Zeitschrift für Erziehungswissenschaft (2023) 26:703–725
https://doi.org/10.1007/s11618-023-01147-x

SCHWERPUNKT

Promoting students' argument comprehension and evaluation skills: Implementation of two training interventions in higher education

Hannes Münchow · Simon P. Tiffin-Richards · Lorena Fleischmann · Stephanie Pieschl · Tobias Richter

Received: 2 February 2022 / Revised: 28 September 2022 / Accepted: 25 November 2022 / Published online: 17 March 2023

Abstract The ability to comprehend and evaluate informal arguments is important for making sense of scientific texts and scientific reasoning. However, university students often lack the skills necessary to comprehend the functional structure and evaluate the structural plausibility of informal arguments. The aim of this study was to evaluate the effectiveness of two training interventions to a) improve students' argument comprehension (identification of argument structure), and to b) improve students' argument evaluation (distinguishing good vs. bad arguments). The training interventions were implemented as a voluntary online add-on to a regular university course. The study used a crossover-experimental design with a pre-test and two training phases in which participants ($N = 29$) alternated between the two training interventions. Students generally improved on the measures of scientific literacy

Data files and analysis scripts for the full analyses are available in the Open Science Framework (https://osf.io/qtg5j/). The training materials and tests are available from the authors upon request.

✉ Dr. Hannes Münchow · Dr. Simon P. Tiffin-Richards · Prof. Dr. Tobias Richter
Department of Psychology IV, Julius-Maximilians-Universität Würzburg, Röntgenring 10, 97070 Würzburg, Germany
E-Mail: hannes.muenchow@uni-wuerzburg.de

Dr. Simon P. Tiffin-Richards
E-Mail: simon.tiffin-richards@uni-wuerzburg.de

Prof. Dr. Tobias Richter
E-Mail: tobias.richter@uni-wuerzburg.de

Lorena Fleischmann
Department of Educational Psychology, Ruprecht-Karls-Universität Heidelberg, Hauptstraße 47, 69117 Heidelberg, Germany
E-Mail: lorena.fleischmann@psychologie.uni-heidelberg.de

Prof. Dr. Stephanie Pieschl
Pädagogische Psychologie, Technical University of Darmstadt, Alexanderstraße 10, 64283 Darmstadt, Germany
E-Mail: stephanie.pieschl@tu-darmstadt.de

that were practiced in each training intervention. The results provide evidence that voluntary online training interventions for components of scientific literacy can be effectively integrated into higher education settings. However, results further showed an interference effect between the training interventions, indicating that students had problems integrating the different aspects of scientific literacy targeted in the two training interventions.

Keywords Argument comprehension · Argument evaluation · Training intervention · Scientific literacy · Higher education

Förderung des Verstehens und Bewertens von Argumenten: Implementierung zweier Trainingsinterventionen in der Hochschulbildung

Zusammenfassung Die Fähigkeit, informelle Argumente zu verstehen und zu bewerten, ist wichtig, um wissenschaftliche Texte zu verstehen und wissenschaftlich zu argumentieren. Universitätsstudierende verfügen jedoch häufig noch nicht in ausreichendem Maße über die notwendigen Fähigkeiten, um die funktionale Struktur informeller Argumente zu verstehen und ihre Plausibilität zu beurteilen. Ziel der vorliegenden Studie war es, die Wirksamkeit zweier Trainingsmaßnahmen zu evaluieren, die a) das Verstehen informeller Argumente (Identifizierung der Argumentationsstruktur) und b) die Bewertung informeller Argumente (Unterscheidung zwischen guten und schlechten Argumenten) fördern. Die Trainingsmaßnahmen wurden den Studierenden als freiwilliges Online-Angebot zusätzlich zu einer regulären Universitätsveranstaltung zur Verfügung gestellt. Die Studie verwendete ein experimentelles Prä-Post-Untersuchungsdesign mit einem Vortest und zwei gekreuzten Trainingsphasen, in denen die Studierenden (N = 29) zwischen den beiden Trainingsinterventionen wechselten. Insgesamt verbesserten sich die Studierenden jeweils in der Fähigkeit, die in der jeweiligen Trainingsphase trainiert wurde. Die Ergebnisse zeigen damit, dass freiwillige Online-Trainingsangebote zur Verbesserung des Argumentverstehens und -bewertens wirksam in die Hochschullehre integriert werden können. Es fand sich jedoch auch ein Interferenzeffekt zwischen den Trainingsinterventionen, was darauf hindeutet, dass die Studierenden Probleme hatten, die trainierten Teilkompetenzen für einen kompetenten Umgang mit wissenschaftlichen Texten miteinander zu verbinden.

Schlüsselwörter Argumentverstehen · Argumentbewertung · Trainingsintervention · Wissenschaftliche Grundbildung · Hochschullehre

1 Introduction

The ability to process information presented in scientific texts is an indispensable skill for students in higher education and a central prerequisite for becoming scientifically literate (Norris and Phillips 2003). According to Britt et al. (2014), core aspects of scientific literacy include students' ability to comprehend and critically

evaluate scientific content for a specific goal. While comprehending and evaluating scientific texts can be constrained by domain-specific knowledge, certain aspects of scientific literacy are of general relevance for working with scientific content. In this study, we focus on university students' ability to comprehend and evaluate arguments typically contained in scientific discourse as domain-general aspects of scientific literacy.

Studies have shown that students usually have the basic capacity to comprehend and evaluate argumentation in scientific texts (e.g., Johnson et al. 2004; OECD 2019), especially when the structure of the arguments is clearly outlined (e.g., Chambliss 1995; Chambliss and Murphy 2002), or when students base their evaluation on judgements about the internal consistency of the arguments, that is, the completeness and relevance of reasons provided to support a claim (e.g., von der Mühlen et al. 2016). However, few students who start higher education have received any formal training in argument comprehension (Osborne 2010), and students usually have little experience in engaging with more complex and less typically structured scientific texts. This is problematic, as these skills are positively associated with academic success (e.g., Münchow et al. 2019; von der Mühlen et al. 2015, 2016). Lab-based studies have demonstrated that training interventions can effectively improve students' argument processing (e.g., von der Mühlen et al. 2018).

However, to our knowledge, little is known about the effectiveness of such interventions in non-laboratory settings in higher education. The goal of the present study was hence to evaluate the effectiveness of two training interventions targeting core aspects of scientific literacy, implemented in an ecologically valid setting. To conceptually replicate the findings from von der Mühlen et al. (2018), we therefore tested the effectiveness of a similar computer-based training intervention to improve students' ability to comprehend the functional structure of scientific arguments. To extend these findings and to address a second core aspect of scientific literacy, we further tested the effectiveness of a second training intervention designed to improve students' ability to evaluate the functional structure of scientific arguments. Our third aim was to provide evidence for the effectiveness of the training interventions in an ecologically valid setting. We therefore investigated whether the two training interventions can be implemented successfully as an online supplementary activity to a regular university course.

