Promoting Sixth Graders' Credibility Evaluation of Web Pages: An Intervention Study

Elina K. Hämäläinen (Corresponding author)
Faculty of Education and Psychology, University of Jyväskylä
tel. +358408053352
email: elina.k.hamalainen@jyu.fi
P.O. Box 35
40014 University of Jyväskylä, Finland
ORCID: 0000-0001-7561-0530

Carita Kiili
Department of Education, University of Oslo, Norway
and Faculty of Education and Culture, Tampere University
P.O. Box 700
33014 Tampere University, Finland
email: carita.kiili@tuni.fi
ORCID: 0000-0001-9189-4094

Miika Marttunen
Department of Education, University of Jyväskylä
P.O. Box 35
40014 University of Jyväskylä, Finland
email: miika.marttunen@jyu.fi
ORCID: 0000-0002-4554-9764

Eija Räikkönen
Faculty of Education and Psychology, University of Jyväskylä
P.O. Box 35
40014 University of Jyväskylä, Finland
email: eija.m.raikkonen@jyu.fi
ORCID: 0000-0003-4450-9178

Roberto González-Ibáñez
Department of Information Technology, University of Santiago, Chile
email: roberto.gonzalez@usach.cl
ORCID: 0000-0001-6849-1799

Paavo H. T. Leppänen
Department of Psychology, University of Jyväskylä
P.O. Box 35
40014 University of Jyväskylä, Finland
email: paavo.ht.leppanen@jyu.fi
ORCID: 0000-0002-8941-2225

Abstract

This study investigated whether a teacher-led intervention program on online inquiry improved sixth graders' performance in a credibility evaluation task. Students (N = 342) were divided into two conditions, an intervention group (190 students) and a control group (152 students). The intervention program (21 x 45 min lessons) was implemented during a six-week course as a part of normal schoolwork. The program included explicit teaching of online inquiry skills: searching for information (3 lessons), evaluating credibility (3 lessons), and synthesizing information (3 lessons). In addition, the skills taught were applied in two online inquiry projects comprising 12 lessons in total. The control group received business-as-usual teaching. Students' performance in the credibility evaluation task was measured before and after the program by pre and post-tests. In both tests, students evaluated three Web pages dealing with two topics (Computer Gaming and Reading on Screen) varying in their perspectives and argumentation. Students rated the credibility of each Web page and justified their ratings. Topic order was counterbalanced in both conditions. The background variables (Pre-test scores, Reading comprehension, Reading fluency, Gender, Topic order, and Test order in the pre-test) were controlled for in the multilevel negative binomial regression analysis. The results showed that the intervention program helped students better justify their credibility ratings by reference to source features but not to the argumentation or other aspects of the content compared to controls. Instructional implications of the findings are discussed.

Keywords: Evaluation - Internet - Intervention - Adolescents - Online Reading

1. Introduction

Given that the Internet, and the variable quality of the resources it offers, has become a dominant information channel, educating consumers to be critical of online information has also become an important goal of basic education (EU Council recommendation on key competences for lifelong learning, 2018; Finnish National Board of Education, 2016). For even the youngest students, the Internet is a crucial information resource (Livingstone et al., 2018). Therefore, instruction targeted at fostering critical evaluation skills should take place already in the early school years (Leu et al., 2015). Recent research findings that adolescents read and use online information in a rather superficial and uncritical manner indicates that supporting students to become critical consumers of online information is of paramount importance (Coiro et al., 2015; Macedo-Rouet et al., 2019).

Unfortunately, school traditions and teaching practices often continue to emphasize the reading of single texts, and hence one-sided learning content, instead of the processing of multiple documents and materials presenting conflicting views (Macedo-Rouet et al., 2013; Paul et al., 2017). The latter approach would better prepare students for the demands of working with unfiltered online information. Along with multiple document literacy practices, students need explicit instruction in how to evaluate the credibility of information when working with multiple information resources on the Web. While increasing efforts have been made to develop intervention programs to promote students' evaluation skills (e.g., Bråten et al., 2019; Pérez et al., 2018), teacher-led interventions remain very rare among elementary school students (see reviews by Brand-Gruwel & van Strien, 2018; Brante & Strømsø, 2018).

To address these requirements, we developed and tested a teacher-led intervention program for elementary school students (sixth graders) in which evaluation of the credibility of Web pages was contextualized as a part of an online inquiry and multiple document reading. The intervention program focused on the central credibility evaluation skills that

only a few adolescents are able to apply when reading and learning online (e.g., Forzani, 2018; Kiili et al., 2018).

1.1 Credibility Evaluation

Credibility evaluation is crucial when reading single texts (Pressley & Afflerbach, 1995) but even more critical when reading multiple texts on paper (Perfetti et al., 1999) or online (Brand-Gruwel et al., 2009; Leu et al., 2015). Evaluating the quality of information presented in multiple texts refers to evaluating the credibility of the source (source-based evaluation) and the credibility of the ideas presented in the text (content-based evaluation) both within and across texts (Forzani, 2019; Stadtler & Bromme, 2014).

The documents model framework (Britt et al., 2014; Perfetti et al., 1999) highlights the importance of information about the source of texts, also often termed *source features,* when readers evaluate and process multiple texts. Source features can include the authors' expertise, credentials, affiliation and motives, and document type and date. When readers engage in *sourcing* they identify and draw on source features in different phases of reading to predict, interpret, and evaluate and use document content in relation to the reading task (Bråten et al., 2018b; Brante & Strømsø, 2018). However, recent studies have shown that integration of source information in written products is challenging for secondary school students (Kiili et al., 2020b; Perez et al., 2018) and especially for students at the elementary school level (Kiili et al., 2020a).

Content-based evaluation can refer to evaluation of the argumentative purpose of the text (Mateos et al., 2018) and the accuracy of the ideas presented in the text, including evaluation of the argumentation used, i.e., the author's claims, reasons and evidence, and explanations given (Forzani, 2019). In evaluating argumentation, readers judge whether arguments are supported with valid reasons (Larson et al., 2009), whether both sides of an issue are considered (Means & Voss, 1996) and the quality of the evidence (e.g., single

experience, research) presented (Hoeken, 2001). To evaluate the content quality, readers can also corroborate information using their prior knowledge or other texts (Forzani, 2019).

It should be noted that source features and content should not be perceived as separate entities when evaluating the credibility of Web pages (Paul et al., 2017). Two models building on Documents Model, the Discrepancy-Induced Source Comprehension (D-ISC) Model (Bråten & Braasch, 2018) and the Content-Source Integration (CSI) Model (Stadtler & Bromme, 2014) explain the reciprocal relationship between evaluation of knowledge claims and evaluation of sources when readers encounter conflicting scientific or socio-scientific information. According to D-ISC Model, readers' attention to sources increases when different sources provide conflicting information about the issue under examination. In these types of situations, linking sources to conflicting pieces of information helps readers to organize conflicting views that further enable them to build a coherent mental representation on the issue. Furthermore, paying attention to the source features can help readers to understand why different Web pages represent diverse perspectives or positions (Brante & Strømsø, 2018).

Further, CSI Model (Stadtler & Bromme, 2014) explains how readers may resolve conflicts when evaluating competing knowledge claims. Readers can evaluate the validity of claims by relying on their own understanding of the issue or evaluating the source of information. Particularly in situations where readers do not have sufficient prior knowledge, they may prefer to rely on their evaluations of author's expertise and intentions (see also Bråten et al., 2018a). Stadtler and Bromme (2014) stress that often these two ways of evaluation tend to complement each other.

1.2 Previous Intervention Research

Given the importance of the ability to evaluate the credibility of Web pages and the limited nature of students' skills for doing so (Kiili et al., 2018; Macedo-Rouet et al., 2019),

several studies have tested instructional methods designed to improve these skills. Recent reviews on sourcing (Brante & Strømsø, 2018) and online inquiry interventions (Brand-Gruwel & van Strien, 2018) show that most studies have been conducted among students in upper secondary or higher education institutions (e.g., Bråten et al., 2019). Controls have less often been used in studies with younger students. Fewer interventions with a control group have been conducted among younger students. Most of the interventions carried out at the lower secondary level (mean student age 15 years) have reported positive results (e.g., Argelagós & Pifarré, 2012; Mason et al., 2014; Pérez et al., 2018) whereas those at the elementary school level (mean student age 10 years) have shown more varying success in enhancing students' credibility evaluation skills (e.g., Kingsley et al., 2015; Macedo-Rouet et al., 2013; Zhang & Duke, 2011). These interventions with younger students have differed in several important respects.

First, the measures used to evaluate the efficacy of the instructional methods employed have differed. While some studies have applied credibility scales to single Web pages or have rank-ordered or rated the credibility of Web pages (Kammerer et al., 2016; Mason et al., 2014), others have asked students not only to rate but also to justify their credibility ratings or Web page evaluations in their own words (e.g., Kroustallaki et al., 2015; Walraven et al., 2013; Zhang & Duke, 2011). While the formulation of justifications can be a demanding task for young students, it may also reveal the criteria they are able to apply when evaluating Web pages. It is of crucial importance that the evaluation tasks are selected so that they are neither too demanding for all the students (Kingsley et al., 2015; Zhang & Duke, 2011) nor too easy for the more skilled students (Macedo-Rouet et al., 2013).

Second, interventions have varied in length, content, and whether they have been led by a teacher or/and a researcher. Teacher-led interventions have lasted for several lessons, often including all phases of online inquiry: searching, evaluating, synthesizing, and

communicating information to others (Argelagós & Pifarré, 2012; Kingsley et al., 2015; Walraven et al., 2013). In contrast, researcher-led interventions have typically lasted for one lesson only and focused only on specific aspects of online inquiry such as evaluating the credibility of Web pages (Kammerer et al., 2016; Macedo-Rouet et al., 2013; Mason et al., 2014). Owing, for example, to teachers' problems in implementing intervention programs (Walraven et al., 2013), teacher-led, long-lasting interventions have not invariably shown improvement in students' credibility evaluation skills (Brante & Strømsø, 2018). In turn, the effects of researcher-led interventions on students' credibility evaluation skills have been either positive (e.g., Mason et al., 2014) or partially positive (Kammerer et al., 2016; Macedo-Rouet et al., 2013).

