Fostering Teenagers' Assessment of Information Reliability:

Effects of a Classroom Intervention focused on Critical Source Dimensions

Word count: 11060

Abstract

Increased amounts of information available from the Internet have triggered new demands for students to evaluate information quality. Our study presents an instructional intervention aimed at fostering ninth grade students' critical evaluation of source reliability. The intervention was grounded into theories of multiple text comprehension and used an analytic framework that defines the core source dimensions of author position (competence), author motivation (intention), and media quality (pre-publication validation). Compared to controls, trained students 1) reduced the score assigned to links containing less reliable information in the three critical source dimensions (knowledge application task), as well as 2) increased the number of references made to a more reliable source (e.g., "*scientific journal*") and decreased references made to a less reliable source (e.g., "*personal blog*"), in a task presenting contradictory information across texts (transfer task). Nonetheless, the intervention outcomes varied according to the type of source evaluation question. We discuss the beneficial effects of implementing classroom intervention on sourcing skills as a means to improve teenagers' critical thinking when comprehending multiple documents.

Keywords: *sourcing skills; classroom intervention; multiple-document comprehension; information reliability; critical thinking*

1. Introduction

In today's world, and particularly since the advent of the Internet, large amounts of information are made available to lay readers. A growing number of adolescents use the Internet to search for information (Eurostat, 2015), either for school or out-of-school purposes (Wartella, Rideout, Zupancic, Beaudoin-Ryan, & Lauricella, 2015). Most teenagers experience difficulties with the comprehension and use of multiple sources of information (e.g., Britt & Aglinskas, 2002; Walraven, Brand-Gruwel, & Boshuizen, 2009). Thus, it is theoretically and educationally important to design effective instructional procedures aimed at fostering teenagers' awareness and use of key dimensions of information quality.

Theories of multiple document comprehension (Perfetti, Rouet, & Britt, 1999) suggest that in some circumstances, readers construct an "intertext model", that is, a representation of the text contents linked to their respective sources. Among other factors (Bråten, Strømsø, Britt, & Rouet, 2011), readers are more likely to link contents to sources when they notice contradictions across texts dealing with the same topic (i.e., the Discrepancy-Induced Source Comprehension or D-ISC effect; Braasch, Rouet, Vibert, & Britt, 2012; see also Braasch & Bråten, 2017). Experiments involving college students have demonstrated that readers of discrepant stories cite information sources more often when summarizing the stories and recall more source information afterwards (Braasch et al., 2012; Rouet, Le Bigot, de Pereyra, & Britt, 2016), suggesting an attempt to restore coherence by integrating source information (see also Kammerer & Gerjets, 2014; Kammerer, Kalbfell, & Gerjets, 2016). Further research has provided details about the processes whereby readers come to integrate sources and contents from conflicting texts. The Content-Source Integration model (CSI, Stadtler & Bromme, 2014) posits that content-source integration involves three processing stages: 1) the detection of a conflict, 2) the regulation of that conflict (e.g., attributing the contradiction to the existence of different sources), and 3) the resolution of the conflict (e.g., evaluating

information reliability). Importantly, readers may restore coherence among conflicting texts by acknowledging that contradictory information comes from distinct sources. In a final stage, readers try to resolve the conflict. This may be accomplished by judging the relative truth value of the statements based on one's own understanding of the subject matter (first-hand evaluation). Stadtler and Bromme (2014) argue that when prior knowledge is low, in particular, readers may rather evaluate the trustworthiness of the sources (second-hand evaluation).

Thus, converging theoretical arguments suggest that the comprehension of multiple-texts demands not only understanding the content of the texts, but also (and especially in the presence of contradictory content) indexing contents onto their respective source, and evaluating source information. Interestingly, research on collaborative discourse and argumentation has shown that content integration improves by prompting elaborative and metacognitive thinking (e.g., through explanations; Nussbaum, 2008), an effect which may extend to content-source integration. However, in order to integrate and evaluate source information, readers have to be able to identify and assess several dimensions of the sources.

1.1. Dimensions of source evaluation

Broadly speaking, the construct of source may be defined as the origin of a message, or as Rieh (2002) puts it, "where the document comes from". A source can be characterized along a number of dimensions regarding the author, the context of production and the communication channel of the document (Britt, Perfetti, Sandak, & Rouet, 1999; Goldman & Scardamalia, 2013; Scharrer & Salmerón, 2016). Research on document comprehension has defined readers' ability to source (or *sourcing*; Wineburg, 1991) as any mental process directed to (explicitly or implicitly) pay attention to, evaluate, integrate, memorize and/or make a decision by using source information (e.g., Bråten, Stadtler, & Salmerón, in press; Scharrer & Salmerón, 2016).