2 Theoretical background

2.1 Scientific Literacy in Higher Education

Scientific literacy includes not only readers' ability to acquire scientific knowledge in the form of facts or theories but also the ability to understand the meaning of the scientific contents (e.g., Norris and Phillips 2003) and to comprehend and evaluate the texts' argumentation (Britt et al. 2014). The strategies required for argument comprehension and evaluation are not, however, explicitly taught in school and are rarely represented in university curricula (Osborne 2010). Nevertheless, students in most undergraduate programs are expected to have attained the skills and know-

ledge to successfully engage with scientific literature as soon as they transition from school to university. These expectations may be unrealistic, as deficits have been found to be pervasive across different aspects of scientific reasoning in university students (Britt et al. 2014). Von der Mühlen et al. (2015, 2016), for instance, compared undergraduate students and academic professionals in terms of their ability to engage in scientific reasoning. The results revealed that professionals were far better than students at comprehending scientific arguments (operationalized as the ability to recognize functional components of scientific arguments) and evaluating the quality of scientific arguments (operationalized as the ability to distinguish good from bad arguments and to recognize argumentation fallacies in scientific arguments). Furthermore, Münchow et al. (2019) showed that students who were better at evaluating the plausibility of informal scientific arguments and detecting argumentation fallacies showed higher academic success, even after controlling for prior school achievement and verbal cognitive capability.

A possible explanation for the observation that students in higher education often struggle with reading scientific texts may be that these readers have not yet acquired necessary reading strategies due to a lack of experience with the peculiarities of scientific texts. In school texts, scientific knowledge is typically presented as an absolute or irrefutable truth. When reading such texts, school students are usually not required to critically reflect on what they have read. Scientific texts, on the other hand, contain relativizing, contradictory, or even conflicting findings or claims. Here, students face the challenge not only of comprehending the information given in the text, but also of putting it in relation to information, including contradictory claims and evidence, that was processed in texts they have previously read, and to construct a coherent mental representation of a scientific phenomenon (e.g., Perfetti et al. 1999).

Students in higher education, especially at the beginning of their studies, often evaluate the contents of scientific texts primarily based on spontaneous assessments of plausibility that are generated as a by-product of routine validation processes (epistemic monitoring, Richter et al. 2009). Students thus tend to ignore the internal consistency of arguments, especially the relevance and completeness of the reasons given (Shaw 1996). Several studies (e.g., Bazerman 1985; Berkenkotter and Huckin 1995) have investigated how science professionals read scientific texts to understand which reading strategies are particularly useful for the rational evaluation of the content of scientific texts, as opposed to an assessment based on intuition or pre-existing beliefs (see Maier and Richter 2013). Results suggest that science professionals generally use a wide repertoire of different reading strategies, depending on their personal reading goals. For example, Bazerman (1985) found that professionals in the field of physics routinely use two important groups of reading strategies to assess the merits of scientific publications. Strategies such as scanning a text for certain keywords, skipping whole sections, or evaluating the quality of a text on the basis of author characteristics (e.g., the number of previous publications), were used to judge a text's trustworthiness, to identify particular sections for a closer reading, or to find certain information quickly. To achieve a deeper understanding of a text's content, however, scientists may use deeper reading strategies (*core reading* in Bazerman 1985) that involve checking the reasoning of a text for its contents' validity

and internal consistency. Such strategies include the readers' ability to identify and assign functional components of scientific arguments, to examine the relevance and completeness of the given reasons, or to evaluate the plausibility of arguments presented. Skilled scientific professionals therefore appear to make strategic use of the structure of scientific texts to evaluate their credibility and sound reasoning (see von der Mühlen et al. 2015). Superficial and deeper reading strategies thereby serve different purposes depending on the particular reading goals. While superficial reading strategies can be of great value for making quick initial judgements about the credibility of scientific texts, deeper reading strategies are generally more important for assessing the content of a text based on comprehending and evaluating its reasoning.

To sum up, it is evident that scientific literacy skills required to comprehend and evaluate scientific arguments are associated with academic success but are not regularly taught in middle school and in higher education. Providing students with a systematic, scientifically sound training to strengthen their scientific literacy skills thus appears a desirable goal in higher education in order to support academic development.

2.2 Comprehending and Evaluating Informal Arguments in Scientific Texts

Arguments, as they are seen for the purpose of this study, are usually made to convince a person to accept a certain assertion or claim (see Galotti 1989). In formal or deductive reasoning, one or several premises form a logically derivable conclusion that is necessarily true if the premises are correct. Informal arguments, on the other hand, cannot be tested beyond reasonable doubt by formal or deductive reasoning and logical considerations (Toulmin 1958). Instead, the claim in informal arguments is more or less likely to be true, based on the quality of the supporting reasons provided (Green 1994; Voss and Means 1991). Sound informal arguments provide relevant, internally consistent evidence that is not compromised by argumentation fallacies. Central aspects of scientific literacy necessary to evaluate the argumentation of scientific texts include being able to *comprehend* and *evaluate* informal scientific arguments.

In the context of this study, *argument comprehension* describes the ability to decode the functional structure of arguments, i.e., to recognize functionally distinct argument components such as a claim or reason. Unlike formal arguments, informal arguments may contain other functional components in addition to claims and reasons. In Toulmin's (1958) model of argumentation, five functionally distinct components are differentiated: a more or less controversial *claim*, of which the reader is to be convinced, one or several *reasons* that provide theoretical, empirical, or practical evidence to support the claim (*datum/data* in Toulmin 1958), a *warrant* that describes the relevance of the reasons for supporting the claim, a *backing* that justifies the warrant, and a *rebuttal* that contains counterarguments to the claim and that provides limitations (i.e., by referring to exceptions).

However, functional argument components are not always easy to detect in texts and may have to be inferred by the reader. This is especially true for less typical components such as warrants (e.g., Chambliss 1995). In addition, the order of the argument components can influence the readers' success in recognizing different

argument components. Readers are typically confronted with informal arguments beginning with the claim followed by their supporting reason(s) (*claim-first arguments*; Britt and Larson 2003). Variations in the order of the argument components may hinder the recognition of components, for example, when arguments start with a reason (*reason-first arguments*; Britt and Larson 2003). Supporting evidence for the processing advantage of claim-first arguments was found in a study by Münchow et al. (2020b), who showed that university students were not only more accurate at recognizing functionally different argument components in claim-first compared to reason-first informal arguments, but also faster. Larson et al. (2004) also found that students were less likely to correctly identify key argument components when arguments were less typically structured. Functional argument components may be introduced by specific linguistic markers (Britt and Larson 2003). For example, reasons can be signaled by words such as 'because of' or 'therefore', whereas a rebuttal is often signaled by words such as 'on the other hand' or 'however'. Knowing about the typical structure of informal arguments and specific linguistic markers can help readers identify the passages in scientific texts that are important for its reasoning, such as the main claim(s) or relevant reason(s).

Argument evaluation describes the ability to judge whether the functionally distinct components of an argument form an internally consistent and plausible argumentation, i.e., that a given reason is relevant for supporting a claim and that no argumentation fallacies have been committed (Richter 2011; Shaw 1996). Despite the importance of an argument's functional structure, several studies have shown that students have difficulties in adequately establishing connections between functionally distinct argument components, making it hard for them to properly evaluate the quality of informal arguments (e.g., Larson et al. 2009; Münchow et al. 2019). Weak informal arguments can contain argumentation fallacies that violate the completeness and relevance of the evidence to support the claim, thus depriving the claim of its plausibility (cf. Shaw 1996). Common argumentation fallacies are the following (see Dauer 1989): *Contradictions* arise when a premise is followed by a false conclusion. A *false dichotomy* occurs when the availability of options is falsely limited. A *wrong example* is an incorrect or inappropriate example given as evidence for a specific claim. *Circular reasoning* occurs when the correctness of a premise is supported by drawing a (logical) conclusion from that very premise. *Overgeneralization* occurs when a false premature conclusion is drawn from a premise because of falsely generalizing or overstating results.