Even though brief researcher-led interventions have shown positive outcomes in students' credibility evaluation skills (Bråten et al., 2018a; Mason et al., 2014), it is nevertheless important to involve teachers in the implementation of intervention programs to learn how instructional methods work in regular classrooms. In the present study, the teaching of credibility evaluation skills was embedded in the teaching program as one component of an online inquiry intervention designed by the researchers but implemented by regular classroom teachers.

1.3 Instruction for Credibility Evaluation

In this section, we introduce several instructional features that we applied in designing the present intervention program: modeling (Coiro, 2011; Davey, 1983), cognitive and metacognitive prompts (Berthold et al., 2007; Quintana et al., 2005), discussions (Applebee et al., 2003; Teasley, 1995), and eliciting sourcing behaviour by presenting sources providing conflicting information (Brante & Strømsø 2018; Pérez et al., 2018).

As recent research has shown, many young students have limited credibility evaluation skills (Coiro et al., 2015; Kiili et al., 2018). One useful way to introduce students

to different evaluation strategies is to *model* them. For example, a teacher may model the processes of skilled evaluators interacting with Web pages (Coiro, 2011; Davey, 1983) or display a video showing a person applying evaluation strategies (van Gog et al., 2014; van Wermeskerken & van Gog, 2017). Modeling seems to be the most effective method for novice learners, and it seems to be important that it is followed by active engagement with the modeled strategies (Frerejean et al., 2018). Students can analyze the target skills by themselves (Fisher & Frey, 2015), contrast their own approach with that of an expert (Frerejean et al., 2018), or compare the performance of two models in which the evaluation skills are at different levels (Bråten et al., 2019). In addition, to strengthen and automate the modeled and analyzed skills, they must be practiced (Fisher & Frey, 2015; Frerejean et al., 2018).

Research has also shown that in many cases students do not spontaneously evaluate Web pages (Paul et al., 2017). These students would probably benefit from *prompting*, a method that seeks to direct their attention to important aspects of learning and to stimulate cognitive activities that students do not otherwise execute spontaneously (Quintana et al., 2005; Zhang et al., 2015). For example, cognitive prompts can be used to help students pay attention to the source features (e.g., Bråten et al., 2018a; Paul et al., 2017) or remind them to stop and consider the quality of the content (e.g., Britt & Aglinskas, 2002) of the Web pages they are reading. These prompts can be embedded in digital tools (Authors, in press; Quintana et al., 2005) or worksheets (Kammerer et al., 2016; Zhang & Duke, 2011).

According to Pérez et al. (2018), written prompts alone are not enough to engage students in productive sourcing behavior, and hence additional group *discussions* are needed. Furthermore, discussions are crucial in facilitating students' awareness of the connections between different source features and the reliability of Web pages (Bråten et al, 2019; Macedo-Rouet et al., 2013; Pérez et al., 2018). It also seems important that such discussions

in the classrooms are structured, focus on the most important aspects to be learned, and carefully implemented with sufficient time resources (Walraven et al., 2010; 2013).

As well as models and prompts, it is important that teachers provide students with Web pages that elicit use of the modeled and prompted evaluation strategies. For example, providing *multiple partly conflicting Web pages* can trigger sourcing activities (Bråten et al., 2019; Paul et al., 2017) and improve learning outcomes related to the evaluation of Web pages (e.g., Mason et al., 2014; Pérez et al., 2018). The evaluation of multiple conflicting Web pages encourages students to pay more attention to source features compared to the evaluation of a single page (cf. Macedo-Rouet et al., 2013) or multiple pages presenting the same point of view (Brante & Strømsø, 2018). Further, the use of pre-selected Web pages, instead of using student-selected pages from open-Web sources, enables deeper-level discussions in which students can compare their evaluations of the same pages with their peers (Brante & Strømsø, 2018).

Applying the instructional features outlined above, this study examined whether a teacher-led intervention improved sixth graders' performance on an online credibility evaluation task.

1.4 The present study

The present study extends the previous intervention research on students' credibility evaluation skills by examining whether sixth graders improved in their online credibility evaluation by a teacher-led intervention. The performance of the intervention group was compared with that of the control group receiving business-as-usual -teaching. Furthermore, we also examined how students' credibility evaluations were reflected in their written products.

The specific research questions were:

1. How did the sixth graders evaluate the credibility of Web pages?

2. Did the teacher-led intervention lead to improvement in the sixth graders' performance on an online credibility evaluation task compared to the control group?
3. How were students' credibility evaluations reflected in their written products? Did the teacher-led intervention result in increase of students' use of justifications for credibility in their written products?

In terms of the second research question, we assumed that the intervention group performs better after the intervention in the credibility evaluation task than the control group. As previous research has shown that adolescents' reading skills contribute to their credibility evaluations (Forzani, 2018; Kanniainen et al., 2019; Kiili et al., 2018), reading fluency and reading comprehension were controlled for in the present study. In addition, recent studies based on PISA-data have revealed that in Finland the gender differences in literacy skills, favoring girls, are especially large (Brozo et al., 2014; Harju-Luukkainen et al., 2016). Further, the study by Forzani (2018) found that girls outperformed boys also in the evaluation of online information. Therefore, gender was also controlled for. Furthermore, it has been shown that the topic of the task plays a role in the evaluation of texts (Bråten et al., 2018a), and for that reason, we also controlled for the topic-order.

Recent studies have shown that sourcing in written products is challenging for students at different ages (e.g., Kiili et al., 2020a; Perez et al., 2018). In terms of the third research question, we therefore assumed that students' credibility evaluations will only rarely be reflected in their written products. Accordingly, we assumed that our teacher-led intervention might not help students to use justifications for credibility in their writings (cf. Perez et al., 2018).

2. Method

This study reports part of an intervention study in which students were taught online inquiry skills (searching, evaluating and synthesizing online information) within two

disciplines, *Social science* and *Science*. Because meanings are constructed in somewhat different ways in different disciplines (Goldman et al., 2016; Shanahan & Shanahan, 2012), students' online inquiry skills were evaluated with separate tasks for each discipline. This study focuses on the *Social science* discipline.[1] In addition, in order to enable the depth of the analyses and careful consideration of instructional implications, this study concentrates on examining the efficacy of the intervention on students' credibility evaluations. Furthermore, this study explores how students' credibility evaluations were reflected in their written products.

2.1 Participants

Sixth graders (N = 364) were recruited from ten primary schools (15 classes) in three suburban areas in Finland. Parental permission was received from 345 students (94.78 % of the recruited students). Two students were absent from all the tests in this study and one student was excluded for another reason. Hence, the final number of participants was 342 (165 girls and 177 boys). Participants' mean age was 12.13 years (SD = .41).

2.2 Research Design

A quasi-experimental design with pre- and post-tests was applied (Figure 1). The participating classes were divided into an intervention group and a control group. For practical reasons, convenience sampling was used: intervention group teachers were recruited based on teachers' opportunities and willingness to implement the intervention program. However, control group teachers were not offered a chance to participate in the intervention and they were recruited after the intervention group teachers. Most of them were from different schools than the intervention group teachers. As a result, the intervention group comprised 192 students (90 girls and 102 boys) in eight classes and the control group 150 students (75 girls and 75 boys) in seven classes.

Figure 1 shows the reading tests and the pre- and post-tests on the online inquiry task that students completed during the intervention study and how the topic order (Topics 1 and 2 in both disciplines) and the test order (Social science and Science) were counterbalanced. The reading tests were conducted one week before the first pre-test. Between the pre- and post-tests, the intervention group participated in a six-week teacher-led intervention program on online inquiry skills while the control group received business-as-usual teaching which follows the Finnish curriculum for basic education. The new curriculum includes broader competencies such as multiliteracies crossing all learning in schools (see Finnish National Board of Education, 2016). In Finland, teachers have a lot of autonomy to realize the aims of the curriculum and they can actualize the curriculum according to their own pedagogical views and their strengths as teachers. The teachers of the control group were not present at the introduction sessions provided for the intervention group teachers, and they received intervention materials after completion of the study. Thus, during the study the control group was not exposed to any of the teaching materials used in the intervention group.

--- FIGURE 1 about here ---

2.3 Measures

2.3.1 Measure of students' credibility evaluation skills. In the pre- and post-tests, students' credibility evaluation skills were assessed as a part of the computer-based online inquiry assessment (González-Ibáñez et al., 2017; Sormunen et al., 2018) that consisted of four phases: searching information, evaluating credibility, identifying main ideas from single texts, and composing a written product. In the Social Science task, the students were asked to explore either *Computer Gaming* (Topic 1) or *Reading on Screen* (Topic 2) in order to compose a newspaper article or an email message (written product) on the advantages and disadvantages of the topic at hand. These topics were chosen owing to their relevance for students' lives and all students were assumed to be somewhat familiar with both.

After receiving the task assignment, students were asked to evaluate their prior knowledge on the topic by answering the 5-point response scale question: *How much knowledge do you think you have about the topic of the given task?* (*1 = I don't know anything about the topic, 5 = I have much knowledge about the topic*). There was no statistically significant difference ($t(304) = -.742$, $p = .458$) between the self-evaluated prior knowledge on the topics (Computer Gaming, $M = 3.10$, $SD = 0.87$ vs. Reading on Screen, $M = 3.03$, $SD = 0.88$) in the pre-tests.