Britt and Aglinskas (2002) contributed an early attempt to categorize relevant source features. Taking into account previous justification responses provided by experts (Rouet, Britt, Mason, & Perfetti, 1996), they identified several source dimensions regarding author information (e.g., position or motivation) and document information (e.g., type or publication date). Interestingly, after offering a short source training workshop to secondary students, Britt and Aglinskas (2002) found that the best recalled source dimension was author position (82%), demonstrating that author's occupation, profession or credentials (e.g., "*professor*") are critical features for source evaluation. Author position influences readers' assessment of text content, whereby information provided by an expert is considered as more accurate than information provided by a non-expert (e.g., Winter, Kramer, Appel, & Schielke, 2010).

Besides the position of the author, their intentions or benevolence also affect experienced readers' appraisal of information quality. Author's motivation seems of particular importance when potentially bad intentions or meanness are perceived (e.g., Porsch & Bromme, 2011). Similarly, texts with historical controversy coming from sources presenting a conflict of interest are evaluated as less trustworthy by both expert and novice readers (Rouet, Favart, Britt, & Perfetti, 1997). Thus, understanding an author's motivation is also important to determine whether the information is (un)reliable.

In addition to author dimensions, Rieh (2002) pointed out the importance of document features such as document type, reputation, or URL. In a study of Web users' evaluation of information, Rieh found that the type of media was present in 28% of users' judgement of information quality, with document type (e.g., "*governmental homepage*") and document reputation (e.g., "*the well-known US-based centre for disease control*") being the most frequently cited features. More recent work has confirmed that young adults make an effort to find reliable websites, even when a Web search engine results page shows unreliable websites at the beginning of the list or in a randomized order (e.g., Kammerer & Gerjets, 2014).

Website reliability largely rests on the existence of an independent process for controlling the validity of information prior to its publication. Such a process is a hallmark of scientific publication, and it is typically lacking in personal publishing (e.g., blogs), or collaborative media (e.g., online encyclopedia). Understanding the level of pre-publication quality control is key to evaluating media quality and arguably an important dimension of readers' ability to critically assess the quality of information.

To summarise, the research to date suggests that three sourcing dimensions are critical to readers' assessment of information quality, namely: author position ("*who says what*"), author motivation ("*why the author says it*") and media quality ("*where it is published*"). In looking for ways to foster teenagers' critical literacy skills, it is important to find out what they know about sources and what kind of instructional intervention have the most significant impact on their strategies.

1.2. Fostering students' source evaluation skills

There is ample evidence that experts spontaneously evaluate sources when reading multiple documents and/or searching for information online (Hilligoss & Rieh, 2008; Rieh, 2002; Rouet, et al., 1997; Wineburg, 1991). In contrast, a growing number of studies have reported that primary and secondary students do not spontaneously evaluate source information when reading (e.g., Braasch, Bråten, Strømsø, Anmarkrud, & Ferguson, 2013; Britt & Aglinskas, 2002; Eastin, Yang, & Nathanson, 2006). Evaluating source information may be seen as a form of critical thinking, a broader competency that also includes the ability to identify, clarify, and solve a problem through the evaluation and judgement of validity and reliability of claims (e.g., Kennedy, Fisher, & Ennis, 1991; Pithers & Soden, 2000). Based on this view, a few intervention studies have attempted to develop students' source awareness and source evaluation skills, as a way to promote their critical thinking about information.

For example, Macedo-Rouet, Braasch, Britt and Rouet (2013, Experiment 2) provided a short classroom intervention (30 minutes) to instigate the evaluation of author position among 4th and 5th graders. Students read a text in which two people (amateur vs. expert) presented discrepant arguments on the same topic, and they were prompted to answer source questions. Subsequently, the concept of "being knowledgeable" (expertise) was discussed and, to conclude, students were asked to answer 1) how they could solve the conflict of the story, and 2) which of the characters could be right and why. Overall, trained students were more aware and made more use of author position (professional status and training in the domain) than control students, demonstrating the ability of young children to learn sourcing skills. Similar beneficial training effects have been found with vocational students dealing with scientific controversies (Stadtler, Scharrer, Macedo-Rouet, Rouet, & Bromme, 2016).

