des Rauchens bei Fünftklässlern ein signifikanter Prädiktor für das Rauchverhalten vier Jahre später sind und dass diese positiven Vorstellungen bei Fünftklässlern das künftige Rauchverhalten stärker beeinflussen als entsprechende Vorstellungen bei Siebtklässlern.	Plausible	
6	2	Die kognitive Vorbereitung auf das spätere Rauchen scheint bereits in der Grundschule zu beginnen und einen nachhaltigen Einfluss auf die weitere Entwicklung des Rauchverhaltens zu besitzen, weil sie bereits in der Kindheit anfängt und einen anhaltenden Effekt hat.	Implausible	Circular Reasoning

7	2	Mit dem Rauchen der ersten Zigarette gelangen Jugendliche in eine Experimentierphase. Da etwa 80% bis 90% aller Jugendlichen wenigstens einmal eine Zigarette rauchen, trägt dieses Experimentieren nicht den Charakter eines abweichenden Verhaltens, sondern eher den einer normativen Entwicklungsaufgabe.	Plausible	
8	2	Kritisch für die Herausbildung des gewohnheitsmäßigen Rauchens ist nicht der Umstand, dass eine erste Zigarette geraucht wird, sondern die Art und Weise, wie anschließend das Erlebnis dieser ersten Zigarette kognitiv und emotional verarbeitet wird.	Plausible	
9	2	Aus der Tatsache, dass am Ende der Jugendzeit der Anteil der gelegentlichen und regelmäßigen Raucher zusammengenommen auf etwa 50% absinkt, lässt sich ableiten, dass das Interesse am Rauchen im Laufe der Jugendzeit sogar noch zunimmt.	Implausible	False Conclusion
10	2	Eine Studie fand, dass Jugendliche eines Jungeninternats vor allem dann mit dem Rauchen anfingen, wenn sie keinen Anschluss an eine Clique fanden oder von einer solchen ausgeschlossen wurden. Der Einfluss der Peergruppe ist daher ein sehr wichtiger Faktor, um mit dem Rauchen anzufangen.	Implausible	Overgeneralisation
11	2	Natürlich gibt es noch viele andere Faktoren, die bei der Entwicklung zum Raucher und der Aufrechterhaltung des Rauchens eine Rolle spielen. Eine umfassende Theorie fehlt bislang.	Plausible	

Test Items– Argument Structure Task

No.	Version	Test Item	Words
1	1	In einer Längsschnittstudie mit 653 Kindern wurde untersucht, wie sich die Bereitschaft, eine Belohnung aufzuschieben, auf die Entwicklung von Schülern auswirkt. Die Autoren stellten fest, dass Kinder, die im Alter von vier oder fünf Jahren eine Belohnung (beispielsweise einen Keks) aufschoben, wenn eine weitere Belohnung (zwei Kekse) lockte, zehn Jahre später bessere kognitive und soziale Kompetenzen aufwiesen als Kinder, die eine sofortige Belohnung vorzogen [*datum*]. Schulischer Erfolg spielt für den weiteren beruflichen Erfolg eine zentrale Rolle [*warrant*]. Abiturienten mit sehr guten Abschlussnoten haben z.B. im Studium oft geringere Schwierigkeiten und Belastungen sowie einen stabileren Studienverlauf [*backing*]. Selbstkontrolle sollte deshalb so früh wie möglich trainiert werden [*claim*]. Natürlich gibt es neben erfolgreicher Selbstkontrolle noch viele andere Faktoren, die für die schulische Entwicklung eines Kindes verantwortlich sind [*rebuttal*].	141
2	2	Entwicklungspsychologische Befunde zeigen deutlich, dass Heranwachsende in vielen psychischen Funktionsbereichen noch wie Jugendliche organisiert sind. Zum Beispiel konnte in einer Studie gezeigt werden, dass Impulsivität, wie auch das „sensation-seeking“, d.h. das Bedürfnis nach Reizstimulation, zwischen mittlerer Adoleszenz und frühem Erwachsenenalter ansteigt und erst danach wieder abfällt [*datum*]. Impulsivität und Sensation-seeking stehen beide im Zusammenhang mit höherem Risikoverhalten im Straßenverkehr [*warrant*]. So verursachen z.B. Menschen, bei denen diese Merkmale besonders ausgeprägt sind, mehr Unfälle [*backing*]. Heranwachsende sollten deshalb bei Vergehen im Straßenverkehr noch dem Jugendstrafrecht unterworfen werden [*claim*]. Auch wenn im Einzelfall gegebenenfalls anders entschieden werden muss [*rebuttal*].	117

Scoring System – Plausibility Task

Main Category	Clarification	Examples
Intuitive	Judgements Based on the Reader's Intuition	"Somehow this does not seem plausible to me."
		"I'm not sure, is this plausible? Maybe not, the wording sounds strange. I'll press "implausible" here."
		"That sounds okay, plausible."
Internal Consistency	Judgements Based on the Strength of Stated Evidence (i.e. Relevance, Sufficiency)	"This makes sense. The claim is supported with a reason."
		"You can't say that this conclusion follows from these findings. The author is too certain about her results."
		"This is a bad argument. The reason stated here is not informative, because it only repeats what was already said before."
Knowledge / Opinion	Judgements Based on the Reader's Prior Knowledge or Beliefs About Only One Component of the Argument (i.e. Accuracy)	"I think the author is wrong! From my experience, everyone gets addicted to nicotine at some point."
		"It starts early in childhood... Seems plausible from what I know about this."
		"Yes, that's true. We discussed that in class."

Other	Embedded References	"There is a reference. That's good."
	Global Text Coherence	"Earlier in the text they said that the picture is not entirely clear. But now they are saying that it's difficult for many people to stay off cigarettes once they started."
	Perceived (In)Completeness	"I think this needs to be described in more detail."

Retrospective Interview – Plausibility Task

No.	Question	Probing Technique
1	War Ihnen die Aufgabenstellung deutlich?	General Probing
1a	Falls nein, was war unklar? [Vgl. Frage 1]	Category Selection Probing
2	Was verstehen Sie unter einem plausiblen Argument? Erklären Sie bitte in Ihren eigenen Worten.	Special Comprehension Probing
3	Wie sind Sie bei der Beurteilung der Plausibilität der verschiedenen Textabschnitte vorgegangen?	Information Retrieval Probing
4	Wie sind Sie bei der Zuordnung der Textabschnitte, die Sie als unplausibel deklariert hatten, zu den entsprechenden Argumentationsfehlern vorgegangen?	Information Retrieval Probing
5	Auf einer Skala von 1 = *sehr einfach* bis 6 = *sehr schwierig*, wie schwierig fanden Sie es, die Plausibilität zu beurteilen?	General Probing
5a	Warum fanden Sie es sehr einfach/ einfach/ eher einfach/ eher schwierig/ schwierig/ sehr schwierig? [Vgl. Frage 5]	Category Selection Probing
6	Auf einer Skala von 1 = *sehr einfach* bis 6 = *sehr schwierig*, wie schwierig fanden Sie es, die Textabschnitte, die Sie als unplausibel deklariert hatten, den entsprechenden Argumentationsfehlern zuzuordnen?	General Probing
6a	Warum fanden Sie es sehr einfach/ einfach/ eher einfach/ eher schwierig/ schwierig/ sehr schwierig? [Vgl. Frage 6]	Category Selection Probing
7	Dachten Sie bei der Zuordnung der Argumentationsfehler noch an andere Fehler als die von uns genannten?	Information Retrieval Probing
7a	Falls ja, welche? [Vgl. Frage 7]	Category Selection Probing

8 Haben Sie sonstige Anmerkungen zu dieser Aufgabe? General Probing

Appendix B

Stimulus Materials for Study 2

Test Items – Credibility Task

No.	Version	Document Type	Document Style	Peer-Review	Author(s)	Structure	Content	Presentation of Science	Rank Order Credibility
1	1	Original Empirical Article	Argumentative	Yes	Scientists	Canonical	Arguments	Uncertain	1
2	1	Review Article	Argumentative	Yes	Scientist	Non-canonical	Arguments	Uncertain	2
3	1	Edited Book	Neutral – Objective	No	Scientists	Non-canonical	Facts	Certain	3
4	1	Textbook	Neutral – Objective	No	Science Educator	Non-canonical	Facts	Certain	4
5	1	Monography	Neutral – Objective	No	Scientist	Non-canonical	Facts	Certain	4
6	1	Popular Science Book	Neutral – Subjective	No	Science Journalist	Non-canonical	Facts with minimum evidence	Certain	5
7	1	Popular Science Article	Neutral – Subjective	No	Science Journalist	Non-canonical	Facts with minimum evidence	Certain	6
1	2	Original Empirical Article	Argumentative	Yes	Scientists	Canonical	Arguments	Uncertain	1

2	2	Review Article	Argumentative	Yes	Scientists	Non-canonical	Arguments	Uncertain	2
3	2	Edited Book	Neutral – Objective	No	Scientist	Non-canonical	Facts	Certain	3
4	2	Textbook	Neutral – Objective	No	Science Educator	Non-canonical	Facts	Certain	4
5	2	Monography	Neutral – Objective	No	Scientist	Non-canonical	Facts	Certain	4
6	2	Popular Science Book	Neutral – Subjective	No	Science Journalist	Non-canonical	Facts with minimum evidence	Certain	5
7	2	Popular Science Article	Neutral – Subjective	No	Science Journalist	Non-canonical	Facts with minimum evidence	Certain	6

Text 1, Version 1	
Document Type	Original Empirical Article
Title	Qualität der Partnerschaft: Ein Produkt von Wertkonsens und Beziehungsdauer
Book / Journal	Zeitschrift für Sozialpsychologie, 34 (2), 91-106
Publisher	Hogrefe
Date of Publication	2003
Author(s)	Brandstätter, H., & Cronberger, N.
Institution	University of Linz
Place	Göttingen
Abstract	Unverheiratete, kinderlose Paare (*n* = 67) in einer mindestens sechs Monate und höchstens zehn Jahre bestehenden Partnerschaft (Alter zwischen 16 und 36 Jahren) beurteilten die emotionale Qualität ihrer Beziehung und beschrieben ihre und ihres Partners Werthaltungen. In einer moderierten Regressionsanalyse mit Ähnlichkeit der Werthaltungen, Beziehungsdauer und Produkt aus Ähnlichkeit und Beziehungsdauer als unabhängigen Variablen und Beziehungsqualität als abhängige Variable stellte sich in Übereinstimmung mit den aus theoretischen Konzepten und empirischen Befunden abgeleiteten Hypothesen heraus, dass die Ähnlichkeit der Werthaltungen für die Beziehungsqualität um so wichtiger wird, je länger die Beziehung besteht.

Text 2, Version 1

Document Type	Review Article
Title	Psychologische Risikofaktoren für Scheidung: Ein Überblick
Book / Journal	Psychologische Rundschau, 52 (2), 85-95
Publisher	Hogrefe
Date of Publication	2001
Author(s)	Bodenmann, G.
Institution	N.A.
Place	Göttingen
Abstract	In diesem Beitrag werden aktuelle empirische Ergebnisse zur Bedeutung von psychologischen Faktoren, vor allem von Persönlichkeitsvariablen, kognitiven Aspekten, Kommunikation, Bindung, Stress und Coping für einen ungünstigen Partnerschaftsverlauf und Scheidung resümiert, wobei nur auf Forschungsbefunde und nicht auf theoretische Ansätze eingegangen wird. Der Überblick zeigt, dass heute eindeutige Risikofaktoren für eine negative Entwicklung der Partnerschaft und Scheidung bekannt sind. Entgegen der in der Bevölkerung vertretenen Meinung, dass es sich dabei um Attraktivität, Status usw. handelt, zeigt die Forschung, dass vor allem emotionale Labilität (Neurotizismus) und ein Mangel an Kompetenzen bezüglich Kommunikation und Stressbewältigung prädiktive Bedeutung haben. Das Wissen über andere Bedingungen, z.B. Bindungsstil, ist dagegen noch zu wenig gesichert. Die klinische Relevanz der Scheidungsursachenforschung für die Prävention bei Paaren wird aufgezeigt, und drei, aus der Grundlagenforschung entwickelte Präventionsansätze werden dargestellt.

Text 3, Version 1	
Document Type	Edited Book
Title	Entwicklung von Paarbeziehungen
Book / Journal	In P. Kaiser (Ed.), Partnerschaft und Paartherapie
Publisher	Hogrefe
Date of Publication	2000
Author(s)	Schneewind, K. A., Graf, J., Gerhard, A. K.
Institution	Die deutsche Bibliothek
Place	Göttingen
Abstract	N. A

Text 4, Version 1	
Document Type	Textbook
Title	Paarbeziehungen
Book / Journal	Familienpsychologie kompakt
Publisher	Beltz
Date of Publication	2009
Author(s)	Jungbauer, J.
Institution	Katholische Hochschule Nordrhein-Westfalen
Place	Weinheim
Abstract	N. A

Text 5, Version 1	
Document Type	Monography
Title	Bewältigung von Stress in Partnerschaften: Der Einfluss von Belastungen auf die Qualität und Stabilität von Paarbeziehungen
Book / Journal	Freiburger Beiträge zur Familienforschung: Band 2
Publisher	Universitätsverlag Bern, Hans Huber
Date of Publication	1995
Author(s)	Bodenmann, G.
Institution	Die deutsche Bibliothek
Place	Freiburg
Abstract	N. A

Text 6, Version 1	
Document Type	Popular Science Book
Title	Schweigen ist Gold – oder Blei
Book / Journal	Du kannst mich einfach nicht verstehen
Publisher	Weltbild Verlag
Date of Publication	1997
Author(s)	Tannen, D.
Institution	Berchtermünz
Place	Augsburg
Abstract	N. A

Text 7, Version 1	
Document Type	Popular Science Article
Title	Mund zu, Herz auf
Book / Journal	NEON
Publisher	Gruner und Jahr
Date of Publication	2013
Author(s)	Weiss, V.
Institution	N. A.
Place	Hamburg
Abstract	N. A

Text 1, Version 2

Document Type	Original Empirical Article
Title	Vermeidung und Depression: Die psychometrische Evaluation der deutschen Version der „Cognitive- Behavioral Avoidance Scale“ (CBAS)
Book / Journal	Diagnostica, 56 (2), 46-55
Publisher	Hogrefe
Date of Publication	2009
Author(s)	Röthlin, P., Holtforth, M. G., Bergomi, C., & Berking, M.
Institution	N. A.
Place	Göttingen
Abstract	Vermeidungsprozesse sind bei der Beschreibung und Erklärung von Angststörungen zentral, spielen aber auch bei Depressionen eine wichtige Rolle. Die Cognitive-Behavioral Avoidance Scale erfasst depressives Vermeiden (Ottenbreit & Dobson, 2004). Die deutsche Übersetzung wurde an einer Stichprobe von insgesamt 657 Probanden (187 ambulante Psychotherapiepatienten, 376 Normalpersonen und 94 Studenten) evaluiert. Die Faktorenstruktur der Originalversion konnte mittels konfirmatorischer Faktorenanalyse bestätigt werden. Psychometrische Analysen ergaben mehrheitlich sehr zufriedenstellende Ergebnisse mit einer internen Konsistenz = .92 und einer Retestreliabilität von r = .80 (Totalskala). Hinweise auf Konstruktvalidität konnten mittels AAQ und EMOREG gewonnen werden. Zusammenhänge zwischen Vermeidung und Depressivität unabhängig von Angst unterstützen die Spezifität depressiven Vermeidens.