In the credibility evaluation phase, students evaluated three Web pages designed for the purpose of this study. The author, document type, perspective and position of each Web page are described in Table 1. First, students were asked to rate the credibility of each Web page on a 5-point response scale consisting of 1 to 5 stars. Second, students were asked to justify their rating by answering the following open question: *Why do you think so?* (see Appendix) Students were allowed seven minutes to evaluate all three of the Web pages they had read in the previous phase of the task. The computer bleeped three minutes before the time elapsed.

Depending on the test topic, students completed the pre-test either one or two weeks after completing the reading fluency and reading comprehension tests (see Figure 1). Similarly, the post-test was completed either one or two weeks after the intervention program. When completing the pre- and post-tests, students worked in a classroom with computers. One or two researchers were present in the classroom and observed the students' actions. If students encountered technical problems, the researcher(s) helped them.

--- TABLE 1 about here ---

2.3.2 Reading measures. To control for students' reading skills, the students' reading fluency and reading comprehension skills were measured before the pre-tests. In the reading test session (45–60 minutes), the students completed the reading fluency and the reading

comprehension tests in the classroom. The researchers gave the instructions and answered students' practical questions.

Reading fluency was measured with a word chain test (Holopainen et al., 2004). The word chain test comprised 25 chains, each containing four words written with no spaces in between. The students were asked to separate as many chains into primary words as possible within 90 seconds. The total score was the number of correctly separated words. The test-retest reliability coefficient for the original test has varied between .70 and .84 (Holopainen et al., 2004). The total score (0–100) was used in the statistical analyses.

In the *reading comprehension* test (Kajamies, 2017), the students read a text on the diversity of nature and answered three open-ended questions on the main ideas presented in it. Students could consult the text while answering the questions. Students were allowed 15 minutes to read the text and another 15 minutes to answer the questions. When needed, students were given five minutes' extra time to finish the test. Students earned 0 to 6 points from each open-ended question, and thus the maximum score was 18 points. The correlations between questions varied from .30 to .37. The total score of the test (0–18) was used in the statistical analyses. To establish inter-rater reliability, two independent researchers scored 20% of the students' answers (n = 68). The following Kappa values (Cohen, 1960) were obtained: .90 (Question 1), .68 (Question 2), and .95 (Question 3).

2.4 Intervention Program

The intervention program for online inquiry contained three modules. Module 1 (9 x 45 min lessons) consisted of explicit teaching on searching information, evaluating the credibility of information, and synthesizing information from multiple sources. Each of these component skills was taught in three lessons that followed five phases: 1) *modeling* effective strategies by showing a video in which two virtual students talked aloud while completing an inquiry sub-task, 2) *analyzing* the strategies modeled in the video with a worksheet and

sharing thoughts with other students, 3) *discussing* the strategies in a teacher-led session, 4) *practicing* the strategies and, 5) *reflecting on* what one had learned. The first three phases were conducted without computers to keep the students' attention on the main points of the strategies. In the lessons regarding searching for information, the students analyzed search queries differing in their quality. Further, the lessons regarding synthesizing information from multiple sources focused on integration of ideas from multiple Web pages. The students were taught how to compose fluent texts and use connecting words to integrate ideas. This study focused on credibility evaluation lessons that are described in more detail below.

In Module 1, three 45 minutes scripted lessons concerning credibility evaluation were implemented in two sessions (one 45 minutes lesson and one 2 x 45 minutes lessons). In designing the lessons, we employed contrasting cases and materials (cf. Braasch et al., 2013; Bråten et al., 2019). In addition, all materials covered source-based and content-based aspects of credibility evaluation at a level appropriate for sixth graders.

In the first 45 minutes lesson, students watched a video where two virtual students modeled evaluation strategies by thinking aloud when evaluating the Web news article. The virtual students varied in the versatility and sophistication of the evaluation strategies they used. The skillful virtual student employed sophisticated strategies, such as quality of evidence on the Web page, and date and type of the Web page. The other, not that skilled virtual student, used more superficial strategies, such as referring to the amount of the text on the Web page and appearance of the Web page. While watching the video, students were prompted to analyze and compare virtual students' evaluation strategies with the worksheet followed by the discussions of the strategies with their partners and the teacher (see Table 2).

In the 90-minutes double lesson, we used two contrasting Web pages that differed in their credibility. The more credible page was written by a researcher who responded to FAQs at the University website. The less credible Web page was a commercial press release written

by a head of marketing (see Kiili et al., 2018). While reading the Web pages, students were prompted to evaluate author expertise, purpose of the text and use of evidence with the worksheet. After that, the credibility of the Web pages was compared in a teacher-led discussion (see Table 2).

--- TABLE 2 about here ---

The taught credibility evaluation strategies were applied and practiced in Modules 2 and 3. In Module 2 (4 x 45 min lessons), the students practiced the online inquiry skills in a social science project on the advantages and disadvantages of social media. In the project, students 1) activated their prior knowledge on social media, 2) searched for information with Google Custom Search Engine, which contained a limited number of authentic Web pages, 3) evaluated the credibility and noted the advantages and disadvantages of social media from their reading of the selected Web pages and, 5) engaged in teacher-led discussion in which different perspectives on social media were compared and contrasted.

In Module 3 (8 x 45 min lessons), the students practiced the taught online inquiry skills in a science project on energy. The project was longer than the social science project (Module 2) as it was more closely adapted to the curriculum. The project followed the same phases as the social science project but with some differences in implementation: the students 1) activated their prior knowledge on energy, 2) searched for information in open Web sources, 3) evaluated the credibility of selected Web pages, and 4) wrote and presented syntheses based on multiple Web pages. During both projects, the students worked individually and in small groups, and their work was supported with worksheets.

Fidelity to the intervention program was assured in several ways (see McKenna et al., 2014). The teachers of the intervention group ($N = 9$) received a detailed intervention manual that they were asked to follow. The manual included a short theoretical introduction to the components of online inquiry, the goals and phases of the lessons, and instructional materials

and methods. We decided to model the strategies using videos instead of teacher modeling to ensure that all students received exactly the same examples of the strategies. Before each Module, the teachers were given a short induction session (30–60 minutes) on the materials, tasks and teaching methods. The teachers were also assigned a researcher to contact if they had any further questions about the lessons.

We observed five teachers' lessons on credibility evaluation in Module 1. In addition, other lessons from Modules 1 to 3 were followed, making a total of 44 observations of lessons. The other four teachers were interviewed after each Module. During the intervention program, the teachers recorded in a diary any deviations from the plan and their observations of the efficacy and adequacy of the tasks, materials and activities. All the students' worksheets were also collected after each Module. The observations, interviews, diaries and worksheets revealed that the intervention program was mostly conducted as planned. The teachers felt that some tasks in the worksheets were too difficult or complex for the students (in Modules 2 and 3) and therefore they made some pedagogical modifications to the plan.

2.5 Data Analyses

2.5.1 Students' justifications for their credibility ratings. Students' justifications for their credibility ratings from the pre- and post-tests were identified, categorized, and counted. The unit of analysis was an expression containing a justification. The identified justifications were divided into four categories. Two categories (Expertise of the Source and Other Source Features) concerned the evaluation of the *Source Features* (e.g., Bråten et al., 2018a; Britt & Aglinskas, 2002; Britt et al., 2014) and two (Argumentation in the Text and Other Aspects of the Content) the evaluation of the *Quality of Content* (e.g., Braasch et al., 2013; Britt et al., 2014; Judd et al., 2006; Metzger, 2007) (Table 3).

--- TABLE 3 about here ---

After categorization, the number of the justifications in each of the four categories (Expertise of the Source, Other Source Features, Argumentation in the Text, and Other Aspects of the Content) was counted. The students received one point for each relevant justification that was in accordance with their credibility rating. Four count variables were formed for both the pre- and post-tests, respectively (8 count variables in total). These count variables were constructed based on the total number of relevant justifications across the three evaluated Web pages. In statistical analyses, the post-test count variables were used as dependent variables and the pre-test count variables used as control variables.

The reliability of the categorization was calculated for a random sample of 15% of the justifications. The first rater identified and categorized the justifications in the students' answers and the other rater categorized the justifications identified by the first rater. The inter-rater agreement for the categorization was .90 (Cohen's kappa; Cohen, 1960). Disagreements were resolved through discussion between the raters.

2.5.2 Students' use of their credibility evaluations in written products. To explore how students' credibility evaluations were reflected in their written products, we examined whether students' justification for their credibility rating or a part of it appeared in their writings. If we found an overlap, it was categorized as representing one of the main categories of the justifications for credibility ratings: Source Features or Quality of Content. The examples below illustrate how students used their justifications in their essays.

> Example 1 (Source Features)
> Student's justification of credibility rating:
> *This is an opinion piece but it is written by a principal.* (ID 2007)
>
> Use of justification in the written product:
> *The principal Ulla-Maija Lehola from Comprehensive School of Helsinki says that reading on screen should be taught for students at school in order to give them sufficient skills to use Internet and different media in their learning.* (ID 2007)
>
> Example 2 (Quality of Content)
> Student's justification of credibility rating:
> *The text gives a wide understanding of advantages and disadvantages of computer*

gaming. (ID 3354)

Use of justification in the written product:
Some people think that sitting by the computer is harmful for your health whereas some people think that computer gaming can develop your skills. Both perspectives are correct. (ID 3354)

The effect of the intervention on the association between students' credibility evaluations and their written products was examined with non-parametric Wilcoxon's test separately for the intervention and control groups. In addition, Cohen's d and its 95% confidence intervals were computed for both groups.

2.5.3 Statistical analyses. The effect of the intervention on the students' credibility evaluation skills was examined by a negative binomial regression analysis (Coxe et al., 2009; Gardner et al., 1995). It is a suitable method here, as the four dependent variables (i.e. the number of justifications in the four post-test categories: Expertise of the Source, Other Source Features, Argumentation in the Text, and Other Aspects of the Content) were by their nature non-normally distributed count variables. These variables also showed large over-dispersion, meaning that the variance of each dependent variable was larger than its mean.