Moreover, Braasch, et al., (2013) carried out another brief intervention (60 minutes) to improve high-school students' ability to evaluate source features in multiple documents. Trained students were presented with two contrasting cases: a) a poor protocol, where there was no reference to source features and distinctions about document's trustworthiness were related to content information; and b) a better protocol, where more sophisticated knowledge about source information and trustworthiness was provided (i.e., evaluation of the author, venue, type and date of publication and its relation to the cognitive authority of the documents). The control group followed regular classroom instruction. Students trained with the contrasting cases method evaluated trustworthiness more often and were more likely to assign this aspect to source features than control students, confirming the benefits of their instructional intervention. Interestingly, these results reflect that the cognitive mechanisms necessary to evaluate author and document source features are readily available to teenagers.

Finally, an instructional unit called Source, Evidence, Explanation and Knowledge (SEEK) successfully improved higher education students' sourcing skills (Wiley et al., 2009).

Wiley et al. divided the intervention in three steps: 1) declarative knowledge to evaluate media quality; 2) the implementation of that knowledge with websites including different types of information quality (e.g., sites from official institutions or personal pages); and 3) feedback after media evaluation. Trained students were better than controls at ranking reliable and unreliable websites. The effects were replicated with ninth grade students (Mason, Junyent, & Tornatora, 2014), and non-university educated adults (Kammerer, Amann, & Gerjets, 2015).

In conclusion, a fruitful approach to train students' source evaluation skills seems to combine: a) the identification of conflicting information dealing with the same topic, b) explicit declarative knowledge about source features and/or dimensions, and c) intensive practice, including comparative cases and informative feedback. Indeed, a meta-analysis has shown that class discussions, problem solving with authentic tasks, and mentorship are relevant aspects to train critical thinking (e.g., Abrami, et al., 2015). Importantly, general abilities such as reading skills and working memory capacity may influence multiple document comprehension (see e.g., Rouet & Britt, 2011). Therefore, intervention studies should make sure to control potential interactions with these factors.

1.3. Rationale and hypotheses of the present study

Past research suggests that students' sourcing skills can be successfully improved by implementing adequate interventions. Thus far, however, the scope of these studies has been restricted in three ways: 1) interventions have generally focused on one or few source dimensions (e.g., Macedo-Rouet et al., 2013; Wiley et al., 2009) and/or they have trained these dimensions as a whole (e.g., Braasch, et al., 2013), ignoring the possible benefits of teaching source dimensions and their interactions more systematically; 2) most published school intervention studies on source evaluation have used just one training session and an immediate post-test, leaving apart possible longer-term training effects. A similar limitation

has also been reported in a recent meta-analysis with interventions on critical thinking (Nordheim, Gundersen, Espehaug, Guttersrud, & Flottorp, 2016); and 3) although some of the previous studies have assessed source evaluation through multiple measures (e.g., essay writing, requiring evaluations, ranking texts usefulness, or justifying decisions), these assessments have been mainly based on a single multiple document task, overlooking possible differences in the intervention effectiveness as a function of the task format and demands.

The first goal of our study was to assess the effects of a theory-based intervention program, which involved the systematic introduction of source characteristics that have been found to play a part in skilled multiple text comprehension. More specifically, we designed a program that focused on the interplay of author position ("*who says what*"), author motivation ("*why the author says it*") and media quality ("*where it is published*"). Based on prior work (e.g., Braasch, et al., 2013; Macedo-Rouet et al., 2013; Wiley, et al., 2009), we created a series of classroom workshops involving a) group discussion about a situation presenting contradictory information, b) introduction of declarative knowledge about one of the three key source dimensions, and c) practice exercises with contrasting items (i.e., the same topic presented by two different sources) and lists of document descriptions to be rated according to source dimensions. In addition, oral discussion provided reflections about possible interactions between dimensions, as well as other possibly confounded constructs (e.g., author competence vs. personal experience).

Our second goal was to characterize precisely the scope of our intervention by using a range of criterion tasks involving both knowledge application and knowledge transfer. In the knowledge application task, students were asked to rate from 0 (certainly not) to 4 (certainly) whether they would use several document references (links) that were contrasted on the three source dimensions. For example, some of the documents were written by experts, whereas others were written by lay persons. This task involved knowledge application because

students were given ample opportunities to rate documents using the same criteria in the course of the three training workshops (see section 2.2).