Text 2, Version 2

Document Type	Review Article
Title	Genetik depressiver Störungen
Book / Journal	Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 36 (1), 27–43
Publisher	Hans Huber
Date of Publication	2008
Author(s)	Schulte-Körne, G., Allgaier, A. K.
Institution	Klinik für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie, Klinikum der Universität München
Place	Bern
Abstract	Depressive Störungen gehören weltweit zu den häufigsten psychiatrischen Erkrankungen, die die psychische und psychosoziale Entwicklung der Erkrankten nachhaltig beeinflussen. Meist beginnen die Erkrankungen im Kindes- und Jugendalter. Anhand der Symptomatik, des Verlaufs und der Ursachen werden unipolare Depressionen von bipolaren Störungen, die durch depressive und manische Erkrankungsphasen gekennzeichnet sind, unterschieden. Für die Entstehung dieser Erkrankungen spielen genetische Faktoren eine entscheidende Rolle. Familien- und Zwillingsstudien konnten das erhöhte Erkrankungsrisiko von Kindern in betroffenen Familien und die hohe Heritabilität, insbesondere von bipolaren Störungen, eindrücklich nachweisen. Die Suche nach prädisponierenden Krankheitsgenen mittels Kopplungs- und Assoziationsanalysen konnte in den vergangenen Jahren beachtliche Fortschritte erzielen. Insbesondere das s-Allel des Serotonintransportergens wurde wiederholt als Risikofaktor bestätigt. Meta-Analysen deuten allerdings auf relativ begrenzte Effekte einzelner Gene hin. Neben genetischen Komponenten sind Umweltfaktoren maßgeblich an der Krankheitsgenese beteiligt: Bei unipolaren Depressionen wird die Erkrankungswahrscheinlichkeit bei entsprechender genetischer Disposition wesentlich durch protektive oder pathogene Umweltfaktoren im Sinne einer engen Gen-Umwelt-Interaktion moduliert.

Text 3, Version 2	
Document Type	Edited Book
Title	Kognitive Modelle der Depression
Book / Journal	In H. Böker (Ed.), Depression, Manie und schizoaffektive Psychosen: Psychodynamische Theorien, einzelfallorientierte Forschung und Psychotherapie
Publisher	Psychosozial-Verlag
Date of Publication	2000
Author(s)	Böker, H.
Institution	Die deutsche Bibliothek
Place	Gießen
Abstract	N. A

Text 4, Version 2	
Document Type	Textbook
Title	Beschreibung und Klassifikation depressiver Störungen
Book / Journal	Depression bei Kindern und Jugendlichen: Psychologisches Grundlagenwissen
Publisher	Ernst Reinhardt
Date of Publication	2002
Author(s)	Essau, C.
Institution	Die deutsche Bibliothek
Place	München
Abstract	N. A

Text 5, Version 2	
Document Type	Monography
Title	Verlust und Trennungserfahrungen
Book / Journal	Psychotherapie der Depression
Publisher	Hans Huber
Date of Publication	2011
Author(s)	Böker, H.
Institution	Psychiatrische Universitätsklinik, Zentrum für Depressions- und Angstbehandlung
Place	Bern
Abstract	N. A

Text 6, Version 2	
Document Type	Popular Science Book
Title	Den inneren Tyrannen bekämpfen
Book / Journal	Depressionen verstehen und bewältigen
Publisher	Verlag für Angewandte Psychologie
Date of Publication	1999
Author(s)	Gilbert, P.
Institution	Berchtermünz
Place	Göttingen
Abstract	N. A

Text 7, Version 2	
Document Type	Popular Science Article
Title	Das Beziehungsdefizitsyndrom: Warum Frauen depressiv werden – und Männer nicht wirklich daran Schuld sind
Book / Journal	Psychologie Heute
Publisher	Beltz
Date of Publication	2012
Author(s)	Nuber, U.
Institution	N. A.
Place	Weinheim
Abstract	N. A

Test Items – Plausibility Task

No.	Version	Test Item	Plausibility	Argumentation Fallacy
1	1	Die Entwicklung zum Raucher und später ggf. auch zum Nicht-Mehr-Raucher basiert auf dem Zusammenwirken einer Vielzahl sozialer, psychologischer und biologischer Faktoren.	Plausible	
2	1	Eine zentrale Rolle scheint dabei das Konstrukt der ererbten Nikotinsensitivität zu spielen. Dieses Konstrukt bezieht sich auf die Tatsache, dass manche Menschen sensitiver auf Nikotin reagieren, weil sie sensibler auf Nikotin ansprechen.	Implausible	Circular Reasoning
3	1	Mit ihm soll erklärt werden, warum manche Menschen – obwohl sie schon eine relativ große Zahl von Zigaretten geraucht haben – nicht vom Nikotin abhängig werden und Gelegenheitsraucher bleiben, während andere eine hochgradige Nikotinabhängigkeit entwickeln.	Plausible	
4	1	Die Frage nach der genetischen Prädisponiertheit für das Rauchen bezieht sich aber nicht nur auf differentielle Unterschiede bei der Nikotinsensitivität, sondern darüber hinaus auch auf angeborene Persönlichkeitsunterschiede.	Implausible	False Dichotomy
5	1	Auf der Grundlage der Eysenckschen Drei-Faktoren-Theorie wurde insbesondere ein positiver Zusammenhang zwischen dem Rauchen und dem Persönlichkeitsmerkmal „Extraversion“ postuliert und auch in einer Vielzahl von Studien empirisch nachgewiesen (z.B. Lipkus, Barefoot, Williams & Siegler, 1994).	Plausible	
6	1	Gleiches gilt auch für den Zusammenhang zwischen Rauchen und dem von Zuckerman postulierten Persönlichkeitsmerkmal „Sensation Seeking“, der Suche nach immer neuen Reizen und Stimulationen.	Implausible	Wrong Example

		Menschen, die eher zurückhaltend wirken und ihre Aufmerksamkeit mehr auf ihr Innenleben richten und solche, die wenig Bedürfnis nach verschiedenen, neuen, intensiven und risikoreichen Sinneserfahrungen verspüren, haben somit ein erhöhtes Risiko, mit dem Rauchen anzufangen.		
7	1	Ob es in der biologischen Grundausstattung des Menschen wirklich Unterschiede innerhalb eines Individuums gibt, die mit der Wahrscheinlichkeit des Rauchens, seiner Initiierung, seiner Aufrechterhaltung und seines Beendens kovariieren, ist unklar.	Plausible	
8	1	Ist das Rauchverhalten erst einmal zur Gewohnheit geworden, ist es für die meisten jedenfalls schwer, wieder davon loszukommen, da es Unterschiede in der Nikotinsensitivität gibt.	Implausible	False Conclusion
9	1	Unklar ist, in welcher Weise solche differentiellen Faktoren Einfluss auf jene aktuell ablaufenden kognitiven und emotionalen Prozesse haben, die für die Herausbildung, die Beibehaltung oder den Abbruch des Rauchverhaltens in einer spezifischen Lebenslage unmittelbar verantwortlich sind.	Plausible	
10	1	Während in den Aneignungsphasen des Rauchverhaltens vor allem biopsychologische Einflussgrößen eine wichtige Rolle spielen, so scheint im Stadium der Aufrechterhaltung das Rauchverhalten in erster Linie eine Funktion interpersonaler Faktoren zu sein.	Plausible	
11	1	In einem Experiment wurden Gewohnheitsraucher über mehrere Wochen hinweg mit Zigaretten versorgt, die entweder stark oder schwach nikotinhaltig waren, ohne dass dies für die Probanden erkennbar war. Die Probanden rauchten im Durchschnitt 25% mehr von den leichten Zigaretten als von den starken. Dieses Ergebnis ist ein eindeutiger	Implausible	Overgeneralisation

		Beweis dafür, dass es Unterschiede in der Nikotinsensitivität gibt.		
1	2	Wenn Kinder bzw. Jugendliche anfangen, darüber nachzudenken, einmal eine Zigarette zu probieren, werden erstmals auf das Rauchen bezogene Vorstellungen und Erwartungen herausgebildet. Man bezeichnet dieses Stadium auch als die Phase der Vorbereitung.	Plausible	
2	2	So führt z.B. die Erwartung, dass einem das Rauchen Stresssituationen erleichtert, eher dazu, dass man in solchen Situationen raucht, als die Erwartung, dass das Rauchen in Belastungssituationen hilfreich ist.	Implausible	False Dichotomy
3	2	Da eigene Erfahrungen mit Zigaretten noch nicht vorliegen, basiert die Bildung solcher Erwartungen hauptsächlich auf der Beobachtung des Modellverhaltens der relevanten Bezugspersonen.	Plausible	
4	2	Lässt der Vater z.B. nach dem Essen erkennen, wie herrlich ihm jetzt die Verdauungszigarette schmeckt, wird bei den anwesenden Kindern eher eine negative Vorstellung des Zigarettenrauchens bekräftigt.	Implausible	Wrong Example
5	2	In einer Längsschnittuntersuchung von Dinh, Sarason, Peterson und Onstad (1995) ist gezeigt worden, dass solche positiven Vorstellungen von den Wirkungen des Rauchens bei Fünftklässlern ein signifikanter Prädiktor für das Rauchverhalten vier Jahre später sind und dass diese positiven Vorstellungen bei Fünftklässlern das künftige Rauchverhalten stärker beeinflussen als entsprechende Vorstellungen bei Siebtklässlern.	Plausible	
6	2	Die kognitive Vorbereitung auf das spätere Rauchen scheint bereits in der Grundschule zu beginnen und einen nachhaltigen Einfluss auf die weitere Entwicklung des Rauchverhaltens zu besitzen, weil sie bereits in der Kindheit anfängt und einen anhaltenden Effekt hat.	Implausible	Circular Reasoning
7	2	Mit dem Rauchen der ersten Zigarette gelangen Jugendliche in eine Experimentierphase. Da etwa 80% bis 90% aller Jugendlichen	Plausible	

		wenigstens einmal eine Zigarette rauchen, trägt dieses Experimentieren nicht den Charakter eines abweichenden Verhaltens, sondern eher den einer normativen Entwicklungsaufgabe.		
8	2	Kritisch für die Herausbildung des gewohnheitsmäßigen Rauchens ist nicht der Umstand, dass eine erste Zigarette geraucht wird, sondern die Art und Weise, wie anschließend das Erlebnis dieser ersten Zigarette kognitiv und emotional verarbeitet wird.	Plausible	
9	2	Aus der Tatsache, dass am Ende der Jugendzeit der Anteil der gelegentlichen und regelmäßigen Raucher zusammengenommen auf etwa 50% absinkt, lässt sich ableiten, dass das Interesse am Rauchen im Laufe der Jugendzeit sogar noch zunimmt.	Implausible	False Conclusion
10	2	Eine Studie fand, dass Jugendliche eines Jungeninternats vor allem dann mit dem Rauchen anfingen, wenn sie keinen Anschluss an eine Clique fanden oder von einer solchen ausgeschlossen wurden. Der Einfluss der Peergruppe ist daher ein sehr wichtiger Faktor, um mit dem Rauchen anzufangen.	Implausible	Overgeneralisation
11	2	Natürlich gibt es noch viele andere Faktoren, die bei der Entwicklung zum Raucher und der Aufrechterhaltung des Rauchens eine Rolle spielen. Eine umfassende Theorie fehlt bislang.	Plausible	

Scoring System – Credibility Task

Main Category	Subcategory	Second Subcategory	Examples (translated from German)
Source Information	Author Information	Author Expertise	"The author is not an expert for cognitive theories of depression, because he is a psychoanalyst."
		Biases / Conflicts of Interest	"The author argues from a psychoanalytical perspective."
	Publication Outlet	Quality of Publisher	"Ah, Diagnostica, a high-quality journal."
		Biases / Conflicts of Interest	"This is a popular journal, they mainly want to reach many people with spectacular results."
	Document Type / Text Genre	Peer-Review vs. Non-Peer-Review	"This document has not been under peer-review."
		Primary vs. Secondary	"This is an original article that can be verified."
		Scientific vs. Popular	"This looks like a self-help book for depression. It has no scientific value."
	Date of Publication	Topicality	"This is a relatively old and probably outdated text. "
	Original Document Language	English vs. German	"Ah, originally, it was published in English... higher impact."
Content	Topic Knowledge	Accuracy	"The description of the gen-environment interaction is correct."