The dependent variables were analyzed separately. In each analysis, the corresponding Pre-test score (number of justifications) was controlled for. In all four analyses, Group (intervention or control) was used as the independent variable, whereas Gender, Reading fluency, Reading comprehension, Topic order, and Test order in the pre-test were controlled for. The descriptive statistics of all the employed variables are presented in Table 4.

Data were hierarchical in nature: that is, the students were nested within classes. Intra-class correlations (Heck, 2001; Muthén, 1991) in the pre-tests (Expertise of the Source = 0.07, Other Source Features = 0.17, Argumentation in the Text = 0.02, Other Aspects of the Content = 0.06) suggested the presence of variation between the classes in the justification scores, particularly for the variable Other Source Features. Therefore, multilevel modeling

(Muthén, 1997) was used with class as a clustering variable. The variation between the classes was taken into account by estimating the means of the dependent variables at the class level (i.e. between-level) as random. The actual negative binomial regression was conducted as a student-level (i.e. within-level) analysis.

Negative binomial regression analysis models the log of the expected count of justifications in each post-test category (dependent variables) as a function of the independent/control variables (Coxe et al., 2009; Gardner et al., 1995). For ease of interpretation, the regression coefficients were presented as incident rate ratios (IRRs) which were obtained by exponentiating the regression coefficients using base *e*. For example, for a dichotomous independent variable (i.e. Group), the IRR represents the change in the expected rate of justifications in a specific post-test category when the value of the independent variable changes from 0 to 1. An IRR greater than 1 indicates how many times greater the expected rate of justifications in the post-test category is for students with an independent variable value of 1 (i.e. Group: intervention) than for those with the value 0 (i.e. Group: control).

In contrast, an IRR smaller than 1 indicates that the expected rate of justifications in the post-test category is greater for those participants with an independent variable value of 0 (i.e. Group: control) than those with the value 1 (i.e. Group: intervention). With continuous control variables (e.g., Reading fluency score), the IRR represents the change in the expected rate of justifications in a post-test category when the value of the control variable (e.g., Reading fluency score) increases by one unit. The statistical significance of the IRRs was determined by computing 95% confidence intervals for each IRR. An IRR differs statistically significantly from the value 1 if its confidence interval does not include the value 1.

All statistical analyses were conducted using the Mplus statistical package (version 8.0; Muthén & Muthén, 1998–2017) with Full-Information-Maximum-Likelihood (FIML)

procedure (Enders, 2010). FIML uses all available information in the data to estimate the model without imputing missing values. Model parameters were estimated using maximum likelihood estimation with non-normality robust standard errors (MLR) (Muthén & Muthén, 1998–2017).

3. Results

3.1 Students' Performance on the Credibility Evaluation Task

Students' performance in the reading tests and pre- and post-tests is presented in Tables 4 and 5. In the pre-tests, the highest scores were for the category Expertise of the Source, suggesting that students most often justified their credibility ratings by reference to the expertise of the source (Table 4, variable 5). The variation in the number of justifications was also largest in this category, indicating that some students found many justifications related to this category and others none. In contrast, students seemed to find evaluation of the other credibility aspects (Other Source Features, Argumentation in the Text and Other Aspects of the Content) rather challenging (Table 4, variables 6–8). The control group outperformed the intervention group only in the category of Other Source Features (Table 5). In all the other pre-test categories and in the reading fluency and reading comprehension tests, the intervention and the control groups performed similarly, indicating no group differences at baseline.

--- TABLE 4 about here ---

--- TABLE 5 about here ---

3.2 Effect of an Intervention on the Students' Performance on the Evaluation Task

After controlling for the background and pre-test variables, the results showed that the intervention helped students to better justify their credibility ratings with reference to Expertise of the Source and Other Source Features (Table 6). The intervention group justified

their credibility ratings 1.52 times more often with expertise of the source and 1.83 times more often with other source features than controls.[2] However, the intervention group and control group did not differ in the numbers of justifications for their credibility ratings related to argumentation in the text or other aspects of the content.

--- TABLE 6 about here ---

Of the control variables (Table 6), the pre-test scores (number of justifications) explained the post-test scores; thus, the more justifications a student had in the pre-test, the more she/he also had in the post-test in each corresponding category. Furthermore, the students' reading comprehension skills were associated with the number of justifications in the post-test in the categories Expertise of the Source and Argumentation in the Text: the better reading comprehension skills a student had, the more often he/she justified his/her credibility ratings with reference to expertise of the source and the argumentation in the text.

Topic order was associated with only one of the post-test scores. The students who completed the Reading on Screen task first scored better in justifying credibility by reference to argumentation in the text than the students who completed the Computer Gaming task first. The other control variables (i.e. Reading fluency, Gender, and Test order in the pre-test) were not associated with the number of justifications in any of the four justification categories in the post-test (Table 6).

3.3 Students' Use of their Credibility Evaluations in Written Products

Table 7 presents the results of students' use of justifications for credibility ratings in their written products between intervention and control groups and results of the Wilcoxon's tests computed separately for the groups. In the pre- and post-tests, students' justifications were rarely reflected in their writings. Of all students, 8.4% in the pre-tests and 10.8% in the post-tests utilized justifications related to Source Features or Quality of Content in their essays. Wilcoxon's tests for change in Source Features and Quality of Content did not show

statistically significant change either in the intervention or control group. In addition, the confidence intervals of the effect sizes (Cohen's d) included 0 in both groups, thus showing that the effect of change in both groups was 0. These confidence intervals also overlapped in both categories (Source Features and Quality of Content). Therefore, it can be concluded that the intervention did not increase students' use of justifications in their written products (Table 7).

--- TABLE 7 about here ----

4. Discussion

The present study, with 340 students, evaluated the efficacy of a teacher-led intervention program that combined different instructional methods, such as modeling, prompting, and discussing evaluation strategies, with the aim of improving students' credibility evaluation skills. We first discuss the main findings and then evaluate the instructional design and implementation of the intervention program. Finally, we consider the limitations of this study and discuss the instructional implications of the findings.

Our intervention program improved students' skills to evaluate source features when considering the credibility Web pages. After participating in the intervention program, students more often justified their credibility ratings with reference to source features. After ruling out plausible alternative data-related explanations for the effects (i.e. Pre-test scores, Reading skills, Gender, Topic order and Test order in the pre-test), the intervention group students presented 1.52 times more justifications related to expertise of the source and 1.83 more justifications related to other source features than controls. Bråten et al. (2018a) emphasize that the ability to use source features in credibility evaluation is particularly important when students use the Internet to explore controversial issues about which they have little prior knowledge. Without relevant prior knowledge, it is very difficult to evaluate,

e.g., the accuracy of information, which suggests that readers are left to rely on source features when evaluating the credibility of Web pages.

However, we observed no improvement in students' content-based evaluation of credibility. There might be several reasons for this finding. First, paying attention to source features, i.e., author and her/his credentials, publisher and date (Britt et al., 2014; Perfetti et al., 1999), can be more concretely modeled and taught compared to content-based evaluation strategies. To be effective, teaching credibility evaluation of argumentation would have required, for example, in-depth knowledge of argument structure and what counts as high quality evidence and why. Second, students can quite easily find source features in Web pages whereas evaluation of argumentation requires careful reading of the texts and identification of claims and related evidence (Britt et al., 2014). In addition, the evaluation of ideas in light of one's prior knowledge or experience requires thoughtful and reflective reading.

Our result on the growth in the use of source features when justifying credibility was favorable and encouraging, as attention to and identification of source features is a prerequisite for other sourcing activities, such as using source features to interpret content (Brante & Strømsø, 2018). The use of source features in interpreting content is, however, a very demanding task for young students (Britt et al., 2014; Macedo-Rouet et al., 2013) and requires teaching that systematically builds on the lower-level skills, i.e., attending to source features, already acquired.

Despite the positive finding in the increase in justifications related to source features in their assessments of credibility, the proportion of students whose performance remained low was quite high (see also Kingsley et al., 2015; Zhang & Duke, 2011). Our results further showed that reading comprehension skills predicted students' performance in their evaluation of the expertise of the source and the argumentation in the text. Hence, it is possible that poor

reading skills also hinder students' performance in their justifications for the credibility of Web pages, as suggested in a recent study by Kanniainen et al. (2019).

In the present study, we also examined how students' justifications for their credibility ratings were reflected in the written products. As assumed, students very rarely used evaluations of sources and content in their writings. This is in accordance with previous studies showing that, in general, citing or evaluating sources in the written products is not that common practice, in particular, among younger students (Kiili et al., 2020a; Perez et al, 2018). In terms of synthesizing information from multiple Web pages, the focus of our intervention was not on teaching how to utilize credibility evaluations in writing but on integration of ideas with connecting words. Accordingly, our intervention did not increase students' use of justifications for credibility evaluations in their written products. As the use of credibility evaluations in the written products was rare both in the pre- and post-tests, it is evident that students need more explicit support to understand the connections between evaluations of Web pages and writing from multiple pages.

4.1 Evaluation of the Intervention

In this section, to contribute further to research in the field, we evaluate the intervention program in terms of its instructional design and implementation. Finally, we evaluate the measure used to assess the efficacy of the program.

In this study, credibility evaluation of Web pages was taught as a part of an intervention program designed to support students' online inquiry skills. This was done to provide an authentic context for the evaluation activities. After explicit teaching of different online inquiry skills, students were able to apply their new skills in two inquiry projects. This meant that the 21-lesson intervention program, which included instruction in several different kinds of skills (searching, evaluating, synthesizing) was rather long. This might have been experienced as overwhelming, especially by students who find online inquiry a struggle.

Furthermore, the explicit teaching of evaluation skills was accomplished in only three lessons, which might not be sufficient time for all students to learn a skill as complex as credibility evaluation (cf. Argelagós & Pifarré, 2012; Walraven et al., 2010).