Awareness of missing or unlikely links between reasons and claims due to argumentation fallacies is important for accurately evaluating the internal consistency of informal arguments. However, the accurate evaluation of informal arguments is cognitively demanding (Münchow et al. 2019). Münchow et al. (2019) asked university students to judge the plausibility of short informal arguments consisting of a claim and a reason. They found that judging the plausibility of implausible arguments was more difficult for the students than judging the plausibility of plausible arguments. However, the evaluations of plausible and implausible arguments not only differed in terms of accuracy but also in terms of processing times. Judgments were slower when the arguments were implausible but only when the judgments were correct. If only one of these conditions was met, i.e., when the arguments were plausible or

when judgments were incorrect, reaction times were not increased. The results of Münchow et al. (2019) indicate that evaluating the content of informal arguments on the basis of their internal consistency does not occur as part of routine validation processes but requires readers to invest mental effort.

It is reasonable to assume that being able to *comprehend* and *evaluate* the functional structure of single informal arguments can also foster students' understanding of scientific texts. The rationale behind this assumption is that scientific texts as a whole often have a structure similar to prototypical informal arguments (Suppe 1998), with a claim being presented in the form of theoretically derived hypotheses regarding a specific scientific problem and evidence usually provided by empirical data. In scientific journal articles, for example, limitations and counterarguments are often described in the discussion section to delineate the framework within which the claim is valid which can be referred to warrants and rebuttals in Toulmin's (1958) model of argumentation.

In sum, comprehending and evaluating informal scientific arguments demands the readers to decode the arguments' functional structure by recognizing distinct argument components and to evaluate connections between those components to form judgements about the arguments' plausibility. Hence, these skills are not only important for argument comprehension and evaluation but also for forming a coherent mental model of the content of scientific texts (Britt and Rouet 2012).

2.3 Improving Lay Readers' Argument Comprehension and Evaluation Skills

University students, especially at the beginning of their studies, often lack effective skills to comprehend and evaluate the quality of informal arguments (e.g., Larson et al. 2004, Experiment 1). However, there is also evidence that scientific literacy skills can be improved with training interventions. Chambliss (1995), for example, found that a substantial percentage of high-school students was able to identify the structure of informal arguments when the arguments were clearly hierarchical and contained strong syntactic cues to their structure. Moreover, the presented information about the structure of the arguments helped students to comprehend the arguments and to construct more accurate mental representations of the arguments' contents. Based on these findings, Larson et al. (2004, Experiment 2) sought to teach university students the structure of more complex informal arguments and of arguments with a less typical structure, by providing them with short educational tutorials on functionally distinct argument components and instructions on how to recognize them. Results showed that students improved in their ability to recognize functionally distinct argument components, but only if their task was to comprehend the arguments. No positive effect was found when students were asked to evaluate the quality of the arguments. Von der Mühlen et al. (2016) further demonstrated that students who considered the internal consistency of the arguments, i.e., the relations between argument components for judging the quality of informal arguments, had less difficulty comprehending and evaluating argumentative statements in scientific texts.

The effectiveness of a computer-based intervention for training scientific literacy skills was demonstrated in a recent training study conducted by von der Mühlen et al.

(2018). The authors trained psychology students' ability to recognize functionally distinct components of informal arguments typically encountered when reading scientific texts in order to enhance their argument comprehension skills. Moreover, the training was intended to promote active learning by providing, for example, tasks for self-generating relevant content and exercises during and after each input block. Such learning principles enable learners to actively construct knowledge at an individual pace (Jonassen 1999) and have been found to improve comprehension and memory for text content (e.g., Chi et al. 1989; Marsh et al. 2001). The training used in the study by von der Mühlen et al. (2018) was particularly effective in fostering students' ability to recognize the functional structure of more complex arguments with a less typical structure (i.e., reason-first arguments) and to correctly identify more uncommon argument components, such as warrants. Moreover, the training was especially helpful for high-achieving students and students who already had relatively good argument evaluation skills before the training. A similar previous intervention study (Larson et al. 2009) also found students' argument evaluation skill to be improved by explicit training.

In conclusion, training interventions to improve students' argument processing that aim to promote interaction with the learning environment appear to be effective in improving students' scientific literacy skills.

3 The Present Study

The main goal of this study was to test whether accessible and easy-to-implement computer-based training interventions can promote students' argument comprehension and evaluation skills as part of their regular university courses. The aim was thus to conceptionally replicate and extend the intervention study conducted by von der Mühlen et al. (2018), which demonstrated the effectiveness of a training intervention to improve students' ability to detect and correctly assign functionally distinct components of informal scientific arguments in a laboratory setting. To this end, we conducted two computer-based training interventions: 1) an argument structure training very similar to the one used in the study by von der Mühlen et al. (2018) to promote students' argument comprehension skills, and 2) an argument judgement training to promote students' ability to accurately judge the plausibility of informal scientific arguments, i.e., their argument evaluation skills. The skills targeted in these trainings have been demonstrated to correlate positively with general cognitive capability, academic success, as well as students' epistemological beliefs in a series of studies (e.g., Münchow et al. 2019, 2020b; see also Münchow et al. 2020a) and can be seen as measurements of core aspects of scientific literacy, that is, the general ability to understand and work with scientific texts.

We hypothesized that participation in the training interventions should enhance students' ability to properly recognize and allocate functionally distinct argument components (Hypothesis 1) and to accurately evaluate the structural plausibility of such arguments based on the evaluation of the interrelations of the argument components and their internal consistency (Hypothesis 2). By integrating the training interventions into students' ongoing academic curriculum, the present study rep-

resents an attempt to transfer research-based, psychologically sound, and tested training procedures to an ecologically valid setting.

4 Method

4.1 Participants

Participation for the study was advertised in a university course in which 372 students were enrolled. Ultimately, 110 university students registered for the online study. Of these, 101 participated at measurement point 1 (T1), 54 at measurement point 2 (T2), and 42 at measurement point 3 (T3). Twenty-nine students participated at all three measurement points and were included in the analyses. The drop-out rate was thus on average 33% between measurement points, resulting in an effective sample size of approx. 30% of the initial sample. The participants in the final sample were predominantly female (65.5%, $n = 19$), 25 years old ($SD = 5.52$, *range* = 19–47), in their fifth semester at university ($M = 4.5$, $SD = 2.22$, range = 1–7), and were studying a range of sciences and humanities (Education = 6, Psychology = 7, Sociology = 6, other STEM subjects = 6, other non-STEM subjects = 3). All participants stated German to be their first language.

4.2 Procedures

The present study was conducted as part of an introductory lecture of Developmental Psychology at a medium-sized German university. The lecture covered typical topics from developmental psychology, such as cognitive, affective, and social development and the development of linguistic skills. To integrate the training interventions into the regular course schedule and to link them to the course content, lecture sessions before and after the measurement points of the study dealt with the importance of students' skills in comprehending and evaluating scientific arguments for the development of scientific thinking and reading. Participation in the training interventions was voluntary and no monetary compensation or course credit was paid. The exception were psychology students, who could earn partial course credit if, among other things, they participated in the study. As an additional incentive to participate in the study, the lecturer advertised the content of the training interventions as relevant to the final course examination. At each measurement point, the students were reminded via e-mails sent from the university's e-learning platform to participate in the study. The students provided information about their age, gender, university programme, university grade average, school grade average, and first and second language. Participants were also asked to estimate their ability (self-concept) on the two scientific literacy tests.