We also designed an additional task in order to find out whether students could transfer the knowledge they had acquired during the workshops to novel situations. The transfer task required students to read two texts presenting conflicting information about a given topic. One text was attributed to a highly reliable source and the other to a less reliable source. In the near transfer question, students were asked to tell which text was best and to write a short justification for their choice. This question was similar to those used for group discussion during the workshops and thus required near transfer of knowledge. In the far transfer question, the students were asked to write a short conclusion from the two texts. Our purpose was to find out if trained students would use source information as a means to integrate discrepant information (Stadtler & Bromme, 2014). The latter question was never practiced during the workshops and could thus be considered farther transfer of knowledge.

Finally, besides the pre-test phase, we included two post-test phases to assess the impact of the intervention in the short term (one week after the last training session) and longer term (three weeks after), and we measured individual differences in reading comprehension and working memory capacity. We expected that the inclusion of a second post-test would add information regarding the effectiveness of our intervention, whereas the control of individual differences would rule out group differences due to general abilities.

As a general hypothesis, we expected trained students to become appreciative of more reliable sources and less appreciative of less reliable sources as evidenced in the knowledge application and transfer tasks.

As regards the knowledge application task (i.e., rate from 0 to 4 whether you would consult each document), we expected all students to rate the sources with more reliable

features more positively than those with less reliable features. However, we hypothesized that trained students would become more critical of ambiguous sources (i.e., those with reliable features on some, but not all dimensions), reflecting their increased ability to scrutinize the source descriptions and to correctly infer information reliability based on a combination of all relevant source dimensions.

As regards the transfer task, in the near transfer question (i.e., select the best of two texts and justify the selection), based on prior studies using similar questions (e.g., Salmerón, Macedo-Rouet, & Rouet, 2016) we expected trained students to be more likely to select the source with the best credentials (more reliable source). As a consequence, trained students should also increase their use of specific source features (e.g., "*the researcher*" or "*the scientific journal*") when justifying their selection of the best document at the post-tests. This finding would indicate a deeper attention to and evaluation of source information when assessing information quality. In the same vein, we expected the far transfer question (i.e., write a short conclusion from two contradictory texts) to yield similar outcomes (more reference to source by trained students in their written conclusions at the post-test), but with perhaps less frequent and vaguer references to the sources since the task did not explicitly require the students to assess the respective qualities of the documents.

Finally, since the training comprised several workshops with contents repeated and consolidated across sessions, we expected that most of the effects would extend across a delay of three weeks (post-test 2). Nonetheless, taking into account the continuum of difficulty related to the three assessment questions, we predicted longer-term effects to be more apparent in the knowledge application compared to the transfer questions.

2. Methods

2.1. Participants

One hundred and eighty-nine students attending one of two ninth-grade classes in each of four different public French secondary schools participated in the study. The purposes and the design of the study were discussed with the school principals and the teachers prior to the intervention. The two classes within the same school were randomly assigned to a trained or a control condition. Participants included in the data analysis met the following criteria: 1) being a native French speaker; 2) no outlier (low) score in the reading comprehension task (see Materials below); 3) participate in the pre-test and post-test phases for both groups; and 4) participate in all training sessions for the trained group. The final sample included 137 students: 73 controls (M = 14.70 years old, range 13-16; 35 girls) and 64 trained (M = 14.73 years old, range 13-17; 41 girls).

2.2. Materials and procedure

The procedure included a total of six one-hour sessions. The first, fifth and sixth sessions (pre-test, post-test 1 and post-test 2, respectively) served to assess students' performance on the sourcing skills tasks, as well as their comprehension and memory skills. Post-test 1 was administered one week after the training, while post-test 2 was run three weeks after the training. The second, third and fourth sessions were used for the training workshops in the trained group and for regular classroom teaching in the control group. The control group also received a short introduction to the core constructs of sourcing in the final session. In the next section, we first describe the sourcing skills training, then the sourcing skills assessment, and finally, the comprehension and memory skill measures that were used to control for group equivalence.

2.2.1. Contents of the classroom sessions in the trained and control groups

Three training sessions of one-hour were administered to the trained classes during regular class sessions, normally devoted to various content areas (i.e., science/technology, history/sociology, or language). The training workshops were run by two members of the

research team (one leader and a helper) together with the teacher, who introduced the researchers at the beginning of each session and helped answering students' questions.