Moreover, the credibility evaluation lessons for were combined with different instructional methods, such as modeling, prompting, and discussing strategies. The aim of combining these was to highlight different aspects of the evaluation of the credibility of Web pages and to deepen students' understanding of critical evaluation. To model the evaluation strategies (Coiro, 2011; Davey, 1983), we provided a video where two virtual students performed a credibility evaluation task. A modeling video might be especially useful for teachers who do not feel comfortable modeling strategies themselves. A video can also be a motivating tool for students (Choi & Johnson, 2005). However, the video modeled multiple evaluation strategies, which may have caused some students cognitive overload and thereby hindered their learning. On the other hand, students were able to watch the video multiple times.

In the lessons, prompts on the worksheets were used to direct students' attention to different aspects of credibility in performing their given tasks during the intervention. To maximize the benefit gained from prompts, students' responses need to be discussed in the classroom (Macedo-Rouet et al., 2019; Pérez et al., 2018). However, according to our observations, the discussions remained rather shallow and teachers did not necessarily take advantage of the guiding questions provided in the intervention manual. Thus, it seems that, for teachers, orchestrating and motivating classroom discussions was one the most challenging features of the intervention (see also Walraven et al., 2010; 2013).

To investigate the efficacy of the intervention, we used a larger online inquiry task during which the students were asked to justify their credibility ratings of three Web pages. The benefit of such a measure is that it reveals the criteria that students apply when

evaluating the credibility of Web pages during authentic online inquiry tasks (cf. Brante & Strømsø, 2018). One downside of this approach is that the students might not have perceived that the evaluation task was as important as the final writing task in which they were expected to report their findings. Moreover, some students might have had difficulties expressing their thinking when writing under the pressure of time (see Macedo-Rouet et al., 2019). Future studies could use a repertoire of measures that would enable all students to express themselves. Finally, while previous intervention studies on evaluation at the primary and lower secondary levels have used different types of outcome measures, such as credibility scales for evaluating single Web pages or rank-ordering Web pages according to their credibility (see Brante & Strømsø, 2018), comparing the efficacy of different interventions is challenging.

4.2 Limitations and Future Research

This study has its limitations, which should be taken into account in future studies. First, the intervention comprised two different domains (Social Science and Science), meaning two pre-tests and two post-tests to be completed by every student. Not all the students may have been equally motivated to complete the post-tests. This was probably true of both the intervention and control groups, especially as the mean performance of the control group declined over time (see Table 5). However, the positive effect was bigger for the intervention than control group in Expertise of the Source and Other Source Features.

Second, the long-term effects of the intervention were not measured. A follow-up could have given information on how well maintained the changes were (e.g., Bråten et al., 2019; Pérez et al., 2018). In our study design, the students had already done an online inquiry test four times, and hence may have found the inclusion of two additional delayed post-tests overwhelming many of them. On the other hand, some studies on evaluation skills

(Kroustallaki et al., 2015; Walraven et al., 2013) have reported that a transfer effect is hard to achieve. In future studies, both the long-term and transfer effects could be measured.

Finally, students' prior knowledge on the topics (Computer Gaming and Reading on Screen) was only measured with one self-evaluation question indicating no statistically significant difference between the topics in the pre-tests. However, self-evaluation question can not be regarded as an objective measure of prior knowledge. In the statistical analysis topic order was controlled for, and therefore, we were able to account for possible differences in the difficulty level of the topics.

4.3 Instructional Implications

The intervention program applied in this study, while promising, could be developed further in several ways. First, although our students found most of the credibility evaluation strategies difficult, they varied in their ability to justify their credibility ratings, especially with reference to source features. Therefore, teachers should tailor their instruction to individual student needs. Low-performing students may need more scaffolding to understand how to evaluate online information. For example, teachers could begin by showing them where to find author information on different types of Web pages (Paul et al., 2017). These students may also benefit from prompts that explicitly guide them to focus on relevant source features (cf. Kammerer et al., 2016). In this study, all the learning materials (e.g., video, worksheets) used in the intervention program mixed source-based and content-based aspects of credibility evaluation. In the future, it might be more effective to concentrate on just a few aspects of evaluation in one lesson (see Kingsley et al., 2015; Pérez et al., 2018). This might help low-performing students to focus more deeply on a limited number of issues at a time. However, the more skillful students could benefit from materials and discussions that demonstrate the interaction of source-based and content-based aspects of credibility evaluation.

Second, more emphasis should be placed on the professional development of teachers who implement intervention programs in classrooms. In order to engage in critical discussions with their students, teachers themselves need to feel comfortable with the various evaluation strategies (cf. Paul et al., 2017). Teachers also need both knowledge-based and pedagogical abilities to adapt their teaching to their students' needs and to react appropriately to students' comments and answers. In our study, the teachers were offered only a short introduction to the intervention manual and learning materials before conducting the lessons. Instead, teachers should be provided with tools not only to help them orchestrate high-quality discussions (Pérez et al., 2018; Walraven et al., 2010; 2013) but also to guide them in giving students feedback on their evaluation skills during lessons (Macedo-Rouet et al., 2019; Paul et al., 2017) and to motivate their students to regularly evaluate the credibility of Web pages (Brante & Strømsø, 2018; Pérez et al., 2018).

References

Applebee, A. N., Langer, J. A., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal, 40*(3), 685–730. https://doi.org/10.3102/00028312040003685

Argelagós, E., & Pifarré, M. (2012). Improving information problem solving skills in secondary education through embedded instruction. *Computers in Human Behavior*, *28*(2), 515–526. https://doi.org/10.1016/j.chb.2011.10.024

Authors. (in press).

Berthold, K., Nückles, M., & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction, 17*(5), 564–577. https://doi.org/10.1016/j.learninstruc.2007.09.007

Braasch, J. L. G., Bråten, I., Strømsø, H. I., Anmarkrud, Ø., & Ferguson, L. E. (2013). Promoting secondary school students' evaluation of source features of multiple documents. *Contemporary Educational Psychology*, *38*(3), 180–195. https://doi.org/10.1016/j.cedpsych.2013.03.003

Brand-Gruwel, S., & van Strien, J. L. H. (2018). Instruction to promote information problem solving on the Internet in primary and secondary education: A systematic literature review. In J. L. G. Braasch, I. Bråten, & M. T. McCrudden (Eds.), *Handbook of Multiple Source Use* (pp. 401–414). New York and London: Routledge.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using Internet. *Computers & Education, 53*(3), 1207–1217. https://doi.org/10.1016/j.compedu.2009.06.004

Brante, E. W., & Strømsø, H. I. (2018). Sourcing in text comprehension: A review of interventions targeting sourcing skills. *Educational Psychology Review, 30*(3), 773–799. https://doi.org/10.1007/s10648-017-9421-7

Bråten, I., & Braasch, J.L.G. (2018). The role of conflict in multiple source use. In J.L.G. Braasch, I. Bråten, & M.T. McCrudden (Eds.), *Handbook of multiple source use* (pp. 184–201). New York, NY: Routledge.

Bråten, I., Brante, E., Strømsø, H. (2019). Teaching sourcing in upper-secondary school: A comprehensive intervention with follow-up data. *Reading Research Quarterly*, *54*(4), 481–505. https://doi.org/10.1002/rrq.253

Bråten, I., McCrudden, M. T., Lund, E. S., Brante, E. W., & Strømsø, H. I. (2018a). Task-oriented learning with multiple documents: Effects of topic familiarity, author expertise, and content relevance on document selection, processing, and use. *Reading Research Quarterly*, *53*(3), 345–365. https://doi.org/10.1002/rrq.197

Bråten, I., Stadtler, M., & Salmerón, L. (2018b). The role of sourcing in discourse comprehension. In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *The Routledge handbook of discourse processes* (2nd ed) (pp. 141–166). New York, NY: Routledge.

Bråten, I., Strømsø, H. I., & Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learning and Instruction*, *21*(2), 180–192. https://doi.org/10.1016/j.learninstruc.2010.02.002

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*(4), 485–522. https://doi.org/10.1207/S1532690XCI2004_2

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

Brozo, W. G., Sulkunen, S., Shiel, G., Garbe, C., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement: Lessons from five PISA countries. *Journal of Adolescent & Adult Literacy*, *57*(7), 584–593. https://doi.org/10.1002/jaal.291

Choi, H. J., & Johnson, S. D. (2005). The effect of context-based video instruction on learning and motivation in online courses. *The American Journal of Distance Education,* 19(4), 215–227. https://doi.org/10.1207/s15389286ajde1904_3

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Coiro, J. (2011). Talking about reading as thinking: Modeling the hidden complexities of online reading comprehension. *Theory Into Practice, 50*(2), 107–115. https://doi.org/10.1080/00405841.2011.558435

Coiro, J., Coscarelli, C., Maykel, C., & Forzani, E. (2015). Investigating criteria that seventh

graders use to evaluate the quality of online information. *Journal of Adolescent & Adult Literacy, 59*(3), 287–297. https://doi.org/10.1002/jaal.448

Council recommendation on key competences for lifelong learning. (2018). The Council of the European Union. https://ec.europa.eu/education/education-in-the-eu/council-recommendation-on-key-competences-for-lifelong-learning_en Accessed 12 August 2019.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment, 91*(2), 121–136. https://doi.org/10.1080/00223890802634175

Davey, B. (1983). Think aloud: modeling the cognitive processes of reading comprehension. *Journal of Reading, 27*(1), 44–47.

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford Publications.

Finnish National Board of Education. (2016). *New national core curriculum for basic education: focus on school culture and integrative approach.* https://www.oph.fi/sites/default/files/documents/new-national-core-curriculum-for-basic-education.pdf Accessed 12 February 2020.