The training interventions were implemented in testMaker 4 (Hartweg et al. 2022), which is a free web-based software for presenting, administering, and evaluating psychological tests. The design of the trainings was similar to the training intervention used by von der Mühlen et al. (2018). In an initial theoretical input block, both trainings first conveyed conceptual knowledge about informal arguments, which

was presented with multimedia elements such as illustrated texts, audio examples, and short video tutorials. During and after the theoretical input block, the students engaged in several tasks and exercises in which they had to generate content relevant to the topic and directly apply and practice what they had learned in the theoretical input block. For example, in the Argument Structure Training, students were asked to explain the different functions of each sentence in short informal arguments after learning about Toulmin's (1958) model of argumentation. In the Argument Judgement Training, students were asked to explain which argumentation fallacy was present after reading short implausible informal arguments. After each exercise, they received detailed feedback on whether they had correctly applied their knowledge of the functional structure of informal scientific arguments (Argument Structure Training) or of the relevance and completeness of the presented reasons to support a claim (Argument Judgement Training). Moreover, if the answers were incorrect, the students were again presented with a brief summary of the input. The last part of the trainings consisted of a series of tasks and exercises to self-check the knowledge acquired in the previous training blocks (knowledge-check block) with feedback about the correctness of the given answers. Students could repeat these tasks and exercises or return to the theoretical input block if their answers were incorrect. Each training session took about 45 to 60 minutes. There was no set time limit.

For reasons of data protection, we ensured that the students were able to anonymously participate in the online study. Each student was assigned a TAN via their matriculation number with which they could log into the system. The assignment of names to matriculation numbers was only accessible to the lecturer who had no insight into which students participated in the study. The TANs additionally served to match the students' test and training sessions across the three measurement points.

4.3 Measures

4.3.1 Argument Structure Test

The Argument Structure Test (Münchow et al. 2020b) assesses individuals' ability to correctly identify the functional structure of informal arguments as a measure of argument comprehension skill. It comprises two parallel versions, both consisting of four arguments each with three to five sentences. All sentences can be classified according to Toulmin's (1958) component model, which distinguishes between claim, reason, warrant, backing, and rebuttal. Respondents' task is to decide for each component from Toulmin's (1958) model whether or not it is represented in the sentences of the four arguments. The arguments used in the Argument Structure Test were adapted from the psychological literature and are similar in content to those typically encountered by students in the humanities and social sciences. They differ not only in length but also in complexity (*claim-first* versus *reason-first* arguments; see Britt and Larson 2003).

According to Münchow et al. (2020b, see also Münchow et al. 2020a), the Argument Structure Test shows acceptable internal consistency (Cronbach's $\alpha = 0.76$). Similar internal consistencies were obtained in the present study (Cronbach's $\alpha = 0.78$

for version 1, 0.82 for version 2 at T2, and 0.87 for version 2 at T3). Mean item difficulties are comparably high for the parallel versions, amounting to 0.69 for version 1 and 0.64 for version 2 (Münchow et al. 2020b). Respondents usually need 10 to 15 minutes to complete the Argument Structure Test. In the present study, the percentage of accurately assigned components served as a score of respondents' ability to identify the functional structure of informal arguments.

4.3.2 Argument Judgement Test

The Argument Judgement Test assesses individuals' ability to accurately evaluate the structural plausibility of informal arguments and to correctly identify argumentation errors (Münchow et al. 2020a) as a measure of argument evaluation skill. Like the Argument Structure Test, it comprises two parallel versions, each comprising two parts. In Part 1, respondents are presented with 15 short arguments with a claim and one or several reasons. Of these 15 arguments, ten are plausible, i.e. they contain strong and internally consistent reasons to support the claim. The remaining five arguments are implausible due to one of five common argumentation fallacies (i.e., contradiction, false dichotomy, wrong example, circular reasoning, or overgeneralization; Dauer 1989). Respondents' task in Part 1 is to evaluate whether the presented arguments are plausible or implausible. In Part 2 of the Argument Judgement Test, respondents are successively presented again with those arguments of Part 1 which they judged to be implausible. Respondents are asked to select from a list of five common argumentation fallacies the one that they think applies to the argument.

The internal consistency of the combined Argument Judgement Test score (Parts 1 and 2) is acceptable (WLE reliability coefficient = 0.63, Münchow et al. 2019; see also Münchow et al. 2020). For Part 1, the internal consistency (Cronbach's α) is 0.64, with 0.56 for plausible arguments and 0.54 for implausible arguments. In the present study, internal consistencies for the Argument Judgement Test were comparable for version 1 (Cronbach's $\alpha = 0.54$), and slightly lower for version 2 (Cronbach's $\alpha = 0.42$ at T2 and 0.40 at T3). Mean item difficulties are comparably high for the parallel versions (see Table 1 in Münchow et al. 2019). The average time to complete the Argument Judgement Test is about 10 minutes. In the present study, we used the percentage of correctly evaluated arguments in Part 1 as a test score.

4.3.3 Argument Structure Training

The Argument Structure Training imparts strategies for recognizing the structural components of informal arguments by training the identification and allocation of functional argument components (Münchow et al. 2020a) in order to enhance students' argument comprehension skills. The theoretical input focuses on the use and purpose of informal arguments, Toulmin's (1958) component model of argumentation, as well as linguistic connectors and markers for correctly identifying different argument components. In the practical phase of the training, the students were presented with a series of 12 different arguments and were asked to match each sentence

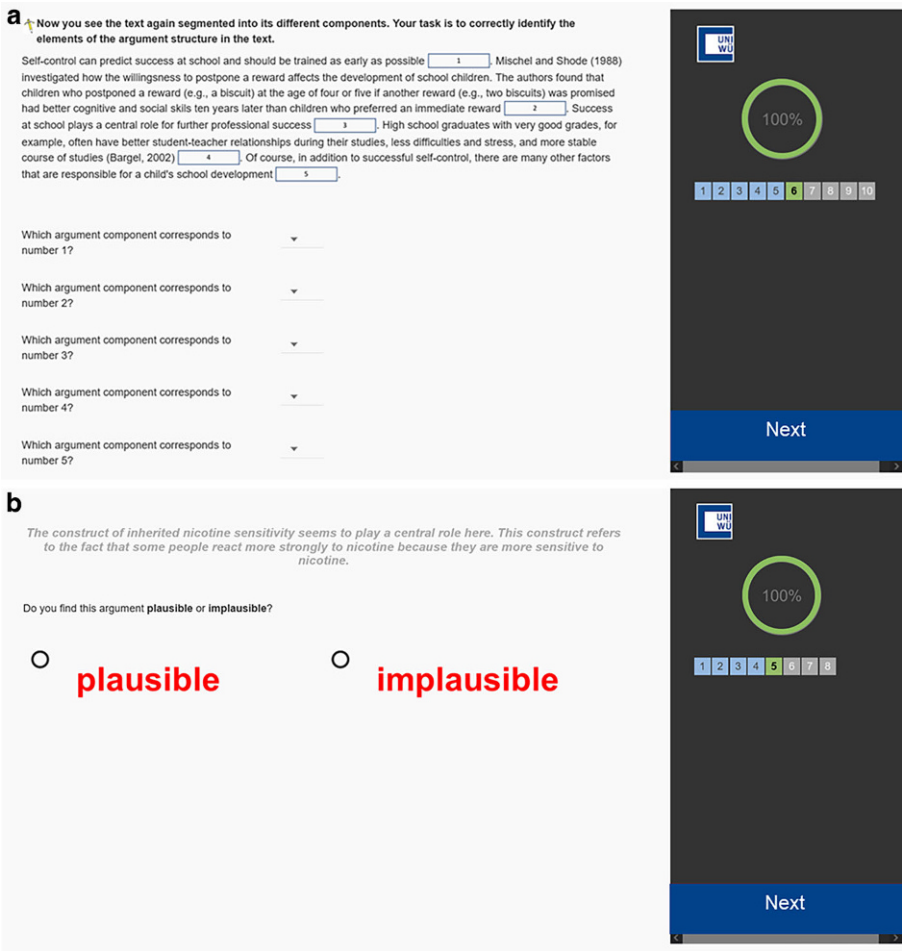

Fig. 1 Example of Practice Items from the Argument Structure Training and the Argument Judgement Training. **a** Practice item in the Argument Structure Training similar to test items. **b** Practice item in the Argument Judgement Training similar to test items