	Topic Beliefs / Opinion		"This is consistent with my own experiences. If you have similar expectations, you don't get disappointed so easily."
	Relevance of Topic		"The central research question is relevant."
	Complexity of Information		"Results are discussed from several perspectives."
	Title Information		"The title is more of an eye-catcher, not very informative."
	Theoretical Foundation		"The concept of avoidance is explained well."
	Abstract		"There is an abstract that nicely summarises the article."
Argumentation	General Line of Argumentation		"The arguments are generally coherent."
		Relevance	"The use relevant empirical findings to support this."
		Sufficiency	"This is very one-sided…They don't discuss any alternative views."
	Quality of Stated Evidence		"The cited literature is relevant, but predominantly not empirical."
	Quantity of Stated Evidence		"The claims are supported with several references."

Writing Style	Comprehensibility		“There are many technical terms I don’t understand.”
	Clarity		“The methods are described clearly.”
	Objectivity		“Some findings are described correctly, but they are exaggerated.“
Method	Research Method		“They used three different samples.”
	Statistical Data Analysis		“They did a confirmatory factor analysis, used different models that were tested against each other”.
Structure	Topical Structure		“This is all clearly structured: Introduction, method section, results, discussion.”
	General Layout		“The layout looks a bit like a scientific article with an experiment.”
Other	Presence of Figures		“Results are explained with a figure.”
	Presence of Tables		“There is a table that gives an overview of the most important results .”
	Formal Aspects	Use and Correctness of APA-Style	“The literature is cited correctly.”
		Spelling and Grammar	“The word “Widerspruch” (contradiction) is incorrectly spelt with an “ie”.”

Retrospective Interview – Credibility Task		
No.	**Question**	**Probing Technique**
1	War Ihnen die Aufgabenstellung deutlich?	General Probing
1a	Falls nein, was war unklar? [Vgl. Frage 1]	Category Selection Probing
2	Was verstehen Sie unter einem glaubwürdigen Text? Erklären Sie bitte in Ihren eigenen Worten.	Special Comprehension Probing
3	Wie sind Sie bei der Beurteilung der Glaubwürdigkeit der verschiedenen Texte vorgegangen?	Information Retrieval Probing
4	Welche Kriterien waren bei Ihrer Entscheidung wichtig? Welche Merkmale und Informationen haben Sie genutzt?	Information Retrieval Probing
5	Auf einer Skala von 1 = *sehr einfach* bis 6 = *sehr schwierig*, wie schwierig fanden Sie es, in der vorgegebenen Zeitbegrenzung ein Urteil zu treffen?	General Probing
5a	Warum fanden Sie es sehr einfach/ einfach/ eher einfach/ eher schwierig/ schwierig/ sehr schwierig? [Vgl. Frage 5]	Category Selection Probing
6	Auf einer Skala von 1 = *überhaupt nicht wichtig* bis 6 = *sehr wichtig*, welche Rolle hat die Textart bei der Beurteilung der Glaubwürdigkeit gespielt?	General Probing
6a	Warum fanden Sie die Textart als Beurteilungskriterium überhaupt nicht wichtig / nicht wichtig/ eher nicht wichtig / eher wichtig / wichtig / sehr wichtig? [Vgl. Frage 6]	Category Selection Probing
7	Auf einer Skala von 1 = *sehr einfach* bis 6 = *sehr schwierig*, wie schwierig fanden Sie es, die Textart zu erkennen?	General Probing
7a	Warum fanden Sie es sehr einfach/ einfach/ eher einfach/ eher schwierig/ schwierig/ sehr schwierig?	Category Selection Probing

[Vgl. Frage 7]

8 Haben Sie sonstige Anmerkungen zu dieser Aufgabe? General Probing

---

Appendix C

Stimulus Materials for Study 3

Test Items for Pretest, Posttest, and Follow-up – Argument Structure Task

No.	Version	Test Item	Components	Complexity	Argument Type	Words
1	1	Verschiedene ethnische Bevölkerungsgruppen unterscheiden sich erheblich in diversen sozio- emotionalen Kompetenzen [*claim*]. In einer amerikanischen Studie wurde das Spielverhalten von 84 Kindern (Asia-Amerikaner, Latein-Amerikaner und Afro-Amerikaner) im Alter von vier Jahren zu Beginn des Kindergarten-Jahres mit der Mac Arthur Story Stem Battery (MSSB) untersucht. Dabei werden Kindern konflikthafte Geschichtsanfänge mit Hilfe von Lego-Figuren präsentiert, mit der Aufforderung, die Geschichten weiter zu erzählen und zu spielen. Die Autoren fanden signifikante Unterschiede bezüglich Problemvermeidung, empathischen und moralischen Themen [*datum*]. Allerdings ist unklar, inwiefern sich die Ergebnisse außerhalb der Vereinigten Staaten von Amerika generalisieren lassen [*rebuttal*].	3	Low	Claim-First	90
2	1	Selbstkontrolle sollte so früh wie möglich trainiert werden [*claim*]. In einer Längsschnittstudie mit 653 Kindern wurde untersucht, wie sich die Bereitschaft, eine Belohnung aufzuschieben, auf die Entwicklung von Schülern auswirkt. Die Autoren stellten fest, dass Kinder, die im Alter von vier oder fünf Jahren eine Belohnung (beispielsweise einen Keks) aufschoben, wenn eine weitere Belohnung (zwei Kekse) lockte, zehn Jahre später bessere kognitive und soziale Kompetenzen aufwiesen als Kinder, die eine sofortige Belohnung vorzogen [*datum*]. Schulischer Erfolg spielt für den weiteren beruflichen Erfolg eine zentrale Rolle [*warrant*]. Abiturienten mit sehr guten Abschlussnoten haben z.B. im	5	High	Claim-First	120

		Studium oft geringere Schwierigkeiten und Belastungen sowie einen stabileren Studienverlauf [*backing*]. Natürlich gibt es neben erfolgreicher Selbstkontrolle noch viele andere Faktoren, die für die schulische Entwicklung eines Kindes verantwortlich sind [*rebuttal*].				
3	1	In einer Studie wurde der Zusammenhang zwischen klinisch relevanten Auffälligkeiten und Grundschulempfehlungen untersucht. Hierzu wurden 3910 Kinder am Ende ihrer Grundschulzeit von deren Eltern anhand anonymisierter Fragebögen (Child Behavior Checklist CBCL) beurteilt. Kinder mit Grundschulempfehlung für die Haupt- oder Förderschule zeigten eine besonders starke, multiple Problembelastung mit konstant höheren Auffälligkeiten in allen Bereichen [*datum*]. Schüler mit einem Hauptschulabschluss haben meist keine guten Chancen auf dem Arbeitsmarkt [*warrant*]. Das bestätigte auch eine Bildungsexpertin des Deutschen Jugendinstitut (DJI) [*backing*]. Auch wenn Korrelationsstudien wie die oben genannte keine kausalen Schlüsse zulassen [*rebuttal*], stellt der Zusammenhang von Problemverhalten und Schulerfolg doch einen wichtigen präventiven Ansatzpunkt zur frühzeitigen Förderung der sozialen und kognitiven Entwicklung von Grundschulkindern dar [*claim*].	5	High	Reason-First	108
4	1	In einer Studie wurde gezeigt, dass sich 44% aller Amokläufe innerhalb von 10 Tagen nach einer ausführlichen Berichterstattung nationaler als auch internationaler Tageszeitungen ereigneten [*datum*]. Die Medien stellen eine zentrale Informationsquelle für die breite Bevölkerung dar [*warrant*]. 48,5 Millionen Deutsche lesen jede Ausgabe einer täglich oder wöchentlich erscheinenden Zeitung [*backing*]. Die Berichterstattung über Amokläufe in den Medien kann somit Amokläufe begünstigen und muss deutlich reduziert werden [*claim*]. Natürlich reicht eine reduzierte	5	High	Reason-First	76

		Berichterstattung allein nicht aus, um einen Amoklauf zu verhindern [*rebuttal*].				
1	2	Religiöse Menschen sind eher bereit dazu, zu verzeihen als nicht religiöse Menschen [*claim*]. Eine Studie zeigte, dass eine Gruppe von regelmäßigen Kirchgängern eine größere Bereitschaft zu verzeihen zeigte als solche, die nicht zur Kirche gingen und diese auch von einer verzeihenden Haltung gegenüber Missetätern berichteten [*datum*]. Allerdings spielen natürlich auch andere Faktoren, wie z.B. das Alter, eine wichtige Rolle beim Verzeihen [*rebuttal*].	3	Low	Claim-First	59
2	2	Unternehmen sollten Maßnahmen ergreifen, um Weiterbildungsangebote zum Thema psychische Belastungen am Arbeitsplatz anzubieten [*claim*]. Bei Unternehmen, bei denen entsprechende Weiterbildungsangebote eingerichtet wurden, konnten laut einer Studie die psychosozialen Risiken um 86.6% gegenüber 64% bei Unternehmen ohne solche Verfahren verringert werden [*datum*]. Für die Gesundheit der Mitarbeiter und den Erfolg eines Unternehmens ist es äußerst wichtig, psychische Belastungen am Arbeitsplatz zu vermeiden [*warrant*]. Menschen sind zufriedener in ihrem Job und arbeiten besser, wenn sie sich nicht psychisch belastet fühlen [*backing*]. Auch, wenn Weiterbildungsprogramme natürlich nicht jedem helfen werden [*rebuttal*].	5	High	Claim-First	83
3	2	In einer Studie wurden 850 gesunde 13- bis 18-jährige Mädchen und Jungen zu ihrem Körpergewicht, ihrer Größe und ihrem Körperselbstbild befragt. Dabei stellten die Autoren der Studie fest, dass die Einschätzung, „zu dick“ zu sein, das größte Risiko einer Essstörung wie Magersucht oder Bulemie darstellt [*datum*]. Diese Einschätzung führt oft zu einem verzerrten Selbstbild [*warrant*]. Viele psychotherapeutische Einrichtungen, die sich mit	5	High	Reason-First	96

		dieser Problematik beschäftigen, berichten von verzerrten Selbstbildern bei Patienten dieser Altersgruppe [*backing*]. Präventionsprogramme sollten sich daher speziell mit dem Selbstbild der Jugendlichen auseinandersetzen [*claim*]. Auch, wenn neben einem verzerrten Selbstbild natürlich viele andere Faktoren zur Entstehung einer Essstörung beitragen [*rebuttal*].				
4	2	In einer Studie wurde bei 116 Patienten mit vorwiegend vaskulär und traumatisch bedingten zerebralen Schädigungen eine neuropsychologische Untersuchung mit besonderer Berücksichtigung von Aufmerksamkeit, Reaktionsfähigkeit und visueller Auffassungsschnelligkeit sowie eine umfangreiche Fahrprobe im öffentlichen Straßenverkehr vorgenommen. Nur 58% der Patienten bestanden nach dem Urteil des Fahrlehrers die Fahrprobe [*datum*]. Ein solch unzureichendes sicheres Fahrverhalten gefährdet den Straßenverkehr [*warrant*]. Diverse Studien bestätigen jährlich viele Tote aufgrund eines unsicheren Fahrverhaltens [*backing*]. Nach einer Hirnschädigung sollte deshalb eine Fahrprobe durchgeführt werden, bevor der Betroffene wieder ein Kraftfahrzeug führt [*claim*]. Allerdings sollte man mit Verallgemeinerungen vorsichtig sein, denn nicht jede Form der Hirnschädigung wirkt sich beeinträchtigend auf das Fahrverhalten aus [*rebuttal*].	5	High	Reason-First	101
1	3	Es ist dringend erforderlich, bei Untersuchungen der Kindheitsentwicklung auch soziokulturelle Bedingungen in unterschiedlichen ethnischen Gruppen miteinzubeziehen [*claim*]. Zwar sind die Klassifikationen von sicherer, unsicherer/abhängiger und unabhängiger Bindungsstildimensionen universell [*rebuttal*]. Studien zeigen jedoch, dass europäische und amerikanische Mütter Wert auf Autonomie in der Beziehung zu ihrem Kind legten. Puertorikanische Mütter hingegen achteten	3	Low	Claim-First	56

		eher auf familiär bezogenes und respektvolles Verhalten [*datum*].				
2	3	Die Befürchtungen, dass Deutschland von einer Welle der Alterskriminalität überrollt wird, sind nicht berechtigt und sollten ausgeräumt werden [*claim*]. Beim sexuellen Missbrauch von Kindern sind ältere Tatverdächtige mit 0.5% des Anteils gegenüber 0.4% in der Gesamtbevölkerung zwar tatsächlich überrepräsentiert [*rebuttal*], aber laut einer Studie des Statistischen Bundesamtes trägt die Gruppe der über 60-Jährigen zu den meisten Delikten (insbesondere Gewaltkriminalität, Raub und Körperverletzung) unterdurchschnittlich stark bei [*datum*]. Alterskriminalität ist also immer noch recht selten [*warrant*]. Das zeigen auch Zahlen aus anderen Ländern [*backing*].	5	High	Claim-First	77
3	3	An einer repräsentativen Stichprobe deutscher Jugendlicher wurden Zusammenhänge zwischen riskanter und pathologischer Internetnutzung mit Depressivität sowie selbstverletzendem und suizidalem Verhalten untersucht. Riskante (14.5%) und pathologische (4.8%) Internetnutzer zeigten im Vergleich zu Schülern mit unauffälliger Internetnutzung (80.7%) signifikant höhere Ausprägungen in Depressivität, selbstverletzendem und suizidalen Verhalten [*datum*]. Eine solche Gefährdung hat auch negative gesellschaftliche Konsequenzen [*warrant*]. Depressive Jugendliche zeigen meist auch Leistungsdefizite und Schwierigkeiten im sozialen Bereich [*backing*]. Es sollte daher mehr Aufmerksamkeit auf Jugendliche mit riskanter Internetnutzung verwendet werden [*claim*]. Es ist jedoch ebenfalls festzuhalten, dass die Internetnutzung für die meisten Jugendlichen keine Gefahr darstellt [*rebuttal*].	5	High	Reason-First	91