Fisher, D., & Frey, N. (2015). Teaching modeling using complex informational texts. *The Reading Teacher*, *69*(1), 63–69. https://doi.org/10.1002/trtr.1372

Forzani, E. (2018). How well can students evaluate online science information? Contributions of prior knowledge, gender, socioeconomic status, and offline reading ability. *Reading Research Quarterly*, *53*(4), 385–390. https://doi.org/10.1002/rrq.218

Forzani, E. (2019). A three-tiered framework for proactive critical evaluation during online inquiry. *Journal of Adolescent & Adult Literacy, 63*(4), 401–414. https://doi.org/10.1002/jaal.1004

Frerejean, J., van Strien, J. L. H., Kirschner, P. A., & Brand-Gruwel, S. (2018). Effects of a modelling example for teaching information problem solving skills. *Journal of Computer Assisted Learning, 34(*6), 688–700. https://doi.org/10.1111/jcal.12276

Gardner, W., Mulvey, E. P., and Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, *118*(3), 392–404. https://doi.org/10.1037/0033-2909.118.3.392

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., et al. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 52*(2), 219–246. https://doi.org/10.1080/00461520.2016.1168741

González-Ibáñez, R., Gacitúa, D., Sormunen, E., & Kiili, C. (2017). NEURONE: oNlinE inqUiRy experimentatiON systEm. *Proceedings of the Association for Information Science and Technology, 54*(1), 687–689. https://doi.org/10.1002/pra2.2017.14505401117

Harju-Luukkainen, H., Vettenranta, J., Ouakrim-Soivio, N., & Bernelius, V. (2016). Differences between students' PISA reading literacy scores and grading for mother tongue and literature at school: A geostatistical analysis of the Finnish PISA 2009 Data. *Education Inquiry, 7*(4), 463–479, https://doi.org/10.3402/edui.v7.29413

Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modelling* (pp. 89–127). Mahwah, NJ: Lawrence Erlbaum.

Hoeken, H. (2001). Anecdotal, Statistical, and Causal Evidence: Their Perceived and Actual Persuasiveness. *Argumentation, 15*(4), 425–437. https://doi.org/10.1023/A:1012075630523

Holopainen, L., Kairaluoma, L., Nevala, J., Ahonen, T., & Aro, M. (2004). *Lukivaikeuksien seulontatesti nuorille ja aikuisille [Dyslexia screening test for youth and adults]*. Jyväskylä, Finland: Niilo Mäki Institute.

Judd, V. C., Farrow, L. I., & Tims, B. J. (2006). Evaluating public web site information: a process and an instrument. *Reference Service Review, 34*(1), 12–32. https://doi.org/10.1108/00907320510631571

Kajamies, A. (2017). *Towards optimal scaffolding of low achievers' learning: Combining intertwined, dynamic, and multi-domain perspectives.* Doctoral dissertation, University of Turku, Finland: Painosalama Oy.

Kammerer, Y., Meier, N., & Stahl, E. (2016). Fostering secondary-school students' intertext model formation when reading a set of websites: The effectiveness of source prompts. *Computers & Education, 102,* 52–64. https://doi.org/10.1016/j.compedu.2016.07.001

Kanniainen, L., Kiili, C., Tolvanen, A., Aro, M., & Leppänen, P. H. T (2019). Literacy skills and online research and comprehension: Struggling readers face difficulties online. *Reading and Writing*, *32,* 2201–2222. https://doi.org/10.1007/s11145-019-09944-9

Kiili, C., Bråten, I., Kullberg, N., & Leppänen, P. H. (2020a). Investigating elementary school students' text-based argumentation with multiple online information resources. *Computers & Education*, *147*. https://doi.org/10.1016/j.compedu.2019.103785

Kiili, C., Brante, E.W., Räikkönen, E., & Coiro, J. (2020b). Citing as a sourcing practice: Students' citing self-selected online sources in their essays. *Journal for the Study of Education and Development.* Advanced online publication: https://www.tandfonline.com/doi/full/10.1080/02103702.2019.1690839

Kiili, C., Leu, D. J., Marttunen, M., Hautala, J., & Leppänen, P. H. T. (2018). Exploring early adolescents' evaluation of academic and commercial online resources related to health. *Reading and Writing, 31*(3), 533–557.

https://doi.org/10.1007/s11145-017- 9797-2

Kingsley, T. L., Cassady, J. C., & Tancock, S. M. (2015). Successfully promoting 21st century online research skills: Interventions in 5th-grade classrooms. *Reading Horizons*, *54*(2), 91–135. https://scholarworks.wmich.edu/reading_horizons/vol54/iss2/5

Kroustallaki, D., Kokkinaki, T., Sideridis, G. D., & Simos, P. G. (2015). Exploring students' affect and achievement goals in the context of an intervention to improve web searching skills. *Computers in Human Behavior, 49,* 156–170. https://doi.org/10.1016/j.chb.2015.02.060

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

Leu, D. J., Forzani, E., Rhoads, C., Maykel, C., Kennedy, C., & Timbrell N. (2015). The new literacies of online research and comprehension: Rethinking the reading achievement gap. *Reading Research Quarterly*, *50*(1), 37–59. https://doi.org/10.1002/rrq.85

Livingstone, S., Mascheroni, G., & Staksrud, E. (2018). European research on children's Internet use: Assessing the past and anticipating the future. *New Media & Society*, *20*(3), 1103–1122. https://doi.org/10.1177/1461444816685930

Macedo-Rouet, M., Braasch, J. L. G., Britt, M. A., & Rouet, J-F. (2013). Teaching fourth and fifth graders to evaluate information sources during text comprehension. *Cognition and Instruction, 31*(2), 204–226. https://doi.org/10.1080/07370008.2013.769995

Macedo-Rouet, M., Potocki, A., Scharrer, L., Ros, C., Stadtler, M., Salmerón, L., et al. (2019). How good is this page? Benefits and limits of prompting on adolescents' evaluation of Web information quality. *Reading Research Quarterly, 54*(3), 299–321. https://doi.org/10.1002/rrq.241

Mason, L., Junyent, A. A., & Tornatora, M. C. (2014). Epistemic evaluation and comprehension of web-source information on controversial science-related topics: Effects of a short-term instructional intervention. *Computers & Education, 76,* 143–157. https://doi.org/10.1016/j.compedu.2014.03.016

Mateos, M., Martín, E., Cuevas, I., Villalón, R., Martínez, I., & González-Lamas, J. (2018). Improving Written Argumentative Synthesis by Teaching the Integration of Conflicting Information from Multiple Sources. *Cognition and Instruction, 36*(2), 119–138. https://doi.org/10.1080/07370008.2018.1425300

McKenna, J. W., Flower, A., & Ciullo, S. (2014). Measuring Fidelity to Improve Intervention Effectiveness. *Intervention in School and Clinic*, *50*(1), 15–21. https://doi.org/10.1177/1053451214532348

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*(2), 139–178. https://doi.org/10.1207/s1532690xci1402_1 .

Metzger, M. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*(13), 2078–2091. https://doi.org/10.1002/asi.20672

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105–125. https://doi.org/10.1037//1082-989X.7.1.105

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338–354. https://doi.org/10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1997). Latent variable modeling of longitudinal and multilevel data. In A. Raftery (Ed.), *Sociological methodology* (pp. 453–480). Boston, MA: Blackwell.

Muthén, L., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed). Los Angeles, CA: Muthén & Muthén.

Paul, J., Macedo-Rouet, M., Rouet, J-F., & Stadtler, M. (2017). Why attend to source information when reading online? The perspective of ninth grade students from two different countries. *Computers & Education, 113,* 339–354. https://doi.org/10.1016/j.compedu.2017.05.020

Pérez, A., Potocki, A., Stadtler, M., Macedo-Rouet, M., Paul, J., Salmerón, L., et al. (2018). Fostering teenagers' assessment of information reliability: Effects of a classroom intervention focused on critical source dimensions. *Learning and Instruction, 58,* 53–64. https://doi.org/10.1016/j.learninstruc.2018.04.006

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

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* New York, NY: Routledge.

Quintana, C., Zhang, M., & Krajcik, J. S. (2005). A framework for supporting metacognitive aspects of online inquiry through software-based scaffolding. *Educational Psychologist, 40*(4), 235–244. https://doi.org/10.1207/s15326985ep4004_5

Shanahan, T., & Shanahan, C. (2012). What is disciplinary literacy and why does it matter? *Topics in Language Disorders*, *32*(1), 7–18. https://doi.org/10.1097/TLD.0b013e318244557a

Sormunen E., González-Ibáñez, R., Kiili, C., Leppänen, P. H. T., Mikkilä-Erdmann, M., Erdmann, N., et al. (2018). A performance-based test for assessing students' online inquiry competences in schools. In S. Kurbanoğlu, J. Boustany, S. Špiranec, E. Grassian, D. Mizrachi, & L. Roy (Eds.), *Information literacy in the workplace. ECIL 2017. Communications in computer and information science*, *810*, 673–682. Springer: Cham.

Stadtler, M., & Bromme, R. (2014). The Content–Source Integration Model: A Taxonomic Description of How Readers Comprehend Conflicting Scientific Information. In Rapp D. & Braasch J. (Eds.), *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences* (pp. 379–402). MIT Press.