In order to ensure symmetry in the topics dealt with in both groups during the class periods, the training materials were designed to match the topics taught by each teacher in their respective classes. For example, "*The scientific arguments not mentioned by [a dairy product]*" helped students to reflect about nutrition, a topic covered in the regular science class, whereas "*The liberation of France during the Second World War*" provided discussions about World War II, a regular history topic at this grade level. Thus, although the control classes did not receive any explicit instruction on source evaluation, they received regular instruction in the same classroom environment, in the same content areas and by the same teacher as the trained classes. This control group allowed us to compare the sourcing abilities acquired during typical content area instruction, and those acquired in similar classroom conditions through our training program. It should also be noted that both groups got a chance to meet with the researchers during the pre-test and the two post-test sessions. Finally, students in both groups were informed that the results of the pre- and post-tests would not be included in the final marks.

Each training workshop was mostly dedicated to one source dimension (see example tasks in the Supplementary Materials, thereafter SM). In addition, the sessions included reminders of previously covered contents as well as tasks requiring a combination of the source dimensions introduced in previous workshops. Consistent with the D-ISC effect (Braasch et al., 2012) and the CSI model (Stadtler & Bromme, 2014), the workshops were largely based on the presentation and discussion of short documents providing conflicting views on a given topic.

Workshop 1: Author position. The students were welcomed by the teacher and sat at their usual place in the classroom. They were given a 5-page booklet containing various

materials and practice tasks. After a brief statement of objectives, the researcher leading the session projected a simulated Web page featuring a short text about the positive effects of brain doping. The Web page was also presented on a page of the student booklet. Students were invited to read the text and to decide whether, according to this web page, brain doping a) was effective ("*Yes*"), and b) had any negative side-effects ("*No*"). Subsequently, students were presented with a second Web page highlighting the lack of benefits and the negative side-effects of brain doping, and they answered the same questions again. This procedure allowed all students to notice the contradiction between the texts. The session leader then asked the students how they could resolve the contradiction. After some group discussion, some students pointed out that one author was a "*neurologist*" (female), while the other was a "*stewardess*". The session leader then built on their comments to introduce the construct of author position, by explaining that whereas one source was competent in the domain, the other was non-competent in the domain but competent in another domain (see SM, Figure 1). The explanation emphasized authors' profession and training in the domain (i.e., years of education and specialization). The students were then given a practice task in which they were asked to rate the competence of various authors with respect to a given statement on a 5-point Likert scale (0-4). The materials took the form of contrasted cases. For instance, "*A mother of four children/a pediatrician explains that cow milk is harmful for babies' health*". The contrasted cases were used to disentangle the construct of author position from related constructs such as personal experience, which are often influential in teenagers' assessment of information quality (Salmerón et al., 2016). The students were invited to share their ratings with the class, and potential discrepancies were discussed. The reasons why a layperson should be rated low on a scale of author position, despite having substantial experience, were discussed and justified with the class until consensus was reached (which took a few seconds to a few minutes in all three experimental classes).

Workshop 2: Author motivation. In the second training session, the lead researcher presented again two Web pages in which two sources presented conflicting information about the effectiveness of a dairy product promoted as "good" for people's health. Both sources were competent in the domain, but one involved a conflict of interest ("*a nutritionist in charge of the dairy firm's laboratory*"), whereas the other did not ("*a research director of European food and agriculture agency"*). This situation brought the question of which of the two sources was more reliable, promoting again a group discussion (see SM, Figure 2). The new construct of author motivation was then explained by referring to the explicit conflict of interest (commercial interest) that a specific source might have. In the rest of the session, students were assigned a series of short practice tasks involving pairs of contrasting items (e.g., "*A Facebook founder/sociologist explains that social networks can reduce isolation and depression among adolescents*"). Once more, students were asked to rate the trustworthiness of the information presented by the author. They were invited to share their ratings. Discrepancies were discussed, which enabled the session leader to further disentangle the constructs of author position and author motivation.

Workshop 3: Media quality. In the final training session, the students were shown a simulated forum about the same dairy product as in the previous session, where a self-proclaimed expert (a contributor introducing himself as a "*doctor*") advocated the health benefits of this product. Students were invited to discuss whether and why the information could be untrustworthy. The construct of pre-publication validation of information was then introduced (see SM, Figure 3). We explained the difference between websites where information is validated before publication (e.g., academic journals or magazines), and websites where information is validated only after publication at best (e.g., blogs or forums). The session also included a practice based on contrasting cases (e.g., "*A website from the Health Ministry/a forum*"), in which students rated to what extent the information contained

in the item had been validated before its publication. Once more, ratings were shared and discussed until the class reached consensus.