4	3	Entwicklungspsychologische Befunde zeigen deutlich, dass Heranwachsende in vielen psychischen Funktionsbereichen noch wie Jugendliche organisiert sind. Zum Beispiel konnte in einer Studie gezeigt werden, dass Impulsivität, wie auch das „sensation-seeking“, d.h. das Bedürfnis nach Reizstimulation, zwischen mittlerer Adoleszenz und frühem Erwachsenenalter ansteigt und erst danach wieder abfällt [*datum*]. Impulsivität und Sensation-seeking stehen beide im Zusammenhang mit höherem Risikoverhalten im Straßenverkehr [*warrant*]. So verursachen z.B. Menschen, bei denen diese Merkmale besonders ausgeprägt sind, mehr Unfälle [*backing*]. Heranwachsende sollten deshalb bei Vergehen im Straßenverkehr noch dem Jugendstrafrecht unterworfen werden [*claim*]. Auch wenn im Einzelfall gegebenenfalls anders entschieden werden muss [*rebuttal*].	5	High	Reason-First	91

Test Items for Pretest, Posttest, and Follow-up – Plausibility Task

No.	Version	Test Item	Plausibility	Argumentation Fallacy
1	1	Die Entwicklung zum Raucher und später ggf. auch zum Nicht-Mehr-Raucher basiert auf dem Zusammenwirken einer Vielzahl sozialer, psychologischer und biologischer Faktoren.	Plausible	
2	1	Eine zentrale Rolle scheint dabei das Konstrukt der ererbten Nikotinsensitivität zu spielen. Dieses Konstrukt bezieht sich auf die Tatsache, dass manche Menschen sensitiver auf Nikotin reagieren, weil sie sensibler auf Nikotin ansprechen.	Implausible	Circular Reasoning
3	1	Mit ihm soll erklärt werden, warum manche Menschen – obwohl sie schon eine relativ große Zahl von Zigaretten geraucht haben – nicht vom Nikotin abhängig werden und Gelegenheitsraucher bleiben, während andere eine hochgradige Nikotinabhängigkeit entwickeln.	Plausible	
4	1	Die Frage nach der genetischen Prädisponiertheit für das Rauchen bezieht sich aber nicht nur auf differentielle Unterschiede bei der Nikotinsensitivität, sondern darüber hinaus auch auf angeborene Persönlichkeitsunterschiede.	Implausible	False Dichotomy
5	1	Auf der Grundlage der Eysenckschen Drei-Faktoren-Theorie wurde insbesondere ein positiver Zusammenhang zwischen dem Rauchen und dem Persönlichkeitsmerkmal „Extraversion“ postuliert und auch in einer Vielzahl von Studien empirisch nachgewiesen (z.B. Lipkus, Barefoot, Williams & Siegler, 1994).	Plausible	
6	1	Gleiches gilt auch für den Zusammenhang zwischen Rauchen und dem von Zuckerman postulierten Persönlichkeitsmerkmal „Sensation	Implausible	Wrong Example

		Seeking“, der Suche nach immer neuen Reizen und Stimulationen. Menschen, die eher zurückhaltend wirken und ihre Aufmerksamkeit mehr auf ihr Innenleben richten und solche, die wenig Bedürfnis nach verschiedenen, neuen, intensiven und risikoreichen Sinneserfahrungen verspüren, haben somit ein erhöhtes Risiko, mit dem Rauchen anzufangen.		
7	1	Ob es in der biologischen Grundausstattung des Menschen wirklich Unterschiede innerhalb eines Individuums gibt, die mit der Wahrscheinlichkeit des Rauchens, seiner Initiierung, seiner Aufrechterhaltung und seines Beendens kovariieren, ist unklar.	Plausible	
8	1	Ist das Rauchverhalten erst einmal zur Gewohnheit geworden, ist es für die meisten jedenfalls schwer, wieder davon loszukommen, da es Unterschiede in der Nikotinsensitivität gibt.	Implausible	False Conclusion
9	1	Unklar ist, in welcher Weise solche differentiellen Faktoren Einfluss auf jene aktuell ablaufenden kognitiven und emotionalen Prozesse haben, die für die Herausbildung, die Beibehaltung oder den Abbruch des Rauchverhaltens in einer spezifischen Lebenslage unmittelbar verantwortlich sind.	Plausible	
10	1	Während in den Aneignungsphasen des Rauchverhaltens vor allem biopsychologische Einflussgrößen eine wichtige Rolle spielen, so scheint im Stadium der Aufrechterhaltung das Rauchverhalten in erster Linie eine Funktion interpersonaler Faktoren zu sein.	Plausible	
11	1	In einem Experiment wurden Gewohnheitsraucher über mehrere Wochen hinweg mit Zigaretten versorgt, die entweder stark oder schwach nikotinhaltig waren, ohne dass dies für die Probanden erkennbar war. Die Probanden rauchten im Durchschnitt 25% mehr von den leichten Zigaretten als von den starken. Dieses Ergebnis ist ein eindeutiger	Implausible	Overgeneralisation

		Beweis dafür, dass es Unterschiede in der Nikotinsensitivität gibt.		
1	2	Wenn Kinder bzw. Jugendliche anfangen, darüber nachzudenken, einmal eine Zigarette zu probieren, werden erstmals auf das Rauchen bezogene Vorstellungen und Erwartungen herausgebildet. Man bezeichnet dieses Stadium auch als die Phase der Vorbereitung.	Plausible	
2	2	So führt z.B. die Erwartung, dass einem das Rauchen Stresssituationen erleichtert, eher dazu, dass man in solchen Situationen raucht, als die Erwartung, dass das Rauchen in Belastungssituationen hilfreich ist.	Implausible	False Dichotomy
3	2	Da eigene Erfahrungen mit Zigaretten noch nicht vorliegen, basiert die Bildung solcher Erwartungen hauptsächlich auf der Beobachtung des Modellverhaltens der relevanten Bezugspersonen.	Plausible	
4	2	Lässt der Vater z.B. nach dem Essen erkennen, wie herrlich ihm jetzt die Verdauungszigarette schmeckt, wird bei den anwesenden Kindern eher eine negative Vorstellung des Zigarettenrauchens bekräftigt.	Implausible	Wrong Example
5	2	In einer Längsschnittuntersuchung von Dinh, Sarason, Peterson und Onstad (1995) ist gezeigt worden, dass solche positiven Vorstellungen von den Wirkungen des Rauchens bei Fünftklässlern ein signifikanter Prädiktor für das Rauchverhalten vier Jahre später sind und dass diese positiven Vorstellungen bei Fünftklässlern das künftige Rauchverhalten stärker beeinflussen als entsprechende Vorstellungen bei Siebtklässlern.	Plausible	
6	2	Die kognitive Vorbereitung auf das spätere Rauchen scheint bereits in der Grundschule zu beginnen und einen nachhaltigen Einfluss auf die weitere Entwicklung des Rauchverhaltens zu besitzen, weil sie bereits in der Kindheit anfängt und einen anhaltenden Effekt hat.	Implausible	Circular Reasoning
7	2	Mit dem Rauchen der ersten Zigarette gelangen Jugendliche in eine Experimentierphase. Da etwa 80% bis 90% aller Jugendlichen	Plausible	

		wenigstens einmal eine Zigarette rauchen, trägt dieses Experimentieren nicht den Charakter eines abweichenden Verhaltens, sondern eher den einer normativen Entwicklungsaufgabe.		
8	2	Kritisch für die Herausbildung des gewohnheitsmäßigen Rauchens ist nicht der Umstand, dass eine erste Zigarette geraucht wird, sondern die Art und Weise, wie anschließend das Erlebnis dieser ersten Zigarette kognitiv und emotional verarbeitet wird.	Plausible	
9	2	Aus der Tatsache, dass am Ende der Jugendzeit der Anteil der gelegentlichen und regelmäßigen Raucher zusammengenommen auf etwa 50% absinkt, lässt sich ableiten, dass das Interesse am Rauchen im Laufe der Jugendzeit sogar noch zunimmt.	Implausible	False Conclusion
10	2	Eine Studie fand, dass Jugendliche eines Jungeninternats vor allem dann mit dem Rauchen anfingen, wenn sie keinen Anschluss an eine Clique fanden oder von einer solchen ausgeschlossen wurden. Der Einfluss der Peergruppe ist daher ein sehr wichtiger Faktor, um mit dem Rauchen anzufangen.	Implausible	Overgeneralisation
11	2	Natürlich gibt es noch viele andere Faktoren, die bei der Entwicklung zum Raucher und der Aufrechterhaltung des Rauchens eine Rolle spielen. Eine umfassende Theorie fehlt bislang.	Plausible	
1	3	Das Selbstbild oder Selbstkonzept ist ein Forschungsgebiet der Sozialpsychologie, bei dem es um das im Langzeitgedächtnis gespeicherte Wissen eines Menschen über sich selbst geht.	Plausible	
2	3	Die Theorie der objektiven Selbstaufmerksamkeit beschäftigt sich damit, was passiert, wenn wir unsere Aufmerksamkeit auf unser Selbstbild richten. Allgemein wird mit "objektiver Selbstaufmerksamkeit" dabei ein	Plausible	

		Zustand bezeichnet, bei dem die Aufmerksamkeit nach innen, auf die eigene Person gerichtet ist.		
3	3	Eine Auswirkung des Zustandes der objektiven Selbstaufmerksamkeit wird darin gesehen, dass durch die Ausrichtung der Aufmerksamkeit auf die eigene Person Diskrepanzen zwischen dem Selbstideal (Anspruchsniveau in verschiedenen Bereichen) und dem realistischen Selbstbild stärker bewusst werden, weil dadurch diese Unterschiede deutlicher wahrgenommen werden.	Implausible	Circular Reasoning
4	3	Dies kann sowohl positive als auch negative Selbstbewertungen zur Folge haben, je nachdem ob man z.B. einen überraschenden Erfolg erlebt (positive Selbstbewertung), oder ob man seinen Ansprüchen nicht gerecht geworden ist (negative Selbstbewertung).	Plausible	
5	3	Die Theorie postuliert, dass objektive Selbstaufmerksamkeit Diskrepanzen, sowohl im negativen als auch im positiven Sinne, zwischen Selbstideal und Wirklichkeit hervorhebt. Wenn eine positive Diskrepanz vorliegt, entstehen positive Emotionen, aber andererseits auch eine positive Selbstbewertung.	Implausible	False Dichotomy
6	3	Eine wichtige Hypothese der Theorie der objektiven Selbstaufmerksamkeit lautet, dass man im Zustand der objektiven Selbstaufmerksamkeit versucht, Diskrepanzen zwischen Anspruch und Wirklichkeit zu reduzieren, z.B. durch Anpassung des Verhaltens an die eigenen Einstellungen und Normen.	Plausible	
7	3	Die Vorhersagen der Theorie der objektiven Selbstaufmerksamkeit zur Wahrnehmung von Diskrepanzen zwischen verschiedenen Aspekten des Selbst sind überwiegend in Form experimenteller Untersuchungen überprüft worden, indem die Versuchspersonen Reizen ausgesetzt	Plausible	

		werden, die die Aufmerksamkeit auf die eigene Person lenken, z.B. Spiegel.		
8	3	In einem Experiment bearbeiteten die Versuchspersonen zunächst verschiedene leistungsthematische Aufgaben. Im Anschluss gab ihnen der Versuchsleiter eine extrem positive Rückmeldung (z.B. einen deutlichen Tadel) über ihre Aufgabenbearbeitung.	Implausible	Wrong Example
9	3	Nach der Rückmeldung füllten die Versuchspersonen einen Selbsteinschätzungsfragebogen aus, indem eigene Leistungen und Fähigkeiten beurteilt werden sollten. Die Hälfte der Versuchspersonen saß dabei vor einem Spiegel, die andere Hälfte vor einer Wand.	Plausible	
10	3	Im Ergebnis beurteilten sich die Versuchspersonen in der Spiegel-Bedingung positiver als die Versuchspersonen ohne Spiegel. Die Theorie der objektiven Selbstaufmerksamkeit wurde somit eindeutig bewiesen.	Implausible	Overgeneralisation
11	3	Aus den Ergebnissen der Untersuchung lässt sich außerdem schlussfolgern, dass, wenn die Versuchspersonen anstatt eines positiven Feedbacks ein negatives Feedback erhalten hätten, dies den eigenen Selbstwert geschwächt hätte.	Implausible	False Conclusion

Example Item – Argument Structure Training

No.	Test Item	Words
1	Es ist wichtig, dieses Training zu absolvieren [*claim*], weil Argumente in der Wissenschaft eine wichtige Rolle spielen. [*datum*]. In diesem Training lernst du den erfolgreichen Umgang mit Argumenten [*warrant*]. Wir zeigen dir anhand von Beispielen, wie man Argumente aufschlüsselt und geben dir Feedback zu den Übungen [*backing*]. Allerdings ist das Training nur dann effektiv, wenn du dir Mühe gibst. [*rebuttal*].	55

Example Items– Plausibility Training

No.	Test Item	Plausibility	Argumentation Fallacy
1	Die Sonne ist bisher jeden Morgen aufgegangen. Deshalb können wir annehmen, dass sie wahrscheinlich auch morgen wieder aufgehen wird.	Plausible	
2	Ein Stierkämpfer sollte ein Mann sein. Deshalb sollten Frauen nicht am Stierkampf teilnehmen.	Implausible	Circular Reasoning
3	Man kann durch noch so viel Übung nicht einfach ein hervorragender Mathematiker werden. Herausragende mathematische Fähigkeiten sind also angeboren.	Implausible	False Conclusion
4	Heute wurde ich von einer Frau zurückgewiesen. Ich habe einfach keine Chance bei Frauen.	Implausible	Overgeneralisation
5	Durch eine künstliche Befruchtung (z.B. bei der Adoption) können ältere Frauen häufig doch noch Kinder bekommen).	Implausible	Wrong Example
6	Ich esse gern asiatisch, aber auch chinesisch.	Implausible	False Dichotomy

Tutorial – Argument Structure Training

Typisches Argument

> Man kann Kindern, die unter Alpträumen leiden, mit sehr einfachen Mitteln helfen [*Behauptung*]. So zeigt eine Studie, dass das wiederholte Malen und anschließende Zerreißen von Zeichnungen bedrohlicher Traumfiguren (z.B. Dracula) den Albtraum verschwinden ließ [*Begründung*]. Die Prozedur ist einfach, weil sie auch von Eltern im Alltag gut angewendet werden kann [*Schlussregel*]. Zum Beispiel können Eltern das Zeichnen und Zerreißen von Traumbildern in ihre tägliche Abendroutine einbinden [*Stützung der Schlussregel*]. Wichtig ist jedoch die Bereitschaft des Kindes, sich mit der Angst zu konfrontieren [*Einschränkung*].