Teasley, S. D. (1995). The role of talk in children's peer collaborations. *Developmental Psychology, 31*(2), 207–220. https://doi.org/10.1037/0012-1649.31.2.207

van Gog, T., Verveer, I., & Verveer, L. (2014). Learning from video modeling examples: Effects of seeing the human model's face. *Computers & Education, 72*, 323–327. https://doi.org/10.1016/j.compedu.2013.12.004

van Wermeskerken, M., & van Gog, T. (2017). Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers & Education, 113*, 98–107. https://doi.org/10.1016/j.compedu.2017.05.013

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. A. (2010). Fostering transfer of websearchers' evaluation skills: A field test of two transfer theories. *Computers in Human Behavior*, *26*(4), 716–728. https://doi.org/10.1016/j.chb.2010.01.008

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. A. (2013). Fostering students' evaluation behaviour while searching the Internet. *Instructional Science, 41(*1), 125–146. https://doi.org/10.1007/s11251-012-9221-x

Zhang, S., & Duke, N. K. (2011). The impact of instruction in the WWWDOT framework on student´s disposition and ability to evaluate web sites as sources of information. *The Elementary School Journal, 112*(1), 132–154. https://doi.org/10.1086/660687

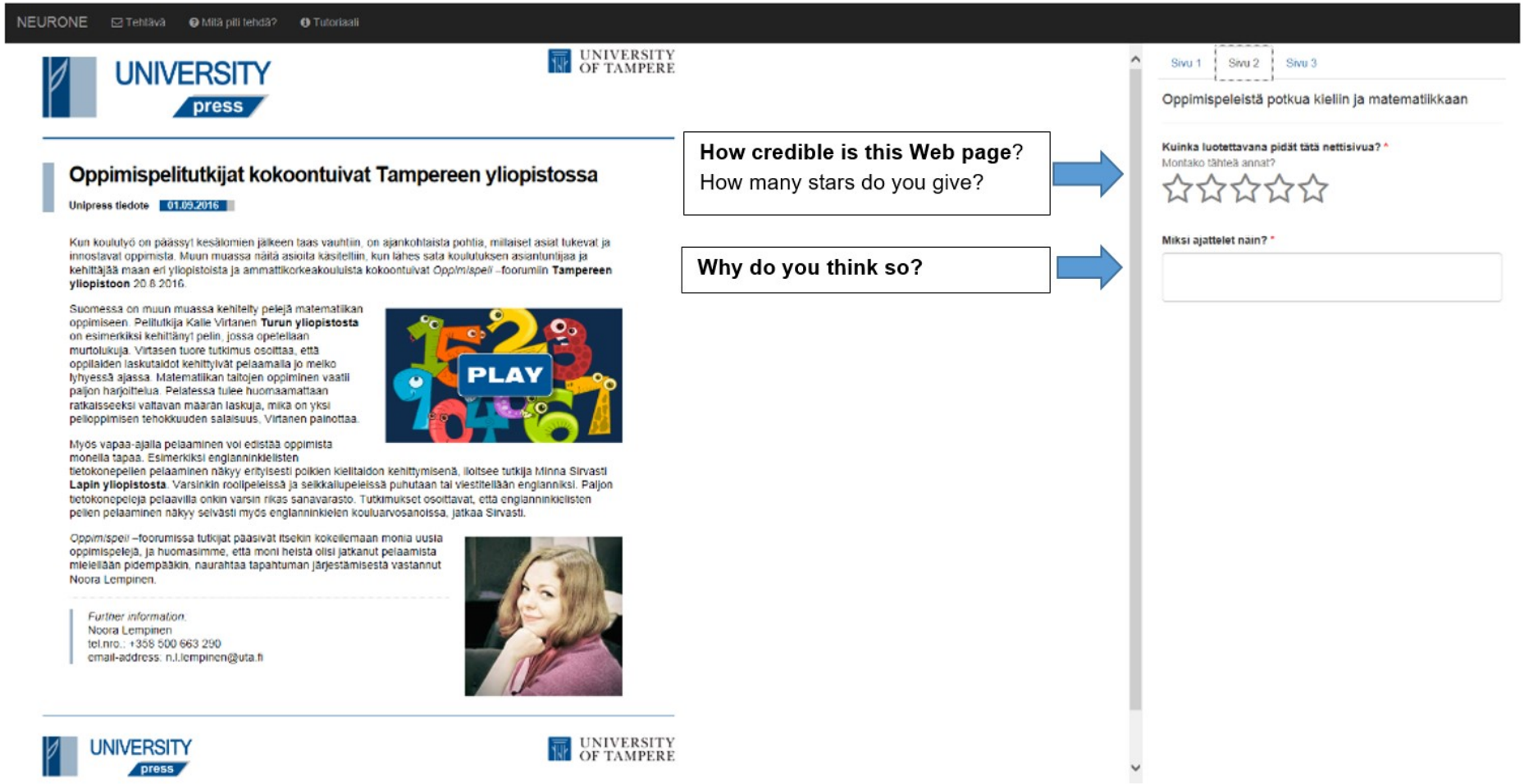
Zhang, W-X., Hsu. Y-S., Wang, C-Y., & Ho, Y-T. (2015). Exploring the impacts of cognitive and metacognitive prompting on students' scientific inquiry practices within an e-learning environment. *International Journal of Science Education, 37*(3), 529–553. https://doi.org/10.1080/09500693.2014.996796

Acknowledgements

The authors thank Michael Freeman for his valuable comments on the language, and the teachers and the students who participated in the study. This research was funded by the Academy of Finland (Project Number: 285817).

Appendix (in colors if possible)

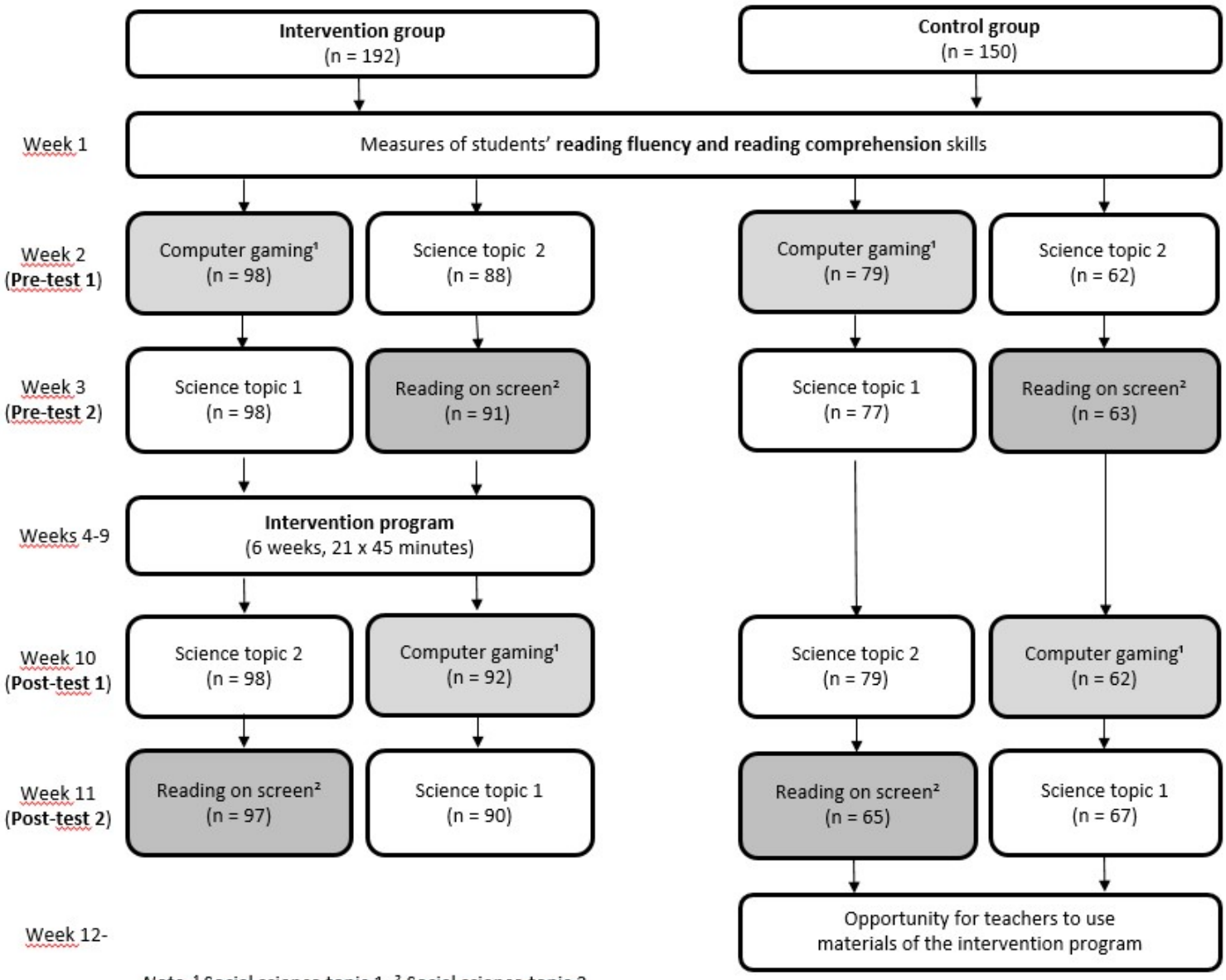
Example of the Web page evaluated by the students and the measure of students' credibility evaluation skills.



Footnotes

[1] Because Science Web pages contained considerably less source features than Social Science Web pages, we excluded Science topics from this study.

[2] To more thoroughly examine the possibility that the intervention effect could have resulted from the lower scores of the control group in Expertise of the Source and Other Source Features, Cohen's *d* for repeated measures was computed separately for the intervention and control groups (Morris & DeShon, 2008). The *d* for Expertise of the Source was 0.35 (95% confidence interval: [0.14; 0.55]) for the intervention group and -0.12 (95% confidence interval: [-0.37; 0.13]) for the control group. The *d* for Other Source Features was 0.32 (95% confidence interval: [0.11; 0.52]) for the intervention group and -0.22 (95% confidence interval: [-0.47; 0.03]) for the control group. As the confidence intervals of the intervention and control groups in either variable do not overlap, it can be concluded that the effect of the intervention group (in Expertise of the Source and Other Source Features) was greater than the corresponding effect of the control group.



Note. ¹ Social science topic 1, ² Social science topic 2

Table 1. Features of the Web pages evaluated by the students.