of these arguments with its correct function. An example of a typical practice task is shown in Fig. 1a. In addition, it was the students' task to find linguistic connectors and markers in the arguments and to enter them into a text box. For each argument, as soon as the students had assigned a function to all sentences, they received feedback on the correctness of their responses, followed by a schematic representation of the correct solution. If they had made a mistake in the matching of the functional components, they were presented with a brief summary of the theoretical input as a reminder.

4.3.4 Argument Judgement Training

The Argument Judgement Training teaches strategies for normatively evaluating the appropriateness of informal arguments (Münchow et al. 2020a) to foster students argument evaluation skills. The theoretical input of the Argument Judgement Training includes especially strategies for evaluating the relevance and completeness of one or several reasons for the justification of an argument's claim as well as for recognizing typical argumentation fallacies. In the practical phase, the students were given a total of 42 short arguments with a claim and one or two reasons. Of these, 32 arguments were plausible and 10 arguments were implausible because they exhibited one of five common argumentation fallacies (i.e., contradiction, false dichotomy, wrong example, circular reasoning, or overgeneralization; Dauer 1989). The students' task was first to evaluate whether the presented arguments were plausible or implausible. Figure 1b shows a typical practice task of the practical part of the training. As soon as the students evaluated an argument, they received feedback on whether their answer was correct or incorrect. If they made a mistake, they were presented with a short summary of the input as a reminder. Regardless of the correctness of their answer, for all implausible arguments, they were then asked to describe the argumentation fallacy in their own words in a text box and to indicate what the argument would have looked like if it had been plausible. Additionally, for all implausible arguments, the students were required to select the exact name of the argumentation fallacy in multiple-choice format.

4.4 Design and Analysis Plan

We employed a randomized crossover trial design (Mills et al. 2009), in which participants took part in both Argument Structure and Argument Judgement training interventions, each alternatingly functioning as an active-control group for the other (see Figure A in the Appendix). Crossover designs are used in clinical studies with at least two alternative interventions to reduce the influence of individual differences. Students were hence randomly assigned to either group A ($n = 13$) or group B ($n = 16$). Data were collected at three measurement points. At measurement point T1, all participants completed version 1 of the Argument Structure Test and the Argument Judgement Test. At measurement point T2, group A received the Argument Structure Training, while the group B received the Argument Judgement Training. Each training was directly followed by version 2 of the Argument Structure Test and the Argument Judgement Test. At measurement point T3, group A received the Argument Judgement Training, while group B received the Argument Structure Training. Each training was again directly followed by version 2 of the Argument Structure Test and the Argument Judgement Test. Participation time was on average 20 minutes at T1 and 60 minutes at T2 and T3, respectively. There was one week between T1 and T2, and one month between T2 and T3.

In this experimental design, group A first received the Argument Structure Training at T2 and then the Argument Judgement Training at T3, while group B first received the Argument Judgement Training at T2 and then the Argument Structure Training at T3. This allowed us to test the gains in argument processing after each

training phase compared to an active-control group, whereby the intervention and active-control group switched between Training Phase 1 (i.e., T1 to T2) and Training Phase 2 (i.e., T2 to T3). For example, the effect of the Argument Structure Training intervention was scored as the change in Argument Structure Test scores between measurement points T1 and T2 for group A, which received the Argument Structure intervention training at T2, and between measurement points T2 and T3 for group B, which received the Argument Structure intervention training at T3. Conversely, group B functioned as an active-control group between measurement points T1 and T2, and group A functioned as an active-control group between measurement points T2 and T3.

To test for sampling bias due to selective drop out, we compared the age, gender, number of semesters at university, average grade at university, average school leaving grade, self-evaluation of scientific literacy skills (percentage of Argument Structure Test and Argument Judgement Test items predicted to be solved correctly), and test-scores on the Argument Structure Test and the Argument Judgement Test before the first training session between participants in the final sample ($N = 29$) and those who did not complete all three measurement points ($N = 72$). The group means, differences tests, and effect size estimates are summarized in Table A in the Appendix. There were no significant differences between groups, except for the Argument Judgement Test, in which the participants in the final sample scored significantly higher than participants who dropped out of the training (Cohen's $d = 0.53$). There were no differences in the distribution of drop-out across groups A and B ($\chi^2 < 2$).

4.5 Availability of Data and Materials

Data files and analysis scripts for the full analyses are available in the Open Science Framework (https://osf.io/qtg5j/). The training materials and tests are available from the authors upon request.

5 Results

The mean scores for the Argument Structure Test and the Argument Judgement Test are presented in Table 1 for group A and B at each measurement point. Change scores were computed and compared between the intervention training and active-control groups between measurement points T1 to T2 (Training Phase 1), and T2 to T3 (Training Phase 2). Analyses were conducted in the R environment (R Core Team 2016) using linear regression models and planned contrasts. Change scores were computed for Training Phase 1 by subtracting participants' Argument Judgement Test/Argument Structure Test scores at T1 from their scores at T2, and for Training Phase 2 by subtracting their Argument Structure Test/Argument Judgement Test scores at T2 from their scores at T3. Positive change scores thus represent an increase in argument processing, which are displayed in Fig. 2. Training condition (intervention training vs. active-control) and training phase (change from T1 to T2 vs. change from T2 to T3) were included as effect-coded factors (–1 vs. 1) to test the between group effects of the training interventions. Results of the regression

Table 1 Mean and Standard Deviations of Argument Judgement Test and Argument Structure Test Scores (percent correct) at each Measurement Point (T1–T3)

Group	T1 Pre-training	T2 Training Phase 1		T3 Training Phase 2		Contrast	Change-score
	Test-score	Condition	Test-score	Condition	Test-score		
			Argument Structure Test				
A	75 (17)	**AS-training**[a]	**87 (11)**	AJ-training[b]	68 (22)	T1 vs. T2	12***
B	72 (17)	AJ-training[b]	66 (18)	**AS-training**[a]	**71 (23)**	T2 vs. T3	5
			Argument Judgement Test				
A	78 (14)	AS-training[b]	70 (17)	**AJ-training**[a]	**84 (10)**	T2 vs. T3	14***
B	76 (17)	**AJ-training**[a]	**80 (10)**	AS-training[b]	79 (14)	T1 vs. T2	4

Test-scores are reported as the average percent of correct responses on the Argument Judgement Training and Argument Structure Test, respectively. Standard deviations are reported in brackets. Test-scores following intervention training are written in **bold**