In diesem Tutorial zeigen wir dir, wie man die verschiedenen Bestandteile eines Arguments richtig zuordnet. Dazu nehmen wir jetzt erst einmal ein ganz typisches Argument, das alle fünf Bestandteile beinhaltet und mit der Begründung beginnt, als Beispiel (siehe oben).

Zu allererst solltest du dir das ganze Argument genau durchlesen.

Anschließend ist es sinnvoll, zunächst einmal nach dem zentralen Element des Arguments zu suchen–der Behauptung. Die **Behauptung** ist definiert als eine „kontroverse These", die durch die weiteren Elemente des Arguments gestützt und eingeschränkt wird. Wenn sie nicht strittig wäre, wäre eine Begründung überflüssig, denn dann müsste man den Leser oder die Leserin ja nicht mehr überzeugen. In unserem Fall ist die Behauptung gleich im ersten Satz zu finden:

„Man kann Kindern, die unter Alpträumen leiden, mit sehr einfachen Mitteln helfen." (Behauptung)

Dies ist zunächst einmal eine These, die vom Autor aufgestellt wird. Sie sollte begründet werden. Man könnte ja beispielsweise auch annehmen, dass es eher schwierig ist, Kindern mit Alpträumen zu helfen, weil man Träume nicht so leicht beeinflussen kann.

In einem zweiten Schritt suchen wir daher nach einer entsprechenden **Begründung**, die die Behauptung mit faktischen, empirischen, oder theoretischen Belegen stützt. Wir finden sie gleich im Anschluss an die Behauptung.

„So zeigt eine Studie, dass das wiederholte Zerreißen von Zeichnungen bedrohlicher Traumfiguren (z.B. Dracula) den Albtraum verschwinden ließ." (Begründung)

In unserem Fall besteht die Begründung aus einem Fallbeispiel. Das Signalwort, das auf eine Verbindung zwischen der Behauptung und der Begründung hindeutet, ist das Wort „So".

Die Frage ist nun, warum diese Begründung für die Behauptung relevant ist. Im nächsten Schritt muss also noch geklärt werden, warum dieses Beispiel die Behauptung stützt, dass man Kindern, die unter Alpträumen leiden, mit sehr einfachen Mitteln helfen kann. Dafür ist die Schlussregel zuständig. Die **Schlussregel** wird im Alltag oft nicht explizit benannt, sondern muss vom Lesenden abgeleitet werden. In der Wissenschaft muss sie jedoch ausdrücklich benannt werden, um die Schlussfolgerungen, die aus bestimmten Forschungsergebnissen gezogen werden oder die Wahl einer bestimmten Forschungsmethode, zu rechtfertigen. In unserem Fall müssen wir uns fragen, ob es sich bei dem genannten Fallbeispiel tatsächlich um eine einfache (und nicht etwa um eine schwierige) Prozedur handelt:

„Die Prozedur ist einfach, weil sie auch von Eltern im Alltag gut angewendet werden kann.“ (Schlussregel)

Die Schlussregel bestätigt, dass es sich um eine einfache Prozedur handelt und ist daher relevant für die Behauptung. Sie bedarf allerdings noch einer weiteren Erläuterung, der **Stützung der Schlussregel**, die die Schlussregel widerum mit Belegen stützt, erläutert oder ergänzt:

„Zum Beispiel können Eltern das Zeichnen und Zerreißen von Traumbildern in ihre tägliche Abendroutine einbinden“ (Stützung der Schlussregel)

Hier wird ein praktisches Beispiel dafür angeführt, wie Eltern die Methode im Alltag einfach anwenden können. Wir erkennen die Stützung wieder an den Signalwörtern „zum Beispiel“.

Am Ende des Arguments fällt sofort ein sehr starkes Signalwort, „jedoch“, ins Auge. Es ist immer wichtig, herauszufinden, ob es möglicherweise noch andere Erklärungen für die angeführten Begründungen gibt, oder ob es **Gegenargumente** oder **Einschränkungen** der Behauptung gibt. Einschränkung können festlegen, wann die Behauptung gilt und wann nicht. In unserem Fall gilt: Die Methode funktioniert nur, wenn sich das Kind mit seiner Angst auseinandersetzen möchte.

„*Wichtig ist jedoch die Bereitschaft des Kindes, sich mit der Angst zu konfrontieren*“.

Man könnte auch Gegenargumente finden, die die Behauptung schwächen. Zum Beispiel könnte man kritisieren, dass von einer einzigen Fallstudie eigentlich keine allgemeinen Aussagen getroffen werden können.

Atypisches Argument

> Eine Studie zeigte, dass ein zielsicheres, dominantes Auftreten von Frauen in Bewerbungssituationen zu einer geringen Bewertung ihrer sozialen Kompetenz führte [*Begründung*]. Soziale Kompetenz ist jedoch in der Arbeitswelt immer mehr gefragt [*Schlussregel*]. So fand eine andere Studie, dass „Soziale Kompetenz“ in 70% der Bewerbungsverfahren ein wichtiges Kriterium im Bewerbungsprozess darstellt [*Stützung der Schlussregel*]. Insofern könnten die zunehmend gestellten Erwartungen an soziale Kompetenz paradoxerweise mehr Diskriminierung von Frauen bei der Auswahl von Führungskräften bewirken [*Behauptung*], sofern diese sich im Bewerbungsprozess als zielsichere Karrierefrau darstellen [*Einschränkung*].

Im zweiten Tutorial schauen wir uns noch ein weniger typisches Argument an, bei dem die Behauptung nicht am Anfang steht, sondern erst im Laufe des Arguments genannt wird (siehe oben). Solche Argumente nicht oft etwas schwieriger zu verarbeiten, weil sie nicht unseren Erwartungen entsprechen.

Zunächst lesen wir das Argument wieder genau durch.

Auch, wenn die **Behauptung** nicht sofort genannt wird, ist es trotzdem sinnvoll, dass wir wieder als Erstes nach diesem zentralen Element suchen, also der kontroversen These, von der uns der Autor oder die Autorin zu überzeugen versucht. Das ist in diesem Fall diese Aussage:

„Insofern könnten die zunehmend gestellten Erwartungen an soziale Kompetenz paradoxerweise mehr Diskriminierung von Frauen bei der Auswahl von Führungskräften bewirken.“ (Behauptung)

Wir erkennen die Behauptung auch an dem Signalwort „insofern“. Sie bedarf einer Begründung, denn wir könnten ja erst einmal vermuten, dass soziale Kompetenz, eine Eigenschaft, die besonders dem weiblichen Geschlecht zugeschrieben wird, Frauen einen Vorteil bei der Bewerbung verschaffen könnte. Die **Begründung** liefert die empirischen Belege für die Behauptung:

„Eine Studie zeigte, dass zielsicheres, dominantes Auftreten von Frauen in Bewerbungssituationen zu einer geringen Bewertung ihrer sozialen Kompetenz führte.“ (Begründung)

Das Problem ist also, dass, wenn Frauen zielsicher und dominant auftreten, dies dazu führen kann, dass ihre soziale Kompetenz gering eingeschätzt wird. Wir können die Begründung auch an dem Wort „zeigte“ erkennen. Warum aber führt diese geringe Einschätzung zu einer Benachteiligung von Frauen bei der Bewerberauswahl?

Die **Schlussregel** beschreibt wieder, warum diese Begründung für die Behauptung relevant ist.

„Soziale Kompetenz ist jedoch in der Arbeitswelt immer mehr gefragt.“ (Schlussregel).

Die schlechtere Bewertung in der sozialen Kompetenz bringt Frauen also einen Nachteil, weil diese Eigenschaft als wichtiges Auswahlkriterium angesehen wird, ebenso wie ein zielsicheres Auftreten. Es wird erwartet, dass Frauen sich behaupten, also dominant auftreten und gleichzeitig auch soziale Kompetenzen aufweisen. Beides können sie jedoch nicht leisten.

Die Schlussregel wird widerum durch die **Stützung der Schlussregel**, in unserem Fall einen empirischen Beleg, untermauert. Dies ist der folgende Satz:

„So fand eine andere Studie, dass „Soziale Kompetenz“ in 70% der Bewerbungsverfahren ein wichtiges Kriterium im Bewerbungsprozess darstellt“ (Stützung der Schlussregel)

Die **Einschränkung** finden wir in unserem Beispiel direkt nach der Behauptung, denn zu einer Diskriminierung von Frauen kommt es dem Argument zufolge nur dann,

„…sofern diese sich im Bewerbungsprozess als zielsichere Karrierefrau darstellen.“ (Einschränkung)

In diesem Fall haben wir es übrigens mit einer situativen Einschränkung des Geltungsbereichs zu tun, gekennzeichnet durch das Signalwort „sofern“.

So, das war‘s erst einmal von uns. Wir hoffen wir haben dir dabei geholfen, die verschiedenen Bestandteile einiges Arguments in Zukunft leichter zu identifizieren. Du kannst dir dieses Tutorial bei den Übungen jederzeit noch einmal anschauen. Wir wünschen dir jetzt viel Erfolg bei der praktischen Phase des Argumentstrukturtrainings!

Tutorial – Plausibility Training

Herzlich Willkommen zu unserem Video-Tutorial, in dem wir dir zeigen möchten, wie man einige typische Argumentationsfehler erkennt.

In den Sozialwissenschaften sprechen wir in der Regel nicht von formellen, logisch validen Argumenten, dessen Schlussfolgerung logisch aus den Prämissen folgt, sondern von informellen Argumenten. Gute informelle Argumente haben wahre Prämissen (oder Begründungen), die relevant für die Schlussfolgerung (oder Behauptung) sind und diese ausreichend stützen. Informelle Argumente sind plausibel, wenn die Behauptung – meist in einem statistisch definierten Rahmen – sehr wahrscheinlich aus den angeführten Begründungen folgt. Wir sprechen dann auch von *starken* Argumenten.

Bei der Beurteilung der Plausibilität eines Arguments ist es daher wichtig, sich nicht nur den Wahrheitsgehalt der Behauptung und Begründung anzuschauen, sondern auch die Relevanz und Vollständigkeit der Begründung für die Behauptung.

Im Alltag und auch in der Wissenschaft begegnen wir häufig nicht nur starken Argumenten, sondern auch solchen, die die eben genannten Kriterien nicht erfüllen. Das erste Beispiel, das wir uns in diesem Zusammenhang anschauen wollen, ist der Zirkelschluss. Ein **Zirkelschluss** tritt immer dann auf, wenn versucht wird, die Richtigkeit einer Behauptung mithilfe einer Begründung zu stützen, die im Prinzip nichts Anderes aussagt als die Behauptung selbst. Dies fällt oft deshalb nicht sofort auf, weil Synonyme verwendet werden. Die Begründung ist jedoch nicht relevant für die Behauptung, da sie keine neuen Informationen enthält. Ein typisches Beispiel für einen Zirkelschluss wäre der Satz:

„*Kaffee wirkt anregend, weil er eine aufputschende Wirkung hat*." (unplausibel)

In diesem Satz wurden die Wörter „anregend" und „aufputschend" als Synonyme verwendet. Somit liefert die Begründung keine neuen Informationen für die Behauptung und ist daher auch nicht relevant. Wir wissen immer noch nicht, warum uns Kaffee aufmuntert. Plausibel wäre z.B. gewesen, zu sagen:

„*Kaffee wirkt anregend, weil er Koffein enthält"*. (plausibel)

Hier wird eine relevante Begründung angeführt, die Informationen enthält, die noch nicht in der Schlussfolgerung enthalten sind.

Schauen wir uns nun einen klassischen **Fehlschluss** an. Es gibt eine ganze Reihe von Fehlschlüssen, bei denen die Behauptung manchmal im direkten Widerspruch zur Begründung steht oder nicht daraus abgeleitet werden kann, weil die Begründung die Behauptung nicht ausreichend stützt. Ein typisches Beispiel für einen Fehlschluss ist zum Beispiel, wenn aus einer Korrelation Kausalität abgeleitet wird, was – wie ihr sicher schon im Studium gelernt habt – nicht zulässig ist. Beispielsweise wird in der Öffentlichkeit häufig vor Cannabis als Einstiegsdroge gewarnt, da die meisten Heroinnutzer auch Cannabis konsumieren, wie in diesem Satz:

„Es wurde ein Zusammenhang zwischen dem Konsum von Cannabis und dem Konsum von Heroin festgestellt. Deswegen werden die heutigen Cannabisnutzer wohl morgen auch Heroinnutzer sein." (unplausibel)

Unabhängig davon, wie man zu politisch relevanten Fragen zu diesem Thema steht, folgt aus der angeführten Begründung nicht, dass Cannabiskonsum zu Heroinkonsum führt, sondern es gibt lediglich eine Korrelation zwischen dem Konsum beider Drogen. Eine relevante Information, nämlich, dass nur eine kleine Gruppe der Bevölkerung Heroin (und womöglich auch Cannabis) konsumiert, während eine sehr viel größere Gruppe Cannabis (aber kein Heroin) konsumiert, fehlt zudem. Die angeführte Begründung ist somit nicht relevant und auch nicht ausreichend für die Behauptung. Plausibel wäre z.B. dieser Satz gewesen:

„Es wurde ein Zusammenhang zwischen dem Konsum von Cannabis und dem Konsum von Heroin festgestellt. Die meisten Cannabisnutzer konsumieren kein Heroin, aber viele Heroinnutzer konsumieren auch Cannabis. Es gibt daher vermutlich einen Zusammenhang zwischen beiden Drogen innerhalb der Gruppe der Heroinnutzer.“(plausibel)

Ein weiterer gängiger Argumentationsfehler ist die **Übergeneralisierung**. Eine Übergeneralisierung liegt dann vor, wenn Schlussfolgerungen voreilig, zu breit oder weitreichend, oder mit zu großer Sicherheit getroffen werden. Ein Beispiel für eine Übergeneralisierung wäre z.B. der Satz:

„Im letzten Jahr war ich zwei Wochen in Manchester. Bis auf einen Tag hat es während meines Urlaubs jeden Tag geregnet. England ist wirklich ein verregnetes Land!“ (unplausibel)

Wenn wir uns dieses Beispiel anschauen, dann fallen gleich mehrere Übergeneralisierungen auf. Zum einen wird von Beobachtungen, die in Manchester gemacht wurden, auf ganz England geschlossen. Das Wetter kann jedoch innerhalb Englands sehr unterschiedlich sein. So regnet es in Manchester z.B. viel häufiger als in London oder Brighton. Des Weiteren wird von der Beobachtung, dass es in Manchester zwei Wochen lang geregnet hat, die verallgemeinerte Aussage, dass England ein verregnetes Land sei, mit sehr viel Sicherheit getroffen. Mehr Beobachtungen sind nötig und es kann auch gut sein, dass ich einfach mit dem Wetter Pech hatte. Besser wäre der folgende Satz gewesen:

„Im letzten Jahr war ich zwei Wochen in Manchester. Bis auf einen Tag hat es während meines Urlaubs jeden Tag geregnet. In Manchester regnet es vermutlich relativ häufig!“ (plausibel)

In der sozialwissenschaftlichen Forschung treten Übergeneralisierungen übrigens oft dann auf, wenn Schlussfolgerungen, die auf einer besonderen Stichprobe mit bestimmten Merkmalen basieren, auf die allgemeine Population angewendet werden. Ein klassisches Beispiel ist die Teilnahme von Psychologiestudierenden an psychologischen Studien. Eigentlich können wir aus solchen Studien keine Schlussfolgerungen auf die gesamte Bevölkerung ableiten, sondern nur auf die Population der Psychologiestudierenden.