Topic	Title	Author	Type	Perspective	Position on the Topic
Computer Gaming	Computer gaming has both advantages and disadvantages for health	Pediatrician	Blog by an expert	Health	For and against
	Learning games researchers met at the University of Tampere	University researcher responsible for organizing the event	Press release	Learning	For
	Violent computer games increase hostile behaviour	Journalist	News	Violence	Against
Reading on Screen	Digital book is easy to carry but digital texts are not easily remembered	Journalist	News	Learning	For and against
	Every student should have a chance to read digitally	School principal	Opinion piece in a newspaper	Equality	For
	Excessive screen time is hard on the eyes	Health research center	Official page of an organization	Health	Against

Table 2. Module 1: Credibility evaluation (3 x 45 min lessons).

Lesson	Instructional methods	Materials	Activities	Prompts for students or guiding questions for teachers
Lesson 1	Modeling strategies Analyzing modeled strategies Discussing strategies	Web news article [1]: "New Study: Later school mornings have significant influence on students"	a. Students read the news article and considered its credibility.	
		Video where two virtual students, Tiina and Toni, evaluated the credibility of the Web news article. The students varied in how versatile and sophisticated evaluation strategies they used.	b. Students watched the video that modelled strategies and analyzed these strategies with the worksheet.	Prompts in the worksheet: 1. What evaluation strategies the virtual students used in the video? (identification from the list)
			c. Students discussed their answers with a partner / partners.	2. Compare how Tiina and Toni evaluated the news article. Which of them did it more thoroughly? Why do you think so?
		Worksheet [1]	d. A teacher-led discussion on the evaluation strategies and the content of the article.	3. What new ways to evaluate Web pages' credibility you learned?

Lessons 2

Table 3. Categories of students' justifications.

Main category	Sub-category	Description	Example	Interpretation
Source Features	Expertise of the Source	Student justifies his or her credibility rating with authors'/publishers' credentials, affiliation, experience or expertise.	*I think this Web page is credible because it includes the ideas of a pediatrician.* (Student 279, topic: Computer gaming)	Being a pediatrician is a credential of the author/publisher (page is a blog by an expert).
			The Web page has been created by a health research center. (Student 292, topic: Reading on screen)	The health research center is an affiliation of the author and publisher.

- How did they talk about the energy drinks?
- Why is it important to compare information from different Web pages?
- Why is it important to evaluate Web pages in different ways?
- How can you assure which Web page contains information that is more credible?

	Learning diary [1]	c. Students individually evaluated their learning in their diaries.	Prompts in the students' diaries: - questions fostering self-evaluation of learning and reflection of learning experience

Note: [1] on paper, [2] digital, [3] Web pages were designed for assessment purposes (Kiili et al., 2018)

Table 4. Descriptive statistics for the variables and their correlations (N = 316-342).

Variables (observed range)	Mean *(SD)* / %	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
Dependent variables														
1. Post-test: Expertise of the Source (0 - 9)	1.84 (2.05)	**.23****	**.18****	**-.17****	**.51****	**.28****	**.18****	-.14*	**.32****	**.30****	-.01	-.10	-.01	.09
2. Post-test: Other Source Features (0 - 6)	0.56 (0.95)		.04	-.05	.15*	**.34****	.06	-.05	**.19****	**.15****	-.07	**-.16****	-.05	.09
3. Post-test: Argumentation in the text (0 - 4)	0.29 (0.64)			-.02	.10	.10	.11	-.04	**.19****	**.20****	**-.24****	-.09	**-.24****	-.01
4. Post-test: Other aspects of the content (0 - 3)	0.39 (0.72)				-.12*	.03	-.02	**.32****	.01	.02	-.01	-.03	-.01	-.03

Control variables				
Quality of Content	in the Text	(negative and positive) of an issue or argumentative purpose of the text on the Web page.	*Altogether 1492 adolescents participated in that study lasting over 4 years.* (Student 206, topic: Reading on screen)	Student refers to the research-basis and quality of evidence on the Web page.
	Other Aspects of the Content	Student justifies his or her credibility rating with correspondence with his/her own experiences or prior knowledge, or amount of the text or writing style on the Web page.	*Gaming affects one's fitness and health, I have noticed it also myself.* (Student 103, topic: Computer gaming)	Student refers to the correspondence with her own experiences.
			I think this Web page is quite credible because its' style of writing is correct. (Student 301, topic: Reading on screen)	Student refers to the writing style on the Web page.

Table 5. Pre-test, post-test and reading test scores in the intervention and control group.

		Intervention group (N = 189-191)	Control group (N = 127-143)								
5.	Pre-test: Expertise of the Source (0 - 10)	1.56 (1.93)		**.27****	**.21****	**-.23****	**.34****	**.16****	**.21****	-.11*	**.22****	-.13*
6.	Pre-test: Other Source Features (0 - 8)	0.52 (1.05)		.13*	.13*	-.07	**.25****	.13*	.11	-.11*	.13*	-.18*
7.	Pre-test: Argumentation in the text (0 - 4)	0.33 (0.71)				-.07	**.16****	**.16****	**.20****	-.11	**.16****	-.07
8.	Pre-test: Other aspects of the content (0 - 5)	0.39 (0.79)					-.08	.11*	.10	.01	.07	-.02
9.	Reading fluency (10 - 89)	38.32 (14.00)						**.40****	-.00	**-.33****	.03	-.07
10.	Reading comprehension (0 - 16)	8.81 (3.62)							-.09	**-.22****	-.09	.04
11.	Topic order (screen - gaming / gaming - screen)	48.37/51.63								-.09	**.88****	-.00
12.	Gender (female / male)	48.25/51.75									-.06	.03
13.	Test order in pre-test (first / second)	53.60/46.40										-.03
Independent variable												
14.	Group (intervention / control)	56.14/43.86										

Note: *= p < .05; **= p < .01

Pre-test measures	*M*	*SD*	*Md*	*M*	*SD*	*Md*	Test result
Expertise of the Source	1.35	1.75	0	1.85	2.17	1	$U = 11936.50$, $p = .066$, $r = -.10$
Other Source Features	0.35	0.65	0	0.74	1.40	0	$U = 12041.00$, $p = .048$, $r = -.11$
Argumentation in the Text	0.29	0.66	0	0.39	0.78	0	$U = 12842.50$, $p = .355$, $r = -.05$
Other Aspects of the Content	0.38	0.81	0	0.40	0.77	0	$U = 13002.50$, $p = .522$, $r = -.04$
Reading test measures							
Reading fluency	37.48	13.52		39.45	14.57		$t(332) = 1.28$, $p = .203$, $d = 0.14$
Reading comprehension	8.93	3.73		8.65	3.47		$t(331) = -0.71$, $p = .480$, $d = 0.08$
Post-test measures							
Expertise of the Source	1.98	2.10	2	1.62	1.98	1	
Other Source Features	0.63	0.93	0	0.46	0.97	0	
Argumentation in the Text	0.29	0.66	0	0.30	0.61	0	
Other Aspects of the Content	0.37	0.68	0	0.41	0.77	0	

Note. The effect of an intervention cannot be calculated based on this table as the background variables are not controlled for.

Table 6. Results of the multilevel negative binomial regression analysis for the association between control variables, independent variable (group) and post-test measures.

	Post-test measures			
	Expertise of the Source	Other Source Features	Argumentation in the Text	Other Aspects of the Content
	IRR* 95% CI	IRR* 95% CI	IRR* 95% CI	IRR* 95% CI
Pre-test: Expertise of the Source	**1.28 [1.19; 1.36]**			
Pre-test: Other Source Features		**1.39 [1.25; 1.55]**		
Pre-test: Argumentation in the Text			**1.35 [1.05; 1.74]**	
Pre-test: Other Aspects of the Content				**1.74 [1.37; 2.20]**
Reading fluency	1.01 [1.00; 1.02]	1.01 [0.99; 1.02]	1.02 [1.00; 1.04]	1.01 [0.98; 1.03]
Reading comprehension	**1.06 [1.03; 1.10]**	1.02 [0.97; 1.07]	**1.09 [1.02; 1.15]**	0.98 [0.92; 1.04]
Gender (0 = female, 1 = male)	1.00 [0.77; 1.30]	0.69 [0.45; 1.05]	0.87 [0.55; 1.38]	0.92 [0.66; 1.29]
Topic order (0 = screen-gaming, 1 = gaming-screen)	1.01 [0.51; 2.01]	0.65 [0.40; 1.06]	**0.43 [0.20; 0.90]**	0.74 [0.25; 2.21]
Test order in the pre-test (0 = second, 1 = first)	0.85 [0.45; 1.61]	1.16 [0.73; 1.82]	0.65 [0.31; 1.35]	1.21 [0.50; 2.93]
Group (0 = control group, 1 = intervention)	**1.52 [1.15; 2.00]**	**1.83 [1.29; 2.60]**	1.01 [0.67; 1.53]	1.00 [0.57; 1.74]

* IRR = incident rate ratio (IRR differs statistically significantly from value 1 if its confidence interval does not include value 1)

Table 7. Students' use of justifications for credibility in their written products.

	Category of justifications used in the written product	Pre-test			Post-test			
		M	*SD*	*Md*	*M*	*SD*	*Md*	Wilcoxon test result (*Z*) Effect size (*d*)
Intervention group (*N* = 189)	Source Features	0.04	0.22	0	0.08	0.29	0	*Z* = 1.56, *p* = .120 *d* = 0.13, 95% *CI* = [-0.08; 0.33]
	Quality of Content	0.05	0.29	0	0.07	0.28	0	*Z* = 0.97, *p* = .333 *d* = 0.07, 95% *CI* = [-0.13: 0.28]
Control group (*N* = 143)	Source Features	0.06	0.23	0	0.02	0.13	0	*Z* = -1.67, *p* = .096 *d* = -0.12, 95% *CI* = [-0.37; 0.13]
	Quality of Content	0.06	0.23	0	0.07	0.26	0	*Z* = 0.24, *p* = .808 *d* = 0.00, 95% *CI* = [-0.25; 0.25]