AS Argument Structure, *AJ* Argument Judgement, *Contrast* comparison to compute change in test-scores following training intervention, *Change-score* change in percent correct responses before and after intervention training

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

[a]Group receiving training intervention

[b]Active-control group

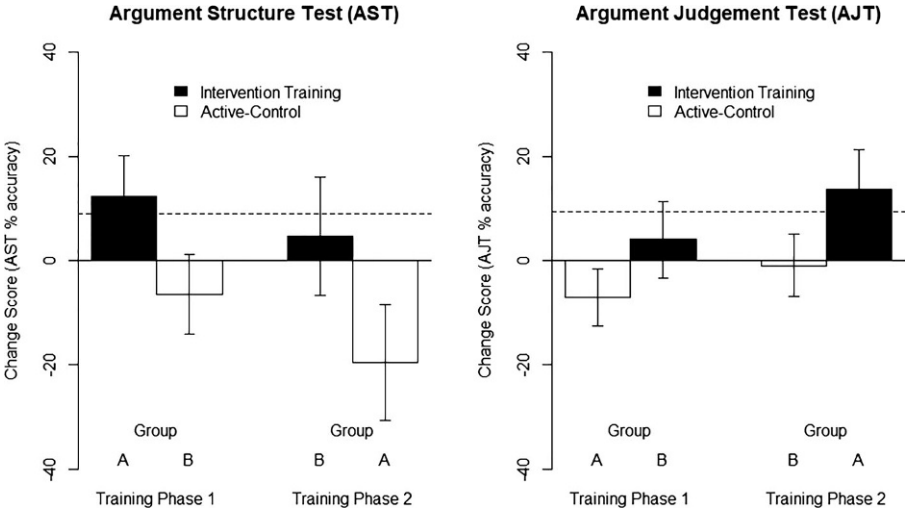

Fig. 2 Change-scores for Intervention Training and Active-Control Groups for Training Phases 1 and 2. Change-scores represent percent correct responses following intervention training. Change-scores for Training Phase 1 were computed as contrasts between T1 and T2, while change-scores for Training Phase 2 were computed as contrasts between T2 and T3. Error bars represent two standard errors from the mean. The dashed line represents the average change-score after the intervention training across groups A and B. (*A* group received Argument Structure Training at T2 and the Argument Judgement Training at T3, *B* group received Argument Judgement Training at T2 and Argument Structure Training at T3)

Table 2 Effects of Group (Control, Training) and Training Period (T1 to T2, T2 to T3) on Change-Scores

Effect	*b*	*SE*	*t*	*p*-value
Argument Structure Test				
Intercept	–0.022	0.024	–0.913	0.365
Training Condition (Training vs. Active-control)	0.108	0.024	4.428	<0.001
Training Phase (1 vs. 2)	0.052	0.024	2.141	0.037
Training Condition × Training Phase	–0.014	0.024	–0.564	0.575
Argument Judgement Test				
Intercept	0.024	0.017	1.452	0.152
Training Condition (Training vs. Active-control)	0.064	0.017	3.836	<0.001
Training Phase (1 vs. 2)	–0.040	0.017	–2.373	0.021
Training Condition × Training Phase	–0.009	0.017	–0.535	0.595

The factor Training Condition was effect coded as 1 = Training and –1 = Active-Control. Training Phase was effect coded as 1 = Training Phase 1 and –1 = Training Phase 2

analyses are summarized in Table 2 and effect sizes in change-scores are reported as model estimates in the following passages.

5.1 Argument Structure Training

For the Argument Structure Test, we found a significant main effect of training condition on Argument Structure Test scores, $t = 4.428$, $p < .001$. This main effect was broken down into the simple main effect of measurement point for each group, using planned contrasts (Hypothesis 1). The change-score for the group receiving the Argument Structure Training between measurement points T1 and T2 was significant and positive with test performance increasing by 12.38 percentage points, $b = 0.124$, $SE = 0.046$, $t = 2.689$, $p = .009$, while it was not significant for the active-control group, $b = –0.064$, $SE = 0.051$, $t = –1.254$, $p = .215$. The difference in change-score between intervention training and active-control group was also significant, $b = 0.188$, $SE = 0.069$, $t = 2.732$, $p = .008$, corresponding to a large effect size of Cohen's $d_s = 1.02$ (Lakens 2013, equation 2). However, the effect of the Argument Structure Training between T2 and T3 was not significant, $b = 0.047$, $SE = 0.051$, $t = 0.922$, $p = .360$, and the change-score for the active-control group was significant and negative with test performance decreasing by 19.57 percentage points, $b = –0.196$, $SE = 0.046$, $t = –4.249$, $p < .001$. The significant positive main effect of Training Phase, $t = 2.141$, $p = .037$, indicated that training effects were greater in the first Training Phase. However, the absence of an interaction of Training Phase and Training Condition, $t < 1$, indicated that there were no training sequence effects. This pattern of results suggests that the Argument Structure Training had some positive influence on the students' argument comprehension skills; however, the active-control group appeared to perform worse in the Argument Structure Test after receiving the Argument Judgement Training.

5.2 Argument Judgement Training

For the Argument Judgement Test, there was a significant main effect of training condition and training phase on Argument Judgement test scores, $t = 3.836$, $p < .001$. These main effects were again broken down into their simple main effects using planned contrasts (Hypothesis 2). The effect of the Argument Judgement Training between T1 and T2 was not significant, $b = 0.039$, $SE = 0.035$, $t = 1.133$, $p = .262$, and the change score for the active-control group was significant and negative with test performance decreasing by 7.08 percentage points, $b = –0.071$, $SE = 0.032$, $t = –2.230$, $p = .030$. However, the change score between the group receiving the Argument Judgement Training between T2 and T3 was significant and positive with test performance increasing by 13.75 percentage points, $b = 0.138$, $SE = 0.032$, $t = 4.328$, $p < .001$, while it was not significant for the active-control group, $b = 0.009$, $SE = 0.035$, $t = 0.260$, $p = .796$. The difference in change-score between training and active-control group was also significant, $b = 0.147$, $SE = 0.047$, $t = 3.091$, $p = .003$, corresponding to a large effect size of Cohen's $d_s = 0.87$. The significant positive main effect of Training Phase, $t = –2.373$, $p = .021$, indicated that training effects were greater in the second Training Phase. However, the absence of an interaction of Training Phase and Training Condition, $t < 1$, indicated that there were no training sequence effects. This pattern of results suggests that the Argument Judgement Training had some positive influence on the students' argument evaluation skills; however, the active-control group appeared to perform worse in the Argument Judgement Test after receiving the Argument Structure Training.

6 Discussion

The aim of the present study was to implement and evaluate a training program to enhance students' scientific literacy skills in an ecologically valid, higher education setting. The training interventions targeted students' ability to comprehend the functional structure of informal arguments and evaluate the structural plausibility of informal arguments in brief scientific texts. The training program was conducted online with pre- and post-training assessments, accompanying the students' regular academic curricula. The results suggest that students who participated in the training interventions performed, on average, significantly higher on a subsequent test of the particular scientific literacy sub-skill directly following training, compared to a control group that received training in a different sub-skill of scientific literacy. However, gains varied substantially and there appeared to be an interference between the training interventions, suggesting that students were overly concentrated on specific aspects of scientific literacy being taught, rather than integrating these skills.