Ein weiterer Fehler, der häufig in der Literatur zu finden ist, sind **falsche Beispiele**. Um eine Behauptung zu stützen, werden oft Beispiele angeführt. Wenn diese unpassend sind, hat dies jedoch auch Auswirkungen auf die Plausibilität der Schlussfolgerung. Schauen wir uns diesen Satz an:

„Beim Autofahren laufen nach einiger Übung viele Bewegungen (z.B. kauen, schlucken) automatisiert ab.“ (unplausibel)

Die genannten Beispiele sind für die Behauptung, dass viele Bewegungsläufe beim Autofahren automatisiert ablaufen, nicht relevant und schwächen daher die Behauptung. Ein plausibles Beispiel wäre der folgende Satz:

„Beim Autofahren laufen nach einiger Übung viele Bewegungen (z.B. das Schalten und Beschleunigen) automatisiert ab“ (plausibel)

Die Beispiele „schalten“ und „beschleunigen“ haben einen direkten Bezug zum Autofahren.

Zum Schluss schauen wir uns noch eine **falsche Dichotomie** an, die immer dann auftritt, wenn ein scheinbarer Gegensatz suggeriert wird, der jedoch eigentlich keiner ist, weil Informationen überlappen, oder es noch weitere Möglichkeiten gibt, die nicht genannt werden. Ein Beispiel:

„Man sollte lieber kohlenhydratarme Lebensmittel statt Lebensmittel mit wenig Zucker essen“. (unplausibel)

Das Wort „statt“ impliziert hier, dass kohlenhydratarme Lebensmittel und Lebensmittel mit wenig Zucker etwas Unterschiedliches sind. Zucker ist jedoch auch eine Form von Kohlenhydraten. Hier ein plausibleres Beispiel:

„Man sollte lieber kohlenhydratarme Lebensmittel statt Lebensmittel mit viel Zucker essen.“ (plausibel)

So, das war‘s erst einmal von uns. Wir hoffen, dass wir dir dabei geholfen haben, Argumentationsfehler in Zukunft besser zu erkennen und sie von plausiblen Argumenten zu unterscheiden. Du kannst dir dieses Tutorial bei den Übungen jederzeit noch einmal anschauen. Wir wünschen dir jetzt viel Erfolg bei der praktischen Phase des Argumentationstrainings!

Practice Items – Argument Structure Training

No.	Test Item	Components	Complexity	Argument Type	Words
1	Die Einbeziehung von Angehörigen kann den Therapieerfolg bei Kindern und Jugendlichen mit Zwangsstörungen verbessern [*claim*]. Familienmitglieder laufen ansonsten oft Gefahr, sich an Zwangsrituale und Zwangsgedanken anzupassen oder in diese eingebunden zu werden [*datum*]. Voraussetzung ist jedoch, dass sich die Familienmitglieder auf die Therapie einlassen [*rebuttal*].	3	Low	Claim-First	42
2	Kinder von sensitiven Müttern reagieren stärker negativ emotional als Kinder weniger sensitiver Mütter [*claim*]. Die Bindungstheorie liefert hierzu theoretische Grundlagen. Sensitive Mütter geben ihren Kindern die Erfahrung, dass es angemessen ist, Unwohlsein auszudrücken und, dass die Unterstützung der Eltern aktiv eingefordert werden kann [*datum*]. Dieses Verhalten führt dazu, dass Kinder darin bestärkt werden, stärker negativ emotional zu reagieren [*warrant*]. Dies bestätigt auch eine Vielzahl von entwicklungspsychologischen Studien [*backing*]. Dies gilt allerdings nur für Kinder mit unsicher vermeidender Bindung [*rebuttal*].	5	High	Claim-First	74
3	Forscher führten eine Meta-Analyse zu exekutiven Funktionen bei ADHS durch. Alle sechs untersuchten Studien zur motorischen Inhibition wiesen eine geminderte Verhaltenshemmung bei ADHS im Vergleich zu gesunden Kindern und Jugendlichen nach [*datum*]. Eine geminderte Verhaltenshemmung wird in der aktuellen Forschung als mögliche Ursache für Lern- und Leistungsschwächen, sowie für diverse soziale Probleme diskutiert [*warrant*]. Beispielsweise ist hier der Zusammenhang mit aggressivem Verhalten auf dem Schulhof zu nennen [*backing*]. Zwar deuten die Effektstärken darauf hin, dass ein inhibitorisches, exekutives Defizit nicht bei allen Kindern mit ADHS besteht [*rebuttal*]. Die genannten laborexperimentellen Studien sprechen aber dafür, dass vielen Kindern mit ADHS durch die Gabe von Inhibitionshemmern der Alltag in der Schule	5	High	Reason-First	108

	erleichtert werden könnte [*claim*].				
4	Eine Konfrontation mit einem unerwarteten und frustrierenden Ereignis löst nicht immer und nicht bei jedem eine nachhaltige Traumatisierung aus [*claim*]. Untersuchungen zeigen beispielsweise, dass ein „Sich Aufgeben“ deutlich mit dem Risiko zusammenhängt, eine chronische posttraumatische Belastungsstörung zu entwickeln, während eine autonome Geisteshaltung vor einer Traumatisierung zu bewahren scheint [*datum*]. Die konkreten Mechanismen dieses Zusammenhangs sind bislang jedoch nicht bekannt [*rebuttal*].	3	Low	Claim-First	57
5	Zwei Metaanalysen zeigen, dass Einstellungen gegenüber älteren Personen insgesamt negativer sind als gegenüber jüngeren Personen [*datum*]: Eine negative Einstellung gegenüber einer Person kann sich negativ auf deren Beurteilung im Bewerbungsprozess auswirken [*warrant*]. Dies wurde in einer umfassenden Studie belegt [*backing*]. Ältere Arbeitskräfte werden somit gegenüber jüngeren systematisch benachteiligt [*claim*]. Aber nicht in jeder Einzelstudie fielen die Einstellungen gegenüber Älteren negativer aus als gegenüber Jüngeren, was einen Einfluss von Moderatorvariablen vermuten lässt [*rebuttal*].	3	Low	Claim-First	69
6	In einer Studie konnte gezeigt werden, dass Unzufriedenheit im Studium durch studienbegleitende Erwerbstätigkeit ausgelöst werden kann [*datum*]. Zufriedenheit gilt als wichtiger Faktor bei der Aufrechterhaltung einer langfristigen Tätigkeit [*warrant*]. So fanden Forscher einen deutlichen Einfluss von Unzufriedenheit auf das Abbruchsrisiko eines Studiums [*backing*]. Studienbegleitende Erwerbstätigkeit scheint somit das Abbruchrisiko zu erhöhen [*claim*]. Jedoch spielt der Umfang der studienbegleitenden Erwerbstätigkeit eine wichtige Rolle [*rebuttal*].	5	High	Reason-First	51
7	Beim komplexen Problemlösen kann ein Zustand innerer Kapitulation entstehen [*claim*], sofern es zu einer Überforderung kommt [*rebuttal*]. Eine Untersuchung wurde an einer Stichprobe von 169 Personen durchgeführt.	5	High	Claim-First	64

	Die Teilnehmenden sollten an einem Spiel teilnehmen, bei dem die Kontrollierbarkeit so manipuliert wurde, dass es zu einem Gefühl der Überforderung kam. Es zeigte sich eine signifikant niedrigere Leistung, wenn das Spiel schwer kontrollierbar war, im Vergleich zum gut kontrollierbaren Spiel [*datum*].				
8	Fremde können dadurch, dass sie weniger involviert sind, bestimmte Emotionen besser wahrnehmen [*datum*]. Eine zu starke Involvierung kann bewirken, dass man seine Emotionen nicht mehr objektiv beurteilen kann [*warrant*]. Man ist gewissermaßen in ihnen „gefangen“ [*backing*]. Selbstbeurteilungsinventare sollten deshalb bei der Beurteilung der Emotionsregulation durch Fremdbeurteilungen ergänzt werden [*claim*]. Dies gilt vor allem für solche Emotionen, die nach außen gut sichtbar sind [*rebuttal*].	5	High	Reason-First	58
9	Eine häufige Wiederholung einzelner Wissensinhalte ist Grundvoraussetzung für eine langfristige Behaltensleistung [*claim*]. Sowohl für instruiertes als auch für beiläufiges Lernen, aber nicht unbedingt für assoziatives Lernen [*rebuttal*], gilt nämlich, dass die Verbindung zwischen zwei Nervenzellen umso stärker wird, je häufiger zwei (oder mehr) Nervenzellen synchron aktiviert werden (so genanntes „Hebb’sches Lernen“) [*datum*]. Wie entsprechende Gedächtnistests zeigen [*backing*], sind vor allem starke Verbindungen für einen schnellen und zuverlässigen Abruf von Informationen und damit auch für eine langfristige Behaltensleistung grundlegend [*warrant*].	5	High	Claim-First	74
10	Der Beitrag gibt eine zusammenfassende Übersicht über verschiedene Faktoren, die den biografischen Verlauf von Drogenabhängigkeit und delinquentem Verhalten beeinflussen können. Er betont dabei den engen Zusammenhang zwischen illegalem Drogenkonsum und Straffälligkeit, der v. a. darauf basiert, dass ein solcher Konsum gehäuft in kriminalitätsbelasteten Milieus stattfindet [*datum*]. Das Milieu, in dem sich Drogenabhängige im Anschluss an eine Therapie bewegen, beeinflusst in	5	High	Reason-First	111

	hohem Ausmaß den langfristigen Erfolg der Therapie [*warrant*]. So erhöht der Kontakt zu kriminalitätsbelasteten Milieus nach einer Therapie die Wahrscheinlichkeit für einen Rückfall um etwa 80% (Braun, 2012) [*backing*]. Es gibt somit Anlass zur Sorge, dass Drogenabhängige im Anschluss an eine Therapie wieder in die Straffälligkeit abrutschen [*claim*], sofern sie das ursprüngliche Milieu nicht wechseln [*rebuttal*].				
11	Forscher fanden, dass die Reaktionszeiten beim Bearbeiten einer Stroop-Aufgabe für gerechtigkeitsbezogene Wörter größer waren als für bedeutungslose Wörter, wenn die Versuchspersonen zuvor eine ungerechte Situation beobachteten [*datum*]. Hoch Ungerechtigkeitssensible werden, wenn sie ein ungerechtes Ereignis mit ansehen müssen, stärker in eine negative Stimmung versetzt als geringer Ungerechtigkeitssensible [*claim*]. Eine negative Stimmung äußert sich in der Regel durch eine verlangsamte Reaktion [*warrant*], da negative Gedanken das Arbeitsgedächtnis blockieren [*backing*]. Die Befunde konnten allerdings nicht repliziert werden [*rebuttal*].	5	High	Reason-First	71
12	Videoanalysen sollten im Trainingsbereich zunehmend eingesetzt werden [*claim*]. Dafür sprechen die Ergebnisse einer Studie, in der sich die Leistungen von Sportler(inne)n um bis zu 10% steigern ließen, wenn sie ihre Trainingseinheiten regelmäßig anhand von Videos analysierten [*datum*]. Im Rahmen solcher Videoanalysen entsteht bei Sportler(inne)n selbstbezogenes Wissen, das im Training produktiv genutzt werden kann [*warrant*]. Dies legt auch eine Metaanalyse mit 2000 Sportler(inne)n verschiedenster Disziplinen nahe [*backing*]. Allerdings ist der Einsatz von Videoanalysen nur sinnvoll, wenn sie von eigens dafür geschulten Trainern durchgeführt werden [*rebuttal*].	5	High	Claim-First	81