The third session ended with a summary of the three source dimensions and a comment about the need to consider all three dimensions in combination when searching the Web for reliable information. An additional comment suggested that checking information quality was especially important for search activities that may involve high stakes (in terms of money, health, safety and so forth).

2.2.2. Sourcing skills assessment tasks

As mentioned earlier (see section 1.3), three questions were created to assess students' ability to evaluate source information. They were always administered at the beginning of each assessment phase, and they took approximately 25 mins.

Knowledge application task. Students were instructed to imagine that they had to search for information about a specific topic (e.g., "*The period of world history called Cold War*"). Subsequently, they were asked to rate on a Likert scale from 0 (certainly not) to 4 (certainly), whether they would consult or not several links that resulted from the search (see Figure 1). Author position, author motivation and media quality were manipulated to create four conditions or types of links: a) "good", links containing reliable features on all three dimensions; b) "fair", links with reliable features on two dimensions (e.g., information given by a competent and trustworthy author but in a non-validated media); c) "poor", links with a single reliable feature (e.g., competent author but with a conflict of interest and in a non-validated media); d) "bad", links with unreliable features on all three dimensions. In addition, the list included a filler link containing relevant keywords but referring to a different topic. The dependent variable in this task was the mean normalized rating score (from 0 to 1) for each link. We created three versions dealing with different topics ("*The period of world history called Cold War*"; "*Opinion polls in democratic societies*"; and "*The health risks of*

electromagnetic waves emitted by mobile phones"). To prevent potential biases due to topic knowledge or interest, these versions were counterbalanced across phases and schools, always keeping the same version in the control and trained groups of the same school.

P M Mq	*For each link, select one digit between 0 and 4*	Certainly not				Certainly
Fair − + +	The article of a film student about the causes of the Cold War, published on the website of an academic journal.	0	1	2	3	4
Good + + +	A political analyst who describes the main events of the Cold War on the official website of the United Nations.	0	1	2	3	4
Fair + − +	A communist militant historian examines the origins of the Cold War on Le Monde Diplomatique newspaper's website.	0	1	2	3	4
Poor − + −	The blog of a student's parent offering a text of his personal reflections about the Cold War.	0	1	2	3	4
Bad − − −	The owner of a company that sells historical videos gives his opinion about the Cold War on his personal website.	0	1	2	3	4
Filler	A site about the social and economic consequences of cold for people in countries at war.	0	1	2	3	4
Poor + − −	A discussion forum where a specialised journalist is promoting his documentary and he analyses the causes of the Cold War.	0	1	2	3	4
Good + + +	The website of the Ministry of Education which provides a record of the Cold War written by a professor of History.	0	1	2	3	4
Bad − − −	A seller who sells history books and comments about the Cold War on the Facebook page of its online bookstore.	0	1	2	3	4

Figure 1. Example of a list of links presented in the knowledge application task. The grey column shows the four conditions (good, fair, poor and bad), created from the dimensions of author position (P), author motivation (M) and media quality (Mq). This column is presented here for illustration purposes and was not shown to the participants.

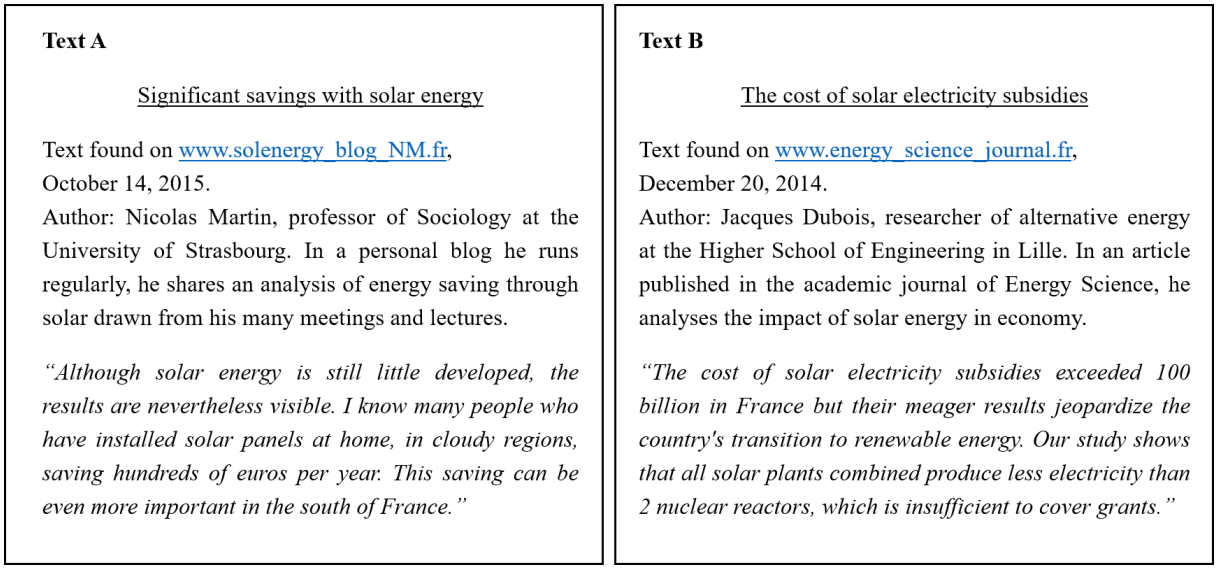
A pilot study was carried out to control for general differences between the three versions of the knowledge application task. The three versions were completed by 12 students (M = 14.59 years old; range: 13-16) who did not take part in the main study. A linear mixed-effects model with Participants and Items as random factors and Version as fixed factor (see section 2.3 for more information about linear mixed-effects models), was ran on the mean normalized rating score obtained in the Likert scale. The model showed no significant effect