6.1 The Effectiveness of Training Scientific Literacy as Part of Higher Education

The positive training effects found for students who participated in the present study are in line with previous findings showing that students' argument comprehension and evaluation skills can be trained with computer-based interventions (Larson et al. 2009; von der Mühlen et al. 2018). Unlike in previous studies, the training interventions in the present study were conducted in an ecologically valid setting rather than in a laboratory. To our knowledge, this is the first study showing that argument processing can be improved with a voluntary online training program offered as an integrated activity in a regular university course. These results have important implications for higher education teaching as they demonstrate that offering easily accessible and economically feasible trainings can, in principle, be employed to compensate for students' lack of previous formal tuition in scientific literacy skills. Our results further indicate that the training interventions presented in this study specifically improved aspects of scientific literacy targeted by the training. When students received the Argument Structure Training, they improved in their skills to recognize functionally distinct components of informal arguments, which we interpret as an indicator of argument comprehension (Münchow et al. 2020b). In contrast, they did not improve in their evaluation of the arguments' structural plausibility as a measure of argument evaluation (Münchow et al. 2019). The reversed pattern of results was found when students worked with the Argument Judgement Training. Thus, the results of the present study indicate that the provided training interventions can be applied effectively in a higher education teaching and each target specific aspects of scientific literacy.

Previous training studies that aimed to improve students' argument processing investigated either argument comprehension (e.g., von der Mühlen et al. 2018) or argument evaluation (e.g., Larson et al. 2009). The present study extends these findings by providing training in both skills required for a reasoned understanding of informal arguments (e.g., Britt and Larson 2003; Shaw 1996). However, we found an interference effect between the two training interventions, in which the ability in one aspect of scientific literacy decreased immediately following the training of the other aspect. This counter-intuitive finding may be explained with the help of mental model theory (e.g., Johnson-Laird 1983). When practicing specific skills in one of the two training interventions, students' current mental model may become more congruent with the task that corresponds to the skill that was trained, possibly leading to less intensive or misguided work in the other task. Similar effects were found in Experiment 2 from Larson et al. (2004), in which students performed better in an argument comprehension task when they were instructed to read scientific arguments with the goal of comprehending (i.e., match between reading goal and task) but not when they were instructed to read those arguments with the goal to evaluate them (i.e., mismatch between reading goal and task). Thus, the interference effect found in the present study could indicate that students did not realize that the skills trained in the training interventions are not mutually exclusive but represent different subskills of the superordinate construct of scientific literacy and that these skills should be used in a complementary manner (Britt et al. 2014).

6.2 Study Limitations

The results of our training study to foster argument comprehension and evaluation should be considered with respect to a number of limitations, some of which are inherent in training programs in applied settings. The first important issue is participant drop-out. The design of the study required three measurement points to include a pre-test assessment, and two training and post-test assessment phases. As the intervention was offered as a voluntary activity accompanying a regular academic course, participation relied entirely on student motivation and perseverance. In the present case, this led to a high drop-out rate, the exact cause of which cannot be specified. However, our analysis of individual differences between participants who completed all three measurement points and those who did not, allow us to narrow plausible explanations. We did not, for instance, find any significant differences in participants' age, time at university, self-concept in either scientific literacy sub-skill, or non-random drop-out across conditions. The only difference between groups was their pre-training scores on the Argument Judgement Test, suggesting that our sample was slightly positively sampled on scientific literacy ability. However, we have no strong evidence suggesting that drop-out was selective based on the students' general academic ability or demographics. However, we did not assess participants' motivations for participation, or other workload factors, and therefore cannot rule out a potential influence of these variables.

A second important issue is that students' tests scores on the assessments of both aspects of scientific literacy did not increase equally in both groups as a result of the training intervention, even though the groups were randomly assigned. This may be due to a number of factors. One explanation might be differences in the general ability and engagement of the participants, resulting in differences in training effectiveness. However, the experimental control of the study and our ability to identify participants who did not seriously participate (e.g., skipped over instruction and test items), makes this explanation unlikely. A different explanation may concern individual differences between participants, as previously discussed, such as workload or the fit of environmental factors (time of day, location) and personal preferences. As we therefore cannot fully discount systematic influences on training success, this remains an important caveat to the generally positive evaluation of the effectiveness of the training programs.

A third limitation regarding the generalizability of our results lies in the domain-specificity of the scientific literacy tests and training programs, as training results may not necessarily transfer to other domains. The materials were designed to test and train scientific literacy skills in cognitive and social science contexts, with content drawn from psychological literature. The topics were, therefore, relevant to a wide range of university degrees, and participating students indeed studied a wide range of subjects. The constraint in generalizability can also be seen as a consequence of the applied setting. While the domain-specificity of our materials may imply that gains in scientific literacy skills are specific to this domain, it arguably also increases the ecological validity of our training study, in that participants received training in skills that were directly relevant to the subject matter in the course that the training accompanied.

Finally, internal consistencies for the Argument Judgement Test were not high in the present study. Test scores from the Argument Judgement Test therefore should be interpreted carefully with respect to reliability constraints. However, Münchow et al. (2019) argue that internal consistency may not be the best estimate for reliability, as these measures are more likely to be lower bounds on the actual reliability of a test. Importantly, the authors report remarkably stable correlations between Argument Judgement Test scores ($r = 0.60$) within a 13-month interval, which may be interpreted as evidence of predictive validity.

6.3 Future Research

The present research replicated results of previous studies that found positive effects of computer-based training interventions on students' argument processing (e.g., von der Mühlen et al. 2018), while implementing the training in an ecologically valid setting. The results show that we were generally able to achieve this goal. As in previous studies, we found that students' argument comprehension and evaluation skills generally improved following computer-based training interventions. Moreover, our study provides initial evidence that such training interventions can be effective when applied as online supplement to regular academic courses. Our results further indicate that the two skills examined in this study can be trained independently, which allows our interventions to be used for targeted training. Two important practical implications can be derived from these findings. First, time and thus cost-effective interventions that are easily incorporated into university curricula can be employed to improve students' understanding of informal reasoning, which is an important aspect of scientific literacy (Britt et al. 2014). Second, aspects of scientific literacy appear to respond to specific training, allowing a targeted approach to interventions offered to students, according to their individual needs.

Nevertheless, more research is needed to further elaborate these findings. One question that remains is how to adequately interpret the interference effect between training programs found in the present study. At least three plausible explanations present themselves, as to why students performed poorly on one scientific literacy assessment, immediately after receiving the training for the other aspect of scientific literacy. First, they may have been unaware that the scientific literacy sub-skills being taught were complementary. Second, they may have had insufficient ability, regardless of awareness, to switch between the scientific literacy tests. Third, the training may have led to a strong focus of motivational engagement, at the cost of the assessment of the sub-skill which was not the target of the training. It may therefore be advisable to explicitly instruct students that argument comprehension and evaluation are different aspects of scientific literacy before training them in these skills or to integrate both aspects of scientific literacy in one training approach. To our knowledge, research on training scientific literacy has focused mainly on specific sub-skills rather than a combination of different skills. Hence, future research could focus more on how different sub-skills of scientific literacy relate.

Moreover, not all students in our study participated for the entire duration of the study. Thus, future research should assess students' reasons for participation and early termination to obtain information about which students can be targeted

by such training programs and which may need other methods of scientific literacy training. Furthermore, future studies should also address possible effects of domain-specific prior knowledge. Although argument comprehension and evaluation skills are, in general, domain unspecific, because the functional structure of informal arguments is independent of content, there are different types of justifications for certain domains that may require domain-specific prior knowledge. In the social sciences, for example, reasons are often provided as quantitative empirical evidence, whereas in other domains, claims may be supported by more theoretical or practical reasons.