Practice Items – Plausibility Training

No.	Test Item	Plausibility	Argumentation Fallacy
1	Die Alltagssprache kennt eine Fülle von Begriffen, mit denen wir zum Ausdruck bringen, welche Beziehung wir zu anderen Menschen haben und welche Gefühle sie in uns hervorrufen.	Plausible	
2	Wir verfügen über ein komplexes und fein abgestuftes Repertoire sprachlicher Verhaltens- und Reaktionsformen (z.B. Mimik, Formen des Blickkontakts, Gestik), mit denen wir versuchen, interpersonale Kontakte zu initiieren, zu regulieren und zu kontrollieren.	Implausible	Wrong Example
3	Interpersonale Attraktion liegt dann vor, wenn zwischen Personen eine Tendenz zur Annäherung beobachtbar ist und wenn die Eigenschaften der beteiligten Personen einen gegenseitigen Belohnungswert besitzen.	Plausible	
4	Zu unterscheiden sind verschiedene Formen der sozialen Attraktion entsprechend der Qualität der bestehenden Beziehung: (1) Flüchtige und kurzzeitige Beziehungen zwischen Personen, die bei den Partnern einen positiven Eindruck hinterlassen, (2) Freundschaftsbeziehungen und (3) Liebesbeziehungen.	Plausible	
5	In der sozialpsychologischen Forschung gibt es inzwischen die häufig zu beobachtende Tendenz, in den Sozialbeziehungen von der Analyse zweiseitiger Beziehungen zur Analyse von Wechselwirkungsvorgängen überzugehen.	Implausible	Circular Reasoning
6	Bei der Entstehung von interpersonaler Attraktion lassen sich die folgenden zentralen Determinanten für soziale Attraktion bestimmen: räumliche Nähe, Ähnlichkeit von Werten, Einstellungen und Persönlichkeitseigenschaften sowie das physische Erscheinungsbild.	Plausible	

7	Die räumliche Nähe hat sich als ein wichtiges Merkmal für das Zustandekommen von Annäherung zwischen Menschen erwiesen. So fand bereits Festinger, dass sich Studenten in einem Studentenwohnheim vor allem zu Personen hingezogen fühlen, die in demselben Wohnheim und auf derselben Etage wohnen.	Plausible	
8	Wenn sich in Studentenwohnheimen Bewohner einer Etage oder eines Wohnblocks besonders attraktiv finden, so kann das sehr wohl damit zusammenhängen, dass sie sich häufig sehen, Interaktionsmöglichkeiten also ohne viel Anstrengung und Aufwand mühelos möglich sind, doch könnten noch weitere und vielleicht wirksamere Faktoren hinzukommen.	Plausible	
9	Es ist höchst wahrscheinlich, dass sich in Studentenwohnheimen Personen finden, die ein hohes Maß an sozialer und personaler Ähnlichkeit, bedingt durch Geschlecht, Alter, sozioökonomischen Status, Herkunftsfamilie, Religion, politische Einstellung usw. besitzen, wodurch die psychische und soziale Distanz möglicherweise verringert wird.	Plausible	
10	In zahlreichen experimentellen Studien, wurde gezeigt, dass beim Vorliegen von Ähnlichkeit die Attraktion erhöht ist. In einer Studie wurden beispielsweise Studenten, die einander nicht kannten, vor ihrem Eintreffen am Studienort gebeten, einen Einstellungs- und Werte-Fragebogen auszufüllen. Während des Zusammenlebens dieser Studenten im Studentenwohnheim wurden mehrfach soziometrische Befragungsdaten erhoben, aus denen sich der Grad gegenseitiger Sympathie und Freundschaft ermitteln ließ. Einstellungs- und Wertähnlichkeit korrelierten deutlich mit Attraktivitätseinschätzungen, was zeigt, dass die interpersonale Attraktion mit Sicherheit davon abhängt, wie ähnlich uns eine andere Person ist.	Implausible	Overgeneralisation

11	Eine weitere, aus vielen alltäglichen Erfahrungen gut bekannte Determinante interpersonaler Attraktion ist das physische Erscheinungsbild. Untersuchungen haben gezeigt, dass der Zusammenhang zwischen physischer und persönlicher Attraktivität nicht linear, sondern umgekehrt u-förmig verläuft, d.h. sowohl wenig attraktive als auch hoch attraktive Personen werden nicht besonders bevorzugt, wohl aber Personen mit einem mittleren bis hohen Attraktivitätsgrad.	Plausible	
12	Interaktionspartner fühlen sich häufig von sehr attraktiven Personen zurückgewiesen, da sie Unterlegenheitsgefühle entwickeln. Das Risiko, Unterlegenheitsgefühle zu entwickeln, ist bei Personen, die einen ähnlichen Grad an physischer Attraktivität aufweisen, häufig geringer ausgeprägt.	Plausible	
13	Zur Erklärung der hier berichteten Befunde über die Entstehungsbedingungen und Determinanten interpersonaler Attraktion wurden im Wesentlichen vier in der Sozialpsychologie weit verbreitete Theorien angewandt, die Balance-Theorie, die Verstärkungstheorie, die Austauschtheorie und die Theorie der distributiven Gerechtigkeit.	Plausible	
14	Die Balance-Theorie beschreibt, dass beide Interaktionspartner füreinander an Attraktivität gewinnen, wenn sie in der Beurteilung bestimmter Personen, Objekte und Ereignisse übereinstimmen, weil sich die interpersonale Anziehung erhöht.	Implausible	Circular Reasoning
15	Nach dem konditionierungstheoretischen Ansatz lösen am Interaktionspartner wahrgenommene, positiv bewertete Persönlichkeitseigenschaften, Ähnlichkeiten von Werten und Einstellungen, sowie wahrgenommene situative Bedingungen angenehme Gefühle aus, die entsprechend dem Prinzip des Assoziativen Lernens so mit dem Interaktionspartner verbunden werden, dass in der Folgezeit bereits die Anwesenheit des Partners diese Gefühle hervorruft.	Plausible	

16	Entsprechend dem austauschtheoretischen Erklärungsansatz bewerten zwei Partner ihre Interaktionen hinsichtlich der für sie entstehenden Kosten und Erträge. Demnach wird Person A für eine Person B umso attraktiver, je höher der Ertrag abzüglich der investierten Kosten aus der Interaktion für Person B ist.	Plausible	
17	Nach der Theorie der distributiven Gerechtigkeit ist die interpersonale Attraktion eines Partners umso größer, je gerechter Kosten und Gewinne in der Interaktionsbeziehung verteilt sind.	Plausible	
18	Die verschiedenen Theorien betonen jeweils gegensätzliche Aspekte des Interaktionsprozesses, wie emotionale versus gefühlsbetonte Faktoren, und sind deshalb nicht in der Lage, eine befriedigende Erklärung für die Entstehung von interpersonaler Attraktion zu liefern.	Implausible	False Dichotomy
19	Untersuchungen über die Wirkung von Freundschaftsbeziehungen auf die interpersonale Attraktion und das Verhalten der Partner zueinander legen nahe, dass Freunde, im Gegensatz zu Bekannten, anders miteinander umgehen.	Plausible	
20	Freunden wird, im Unterschied zu Bekannten, häufig mehr Verantwortung für gute und weniger Verantwortung für schlechte Taten zugeschrieben.	Plausible	
21	Freunde reden untereinander oft mehr über Gedanken und Gefühle und zeichnen sich so durch eine geringere emotionale Offenheit aus.	Implausible	False Conclusion
22	Bei Belohnungsaufteilungen unter Freunden geht es meist gerechter zu als bei Aufteilungen unter Bekannten.	Plausible	
23	Bei Vorliegen einer hohen Arbeitsmotivation sind Freundschaften oft leistungsförderlich. Besteht jedoch eine geringe Arbeitsmotivation und eine Divergenz hinsichtlich der Leistungsziele, so behindern Freundschaften immer die Gruppenleistung.	Implausible	Overgeneralisation

24	Der Grad der Ich-Beteiligung eines Individuums in einer Freundschaft hängt davon ab, inwieweit es sich selbst zur Aufnahme dieser Beziehung ermuntert hat und wie hoch es den Grad der eigenen Wahlfreiheit für diese Beziehung bewertet.	Plausible	
25	Weiterhin spielt die Einschätzung des eigenen Selbstwertgefühls bei der Auswahl der Partner und der Entwicklung einer Freundschaftsbeziehung eine wichtige Rolle.	Plausible	
26	Bisher konnte allerdings nicht eindeutig geklärt werden, ob sich Personen mit einer größeren Diskrepanz zwischen tatsächlichem und idealen Selbst eher Freunde aussuchen, die dem idealen Selbst entsprechen oder solche, die im Sinne der Ähnlichkeitshypothese eher dem tatsächlichen Selbst nahekommen.	Plausible	
27	Nach Levinger lässt sich die die Entwicklung von Liebesbeziehungen in drei zeitlich aufeinander folgende Phasen unterteilen. In der ersten Phase ist es notwendig, dass die Partner aufeinander aufmerksam werden.	Plausible	
28	Ist die Phase des Aufeinander-Aufmerksam-Werdens abgeschlossen, so beginnt die Phase des oberflächlichen Kontaktes, die durch den Faktor „physische Attraktivität“ bestimmt ist.	Plausible	
29	Finden sich die beiden Personen gegenseitig aufgrund äußerer, meist körperlicher Erscheinungsmerkmale (z.B. das Prestige des Berufs) attraktiv, so erhöht sich die Kontaktbereitschaft.	Implausible	Wrong Example
30	Wahrgenommene Ähnlichkeit bezogen auf Persönlichkeitseigenschaften, Werte und Einstellungen können zu einer Intensivierung der Beziehung führen.	Plausible	
31	Da eine Übereinstimmung von Werten und Einstellungen eine bedeutende Rolle für die Aufrechterhaltung einer Beziehung spielt, führt eine stark ausgeprägte Einstellungsdifferenz eher zum Erhalt einer Beziehung.	Implausible	False Conclusion

32	In der Phase des oberflächlichen Kontaktes entwickeln die beiden Partner Schemata voneinander und von ihrer interaktiven Beziehung, die in Verbindung mit dem Auftreten romantischer Liebesgefühle und idealisierter Partnervorstellungen den weiteren Interaktionsverlauf bestimmen	Plausible	
33	In der anschließenden Phase der Gemeinsamkeit nehmen die Interdependenzgefühle der Partner untereinander an Intensität zu und man lernt in zunehmendem Maße, sich auf den Partner zu verlassen.	Plausible	
34	Obwohl die Vorteile des Attraktionskonzepts, nämlich Einfachheit und empirische Brauchbarkeit zur Unterscheidung von Einflussfaktoren, wie dem Bedürfnis nach sozialem Anschluss, nach sozialer Zuwendung und nach sozialem Vergleich, nicht übersehen werden können, so hat die Forschung doch zu widersprüchlichen Befunden der Erklärung der Zusammenhänge zwischen den verschiedenen Determinanten geführt.	Plausible	
35	Ein wichtiger Grund für das Zustandekommen widersprüchlicher Befunde ist die unpräzise Bestimmung des Attraktionskonstrukts. Durch die unpräzise Bestimmung des Attraktionskonstrukts sind außerhalb des Individuums liegende, aber auch externale Determinanten in ihrem Einfluss auf die Unterstützung und Behinderung interaktiver Handlungsformen zu wenig beachtet worden.	Implausible	False Dichotomy
36	Allerdings sind es gerade solche externalen Faktoren, die sozialen Kontakt, Anschluss, Anpassung, Vergleiche usw. erzwingen und in deren Folge erst die bisher behandelten Prozesse der interpersonalen Attraktion handlungswirksam werden.	Plausible	

Overview Training Environment

Übersicht Wie möchtest Du fortfahren? (Wähle aus, indem du auf einen der gelben Buttons klickst.)

	**Input**	Hier greifst du auf den Input zurück.
	**Check**	Hier siehst du, was du selbst während des Lernens notiert hast.
	**Tutorial**	Hier kannst du nochmal das Video-Tutorial sehen.
	**Feedback**	Hier erhältst du Feedback zu deinem bisherigen Fortschritt.

	**Übung**	Mit dem Üben fortfahren.

Erklärung zum Eigenanteil

Erklärung über den Eigenanteil an den veröffentlichten oder zur Veröffentlichung eingereichten wissenschaftlichen Schriften.

1. Allgemeine Angaben

Sarah von der Mühlen, Department of Psychology, University of Würzburg.

Titel Dissertation: „Fostering Students' Epistemic Competences when Dealing with Scientific Literature" ["Die Förderung epistemischer Kompetenzen von Studierenden im Umgang mit wissenschaftlicher Literatur"].

2. Anschriften der jeweiligen Mitautor(inn)en

tobias.richter@uni-wuerzburg.de

seb.schmid@paedagogik.uni-regensburg.de

kirsten.berthold@uni-bielefeld.de

elisabeth_marie.schmidt@uni-bielefeld.de

3. Liste der Publikationen

Publikation: von der Mühlen, S., Richter, T., Schmid, S., & Berthold, K. (2017). How to Improve Argumentation Comprehension in University Students: Experimental Tests of Two Training Approaches. *Manuscript submitted for publication.*					
Beteiligt an	**Autoren-Initialen,** Verantwortlichkeit abnehmend von links nach rechts				
Planung der Untersuchungen	SvdM (65%)	TR(15%)	SS(10%)	KB (10%)	
Datenerhebung	SvdM (100%)				
Daten-Analyse und Interpretation	SvdM (75%)	TR(25%)			
Schreiben des Manuskripts	SvdM (80%)	TR(10%)	SS (5%)	KB (5%)	

Erläuterung:

Die empirischen Untersuchungen basieren auf Daten, die im Rahmen des durch das Bundesministerium für Bildung und Forschung geförderten Verbundprojekts „Kompetenzen

Studierender im Umgang mit Wissenschaftlicher Originalliteratur“ (Förderkennzeichen: 01PK11017A, 01PK11017B) erhoben wurden. Die verwendeten Interventionsansätze zur Kompetenzförderung bei Studierenden wurden durch Prof. Dr. Tobias Richter, Prof. Dr. Kirsten Berthold und Dr. Sebastian Schmid zu etwa gleichen Anteilen (ca. 33%) konzeptualisiert und durch Sarah von der Mühlen (ca. 90%) und Prof. Dr. Tobias Richter (ca. 10%) (weiter)entwickelt. Die Datenerhebung wurde durch Sarah von der Mühlen durchgeführt. Die Entwicklung der Konzeption der Manuskripte, die Literaturrecherche, die Auswertung der Daten, die Diskussion der Ergebnisse, die Beweisführung sowie das Niederschreiben der Manuskripte wurde überwiegend durch die Erstautorin Sarah von der Mühlen (ca. 80%) und Prof. Dr. Tobias Richter (ca. 10%) durchgeführt. Die Koautor(inn)en Prof. Dr. Kirsten Berthold und Dr. Sebastian Schmid haben die Manuskripte vor der Einreichung bei der jeweiligen Fachzeitschrift überprüft und kommentiert (jeweils ca. 5%).