of version, χ^2 (2, N = 12) = 0.40, p = .82, where none of the three versions differed from the other two (all *ps* > .85): war (M = 0.50, SE = 0.63), polls (M = 0.53, SE = 0.63), and mobiles (M = 0.50, SE = 0.63). These results supported the equivalence of the versions. In addition, the same model including condition (type of link) as the fixed factor demonstrated a significant main effect, χ^2 (3, N = 12) = 169.49, p < .001, dv = 10.42, where rating differed according to the type of links (bad: M = 0.13, SE = 0.06; poor: M = 0.39, SE = 0.06; fair: M = 0.63, SE = 0.06; and good: M = 0.81, SE = 0.05): bad vs poor, *t.ratio* (39) = 3.65, p < .01; bad vs fair, *t.ratio* (22) = 6.44, p < .001; bad vs good, *t.ratio* (118) = 12.06, p < .001; poor vs fair, *t.ratio* (21) = 3.01, p = .03; poor vs good, *t.ratio* (94) = 7.06, p < .001, and fair vs good, *t.ratio* (29) = 2.62, p = .06. These effects indicated that students were able to discriminate more reliable (good and fair) from less reliable (poor and bad) links at least to some extent.

Transfer task. The transfer task required students to read two short texts referring to social, technical or health-related issues. These texts were presented side by side at the top of an A4 page (21x29.7 cm). Each text was composed of 1) a title summarizing text content; 2) a paragraph describing the source; and 3) a paragraph stating content information that caused a contradiction across texts (document level; see Figure 2). Importantly, one of the two source descriptions featured a competent author communicating through a valid media (more reliable), whereas the other source featured a less competent author communicating through a non-valid media (less reliable)[1]. Three pairs of texts dealing with different topics were created (i.e., "*The economic impact of solar energy*"; "*The effects of Aspartame (fake sugar) on health*"; and "*The effects of Urban sprawl (suburbanization) on sustainable development*"). The assignment of sources and topics were counterbalanced across schools and assessment phases. A prior enquiry with teachers as well as a subsequent informal group question to

[1] Author motivation was controlled in this task, by keeping an implicit "intention to inform" in both sources. This had a two−fold purpose: a) to avoid ceiling effects in accuracy due to highly contrasted sources, and b) to reduce task complexity by keeping the number of texts in two.

students, suggested that the students had very little knowledge of the topics presented in the two sourcing assessment tasks.



Text A

Significant savings with solar energy

Text found on www.solenergy_blog_NM.fr,
October 14, 2015.
Author: Nicolas Martin, professor of Sociology at the University of Strasbourg. In a personal blog he runs regularly, he shares an analysis of energy saving through solar drawn from his many meetings and lectures.

"Although solar energy is still little developed, the results are nevertheless visible. I know many people who have installed solar panels at home, in cloudy regions, saving hundreds of euros per year. This saving can be even more important in the south of France."

Text B

The cost of solar electricity subsidies

Text found on www.energy_science_journal.fr,
December 20, 2014.
Author: Jacques Dubois, researcher of alternative energy at the Higher School of Engineering in Lille. In an article published in the academic journal of Energy Science, he analyses the impact of solar energy in economy.