Also, as the functional structure of scientific texts as a whole often resembles prototypical informal arguments, it is reasonable to assume that training students in sub-skills of scientific literacy in single informal arguments can improve their ability to comprehend and evaluate the reasoning in whole scientific texts. However, future research is necessary to specifically test for this assumption.

Finally, as expertise in argumentation needs time and practice to develop (Sadler 2004), future research could examine the effectiveness of trainings that include several practice sessions distributed throughout the semester. Making such trainings an obligatory part of a course (e.g., an introductory course for teaching academic and scientific working techniques) might add to its effectiveness.

6.4 Conclusion

The current study investigated the effects of training interventions, which promoted students' argument comprehension and evaluation skills, as part of their regular university curriculum. We argue that establishing such training interventions in a sustainable way in university teaching is a highly desirable long-term goal to help students familiarize themselves with the link between content knowledge, domain-specific knowledge, and knowledge about the rational evaluation of informal arguments in scientific texts. In this way, students not only absorb conceptual knowledge, but are empowered to understand the arguments on which scientific knowledge is based and to evaluate them in terms of their plausibility and credibility.

Supplementary Information The online version of this article (https://doi.org/10.1007/s11618-023-01147-x) contains supplementary material, which is available to authorized users.

Funding This research was supported by the German Federal Ministry of Education and Research (Grant no. 01PK19009).

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest H. Münchow, S.P. Tiffin-Richards, L. Fleischmann, S. Pieschl and T. Richter declare, that they have no competing interests.

References

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication*, *2*(1), 3–23. https://doi.org/10.1177/0741088385002001001.

Berkenkotter, C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Hillsale, NJ: Erlbaum. https://doi.org/10.1007/s10648-010-9124-9.

Britt, M. A., & Larson, A. (2003). Construction of argument representations during on-line reading. *Journal of Memory and Language*, *48*(4), 749–810. https://doi.org/10.1016/S0749-596X(03)00002-0.

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). Cambridge: University Press. https://doi.org/10.1017/CBO9781139048224.017.

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*(2), 104–122. https://doi.org/10.1080/00461520.2014.916217.

Chambliss, M. J. (1995). Text cues and strategies successful readers use to construct the gist of lengthy written arguments. *Reading Research Quarterly*, *30*(4), 778–807. https://doi.org/10.2307/748198.

Chambliss, M. J., & Murphy, P. K. (2002). Fourth and fifth graders representing the argument structure in written texts. *Discourse Processes*, *34*(1), 91–115. https://doi.org/10.1207/S15326950DP3401_4.

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*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1.

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

Galotti, K. M. (1989). Approaches to studying deductive and everyday reasoning. *Psychological Bulletin*, *105*(3), 331–351. https://doi.org/10.1037/0033-2909.105.3.331.

Green, D. W. (1994). Induction: Representation, strategy and argument. *International Studies in the Philosophy of Science*, *8*(1), 45–50. https://doi.org/10.1080/02698599408573479.

Hartweg, V., Milbradt, A., Zimmerhofer, A., & Hornke, L. F. (2022). *testMaker — A computer software for web-based assessments*. RWTH Aachen University, Department of Industrial and Organizational Psychology. https://doi.org/10.1016/j.learninstruc.2015.05.002.

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). Hillsdale, NJ: Erlbaum. https://opencommons.uconn.edu/chip_docs/14.

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

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863.

Larson, A. A., Britt, M. A., & Kurby, C. (2009). Improving students' evaluation of informal arguments. *Journal of Experimental Education*, *77*(4), 339–365. https://doi.org/10.3200/JEXE.77.4.339-366.

Larson, M., Britt, M. A., & Larson, A. (2004). Disfluencies in comprehending argumentative texts. *Reading Psychology*, *25*(3), 205–224. https://doi.org/10.1080/02702710490489908.

Maier, J., & Richter, T. (2013). Text-belief consistency effects in the comprehension of multiple texts with conflicting information. *Cognition and Instruction*, *31*(2), 151–175. https://doi.org/10.1080/07370008.2013.769997.

Marsh, E. J., Edelman, G., & Bower, G. H. (2001). Demonstrations of a generation effect in context memory. *Memory and Cognition*, *29*(6), 798–805. https://doi.org/10.3758/BF03196409.

Mills, E. J., Chan, A. W., Wu, P., Vail, A., Guyatt, G. H., & Altman, D. G. (2009). Design, analysis, and presentation of crossover trials. *Trials*, *10*(1), 1–6.

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, L. M., & Berthold, K. (2015). The use of source-related strategies in evaluating multiple psychology texts: A student-scientist comparison. *Reading and Writing*, *29*(8), 1677–1698. https://doi.org/10.1007/s11145-015-9601-0.

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). Judging the plausibility of argumentative statements in scientific texts: A student-scientist comparison. *Thinking and Reasoning*, *22*(2), 221–246. https://doi.org/10.1080/13546783.2015.1127289.

von der Mühlen, S., Richter, T., Schmid, S., & Berthold, K. (2018). How to improve argumentation comprehension in university students: Experimental test of a training approach. *Instructional Science*, *47*(2), 215–237. https://doi.org/10.1007/s11251-018-9471-3.

Münchow, H., Richter, T., von der Mühlen, S., & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network and relevance for academic success at the university. *British Journal of Educational Psychology*, *89*(3), 501–523. https://doi.org/10.1111/bjep.12298.

Münchow, H., Richter, T., & Schmid, S. (2020a). What does it take to deal with academic literature? Epistemic components of scientific literacy. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper & C. Lautenbach (Eds.), *Student learning in German higher education: Innovative modelling and measurement approaches and research results* (pp. 241–260). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-27886-1_12.

Münchow, H., Richter, T., von der Mühlen, S., Schmid, S., Bruns, K., & Berthold, K. (2020b). Verstehen von Argumenten in wissenschaftlichen Texten: Reliabilität und Validität des Argumentstrukturtests (AST) [Comprehension of arguments in scientific texts: Reliability and validity of the Argument Structure Test (AST)]. *Diagnostica*, *66*(2), 136–145. https://doi.org/10.1026/0012-1924/a000225.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240. https://doi.org/10.1002/sce.10066.

OECD (2019). *What students know and can do* (PISA 2018 Results, Vol. 1). Paris: OECD Publishing. https://doi.org/10.1787/5f07c754-en.

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463–466. https://doi.org/10.1126/science.1183944.

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). Hillsdale, NJ: Erlbaum.

R Core Team (2016). *R: A language and environment for statistical computing. [Computer software]*. R Core Team. https://www.R-project.org/

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* (pp. 125–140). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1793-0_7.

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*(3), 538–558. https://doi.org/10.1037/a0014038.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513–536. https://doi.org/10.1002/tea.20009.

Shaw, V. F. (1996). The cognitive processes in informal reasoning. *Thinking and Reasoning*, *2*(1), 51–80. https://doi.org/10.1080/135467896394564.

Suppe, F. (1998). The structure of a scientific paper. *Philosophy of Science*, *65*(3), 381–405. https://www.jstor.org/stable/188275.

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

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*(4), 337–350. https://doi.org/10.1016/0959-4752(91)90013-X.