Publikation: von der Mühlen, S., Richter, T., Schmid, S., Berthold, K., & Schmidt, E.M. (2016). The use of source-related strategies in evaluating multiple psychology texts: A student-scientist comparison. *Reading and Writing, 8*, 1677-1698.					
Beteiligt an	**Autoren-Initialen,** Verantwortlichkeit abnehmend von links nach rechts				
Planung / Entwicklung der Untersuchungen	SvdM (60%)	TR (15%)	SS (10%)	KB (10%)	EMS (5%)
Datenerhebung	SvdM (100%)				
Daten-Analyse und Interpretation	SvdM (75%)	TR (25%)			
Schreiben des Manuskripts	SvdM (80%)	TR (10%)	SS (5%)	KB (2.5%)	EMS (2.5%)

Erläuterung:

Die empirischen Untersuchungen basieren ebenfalls auf Daten des oben genannten Verbundprojekts. Die in dieser Studie verwendeten Tests zur Kompetenzerfassung bei Studierenden wurden durch Prof. Dr. Tobias Richter, Prof. Dr. Kirsten Berthold und Dr. Sebastian Schmid zu etwa gleichen Anteilen (ca. 33%) konzeptualisiert und vorwiegend

durch Sarah von der Mühlen (ca. 85%), Prof. Dr. Tobias Richter (ca. 10%) und Elisabeth Marie Schmidt (ca. 5%) (weiter)entwickelt. Die Datenerhebung wurde durch Sarah von der Mühlen durchgeführt. Die Entwicklung der Konzeption der Manuskripte, die Literaturrecherche, die Auswertung der Daten, die Diskussion der Ergebnisse, die Beweisführung sowie das Niederschreiben der Manuskripte wurde überwiegend durch die Erstautorin Sarah von der Mühlen (ca. 80%) und Prof. Dr. Tobias Richter (ca. 10%) durchgeführt. Die Koautor(inn)en Prof. Dr. Kirsten Berthold, Dr. Sebastian Schmid und Elisabeth Marie Schmidt haben die Manuskripte vor der Einreichung bei der jeweiligen Fachzeitschrift überprüft und kommentiert (jeweils ca. 2.5-5%).

Publikation: von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E.M., & Berthold, K. (2016). Judging the plausibility of arguments in scientific texts: A student-scientist comparison. *Thinking & Reasoning, 22,* 221-246.					
Beteiligt an	**Autoren-Initialen,** Verantwortlichkeit abnehmend von links nach rechts				
Planung der Untersuchungen	SvdM (60%)	TR (15%)	SS (10%)	KB (10%)	EMS (5%)
Datenerhebung	SvdM (100%)				
Daten-Analyse und Interpretation	SvdM (75%)	TR (25%)			
Schreiben des Manuskripts	SvdM (80%)	TR (12.5%)	SS (5%)	EMS (5%)	KB (2.5%)

Erläuterung:

Die empirischen Untersuchungen basieren ebenfalls auf Daten des oben genannten Verbundprojekts. Auch die in dieser Studie verwendeten Tests zur Kompetenzerfassung bei Studierenden wurden durch Prof. Dr. Tobias Richter, Prof. Dr. Kirsten Berthold und Dr. Sebastian Schmid zu etwa gleichen Anteilen (ca. 33%) konzeptualisiert und vorwiegend durch Sarah von der Mühlen, Prof. Dr. Tobias Richter (ca. 10%) und Elisabeth Marie Schmidt (ca. 5%) (weiter)entwickelt. Die Datenerhebung wurde durch Sarah von der Mühlen durchgeführt. Die Entwicklung der Konzeption der Manuskripte, die Literaturrecherche, die

Auswertung der Daten, die Diskussion der Ergebnisse, die Beweisführung sowie das Niederschreiben der Manuskripte wurde überwiegend durch die Erstautorin Sarah von der Mühlen (ca. 80%) und Prof. Dr. Tobias Richter (ca. 10%) durchgeführt. Die Koautor(inn)en Prof. Dr. Kirsten Berthold und Dr. Sebastian Schmid haben die Manuskripte vor der Einreichung bei der jeweiligen Fachzeitschrift überprüft und kommentiert (jeweils ca. 2.5-5%).

Für alle in dieser „Dissertation unter Einschluss mehrerer publizierter Manuskripte“ verwendeten Manuskripte liegen die notwendigen Genehmigungen der Verlage und Co-Autoren für die Zweitpublikation vor. Mit meiner Unterschrift bestätige ich die Kenntnisnahme und das Einverständnis meines direkten Betreuers.

Datum, Unterschrift der Antragstellerin

.....	.....	.....
Name Doktorandin	Ort, Datum	Unterschrift

Ich bestätige die von Frau Sarah von der Mühlen unter Punkt 3. abgegebene Erklärung zum Eigenanteil an den veröffentlichten oder zur Veröffentlichung eingereichten Publikationen:

.....	.....	.....
Prof. Dr. Tobias Richter	Ort, Datum	Unterschrift

.....	.....	.....
Dr. Sebastian Schmid	Ort, Datum	Unterschrift

.....	.....	.....
Prof. Dr. Kirsten Berthold	Ort, Datum	Unterschrift
.....	.....	.....
Elisabeth Marie Schmidt	Ort, Datum	Unterschrift

Eidesstattliche Versicherung und sonstige Erklärungen

gemäß § 6 PromO 2014 zum Antrag auf Zulassung zur Doktorprüfung

Frau Sarah von der Mühlen

Erklärungen

§ 6 Abs. 2 Nr. 5: Hiermit versichere ich an Eides statt,

►dass ich die Dissertation selbständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe,

►dass ich die Gelegenheit zum Promotionsvorhaben nicht kommerziell vermittelt bekommen und keine Person oder Organisation eingeschaltet habe, die gegen Entgelt Betreuer bzw. Betreuerinnen für die Anfertigung von Dissertationen sucht.

§ 6 Abs. 2 Nr. 6: Hiermit erkläre ich, dass ich die Regeln der Universität Würzburg über gute wissenschaftliche Praxis eingehalten habe.

§ 6 Abs. 2 Nr. 8: Die vorgelegte Dissertation wurde bisher bei keinem Prüfungsverfahren eingereicht; sie ist nicht identisch mit einer früher abgefassten wissenschaftlichen Arbeit, z. B. einer Magister-, Diplom-, Master, Bachelor- oder Zulassungsarbeit.

Bei der vorliegenden Arbeit handelt es sich um eine publikationsbasierte Dissertation. Die entsprechenden Kriterien der Fakultät für Humanwissenschaften wurden berücksichtigt.

....

Ort, Datum Unterschrift

Lebenslauf

Persönliche Daten

Name	Sarah von der Mühlen
Geburtsdatum	22. 09. 1986 in Herdecke
10/2017 – heute	Doktorandin an der Julius-Maximilians-Universität Würzburg (Fach Psychologie)

Erfahrungen in der Wissenschaft

11/ 2012 – 02/2016	Wissenschaftliche Mitarbeiterin an der Universität Kassel, Institut für Psychologie, Allgemeine Psychologie Tätigkeiten: o Konzeption, Durchführung und Auswertung verschiedener qualitativer und quantitativer empirischer Untersuchungen zur Erfassung und Förderung epistemischer Kompetenzen (BMBF-Projekt) o Konferenzbeiträge o Publikationen

Lehrerfahrungen

Wintersemester 2014/15	Seminar „Biologische Psychologie“ (Bachelor Psychologie)
Wintersemester 2014/15	Seminar „Biologische Psychologie“ (Bachelor Psychologie)
Sommersemester 2014	Seminar „Motivation und Emotion“ (Bachelor Psychologie)
Wintersemester 2013/14	Seminar „Biologische Psychologie“ (Bachelor Psychologie)

Ausbildung

01/2015 – 04/2015	Ausbildung am Zentrum für Positive Psychologie Berlin zur Zertifizierten Anwenderin der Positiven Psychologie (Deutschsprachiger Dachverband für Positive Psychologie e.V.)
30. 06. 2012	Abschluss: Master of Science (M.Sc.), Note: 1.2 (cum laude)
09/2011 – 06/2012	Masterstudium der Gesundheits- und Sozialpsychologie, Universität Maastricht, Maastricht, Niederlande
31. 08. 2011	Abschluss: Bachelor of Science (B.Sc.), Note: 1.7

09/2006 – 08/2011	Bachelorstudium der Psychologie (Vertiefungsrichtung: Kognitive Psychologie), Universität Maastricht, Maastricht, Niederlande
22.06. 2006	Abschluss: Abitur, Note: 1.9
08/2003 – 12/2003	Brockwood Park School, Bramdean, Hampshire, UK
08/1993 – 05/2006	Rudolf-Steiner Schule, Dortmund

Weitere Qualifikationen

Sprachkenntnisse	Deutsch (Muttersprache)
	Englisch (fließend),
	Niederländisch (fließend)
	Russisch (Grundkenntnisse)
EDV-Kenntnisse	verschiedene Betriebssysteme (Windows / Macintosh)
	MS-Office (Word, Power-Point, Excel)
	SPSS (erweiterte Kenntnisse); R, MPlus (Grundkenntnisse)
	Inquisit, E-Prime (Grundkenntnisse)

Publikationen

Münchow, H., Richter, T., von der Mühlen, S., Schmid, S., Berthold, K., & Bruns, K. (eingereicht). Verstehen von Argumenten in wissenschaftlichen Texten: Reliabilität und Validität des Argumentstrukturtests (AST). *Manuskript zur Publikation eingereicht.*

von der Mühlen, S., Richter, T., Schmid, S. & Berthold, K. (eingereicht). How to Improve Argumentation Comprehension in University Students: Experimental Tests of Two Training Approaches. *Manuskript zur Publikation eingereicht.*

von der Mühlen, S., Richter, T., Schmid, S., Berthold, K. & Schmidt, E. M. (2016). The use of source-related strategies in evaluating multiple psychology texts: A student-scientist comparison. *Reading and Writing, 8*, 1677–1698.

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M. & Berthold, K. (2016). Judging the plausibility of arguments in scientific texts: A student-scientist comparison. *Thinking & Reasoning, 22,* 221–246.

Konferenzbeiträge

Richter, T., von der Mühlen, S., Schmid, S. & Berthold, K. (2014, Oktober). *Credibility judgements and source-related strategies in multiple text comprehension: An expert- novice comparison in the domain of psychology*. Vortrag im Rahmen des Workshops Multiple Document Literacy, Valencia, Spanien.

Schmid, S., von der Mühlen, S., Schmidt, E. M., Bruns, K., Richter, T. & Berthold, K. (2014, März). Diagnostik von Kompetenzen Studierender im Umgang mit wissenschaftlicher Originalliteratur. Talk given at the 2nd GEBF Meeting (Symposium *Lesen und Schreiben – Der Umgang mit wissenschaftlicher Literatur im Hochschulkontext*), Frankfurt/Main, Deutschland.

Schmid, S., von der Mühlen, Richter, T., S., Bruns, K. & Berthold, K. (2013, Oktober). Kompetenzen Studierender im Umgang mit wissenschaftlicher Originalliteratur. In K. Trempler (Chair), *Umgang mit wissenschaftlicher Evidenz*. Vortrag im Rahmen des Kolloquiums der KoKoHs-Verbünde KOSWO, KOMPARE und E4Teach, Wuppertal, Deutschland.

Schmidt, E. M., von der Mühlen, S., Bruns, K., Schmid, S., Richter, T. & Berthold, K. (2014, April). *Useful strategies in dealing with primary scientific literature: A student-scientist comparison.* Poster-Präsentation auf dem Kongress der AERA, Philadelphia, PA, USA.

Schmidt, E. M., von der Mühlen, S., Schmidt, E. M., Bruns, K., Schmid, S., Richter, T. & Berthold, K. (2014, März). *Muss ich das denn alles lesen? Kann ich das denn alles glauben? Kompetenzen Studierender im Umgang mit psychologischer Originalliteratur.* Vortrag auf dem DGfE- Kongress, Berlin, Deutschland.

von der Mühlen, S. (2013, November). *Students' competencies when dealing with primary scientific literature – first results*. Vortrag auf dem Internationales Doktoranden-Kolloquium, Mainz, Deutschland.

von der Mühlen, S., Richter, T., Schmidt, E. M., Schmid, S., Berthold, K. & Bruns, K. (2015, September). *Plausibilitätsbeurteilungen und das Erkennen von Argumentationsfehlern bei argumentativen Texten – ein Experten - Novizen - Vergleich.* Poster-Präsentation auf der Tagung der Fachgruppe Pädagogische Psychologie, Kassel, Deutschland.

von der Mühlen, S., Richter, T., Schmidt, E. M., Schmid, S. & Berthold, K. (2015, Juli). *The use of source-related strategies in reading multiple psychology texts: A student-scientist comparison.* Poster-Präsentation auf dem Kongress der STnD, Minneapolis, MN, USA.

von der Mühlen, S., Richter, T., Schmidt, E. M., Schmid, S., Berthold, K. & Bruns, K. (2014, September). *Glaubwürdigkeitseinschätzungen bei der Rezeption wissenschaftlicher Texte im Fach Psychologie: Ein Experten-Novizen-Vergleich als Teil der Entwicklung eines Instruments zur Erfassung epistemischer Kompetenzen*. Symposiumsbeitrag auf dem DGPS-Kongress, Bochum, Deutschland.

von der Mühlen, S., Richter, T., Schmidt, E. M., Schmid, S., Berthold, K. & Bruns, K. (2014, August). *Judging the plausibility of argumentative statements in scientific texts: A student-scientist comparison* Poster-Präsentation auf dem Kongress der STnD, Chicago, IL, USA.

von der Mühlen, S., Richter, T., Schmidt, E. M., Schmid, S., Berthold, K. & Bruns, K. (2014, August). *Judging the plausibility of argumentative statements in scientific texts: A student-scientist comparison.* Poster-Präsentation auf dem Kongress der EARLI- SIG2, Rotterdam, Niederlande.

....

Ort, Datum Unterschrift