"The cost of solar electricity subsidies exceeded 100 billion in France but their meager results jeopardize the country's transition to renewable energy. Our study shows that all solar plants combined produce less electricity than 2 nuclear reactors, which is insufficient to cover grants."

Figure 2. Example of a pair of texts presented in the transfer task, with Text A containing the less reliable source and Text B the more reliable source.

In the near transfer question, students were asked to indicate which of the two texts was better and why, by selecting one of three alternatives (i.e., "*Text A*", "*Text B*", or "*Both texts are at the same level*"), and justifying their answer. Responses were scored for accuracy (selection of the text featuring the more reliable source) and for the inclusion of source features in the justification (see section 2.3). In the far transfer question, students were asked to write a short conclusion about the topic discussed in the texts. Again, responses were scored for the inclusion of source features.

In the pre- and post-test phases, the sourcing assessment questions were always assigned in the following order: far transfer, near transfer, and knowledge application. This was done in order to rule out a potential contamination from the most explicit (knowledge application) to the more implicit (far transfer) question.

2.2.3. Control measures

To further control for potential group differences, we tested two basic abilities: reading comprehension and working memory capacity.

Reading comprehension. An adapted version of the test "Protocole Emilie" (Emilie's protocol; Duchêne, 2010), the only text comprehension task adapted to teenagers available in French (Pourcin & Colé, 2016), was used to evaluate students' reading comprehension. This protocol assessed narrative rather than scientific comprehension, which was preferred to better cover basic reading processes like inference making, and ensure a similar amount of prior knowledge across students. A long narrative text (897 words) was presented and students were instructed to read it during eight minutes. Subsequently, the text was removed and the test-takers had to answer 25 questions assessing literal and inferential comprehension. This protocol was administered just after the sourcing skill assessment of the pre-test phase, and took approximately 20 minutes.

Working memory. To test working memory capacity we used the standardized letter-number sequencing task (WISC-IV; Wechsler, 2005). The test administrator uttered a series of alternating numbers and letters, and students had to first report the numbers in ascending numerical order, and then the letters in alphabetical order. Difficulty increased progressively from a block of 2 to a block of 7 items, including three trials per block. A practice with a block of 2 and a block of 3 items was provided to ensure that students understood the instructions. The total score was the sum of all items included in the trials correctly recalled, with a maximum score of 81. This task was conducted after the sourcing skill assessment of the post-test 1 phase, and took approximately 10 minutes.

2.2.4. Homogeneity and fidelity of the trained and control groups

The training workshops were run jointly by the teacher and two of the six members of the research team. To ensure homogeneity and fidelity of the intervention, the research team met on various occasions prior to and during the intervention in order to review and discuss

the script of each session. Moreover, different pairs of researchers were assigned to different sessions in distinct classrooms, to avoid any researcher-related bias in intervention effectiveness. Finally, during the training workshops, students received a student booklet and were actively monitored in their completion of the practice tasks. The researchers made sure to take questions and comments from as many students as possible during each session.

Although we could not directly assess the fidelity of the control group, we had no report of teachers missing classes or any other type of disruption in the regular class activities. Conversely, the teachers were never provided with the materials ahead of the training sessions, which decreased the risk for them to accidentally communicate training materials to the control classes.

2.3. Data scoring and data analysis

Students' justifications for why a document was better (near transfer) and their conclusions about the topic (far transfer) were content-analyzed according to a scoring rubric similar to that of Britt and Aglinskas (2002). The scoring category of interest was "source", that is, reference to information presented in the source paragraph (e.g., "*A scientific journal provides better information*", our underlining). A random sample of 9% (74 out of 822 cases) of the responses was scored by two independent raters, with satisfactory inter-rater agreement for both near and far transfer questions (all Cohen's $\kappa s > .70$). Discrepancies were resolved through discussion.

Different statistical analyses were conducted for each assessment task. First, students' normalized rating scores in the knowledge application question were analyzed by a linear mixed-effects (LME) model by using the *lmer* function of the same lme4 R package. Second, we analyzed the proportion of accuracy and justification of the near transfer question through mixed-effects logistic regression (MELR) models by using the *glmer* function of the lme4 R package (Bates, Maechler, Bolker, & Walker, 2015). The far transfer question did not allow

statistical analyses, due to a low frequency of source reference (see section 3.3). Both LME and MELR accounted for random and fixed effects, with Participants and Items[2] as the random factors, and Group (control and trained), Phase (pre-test, post-test 1 and post-test 2) and Condition (good, fair, poor and bad or more reliable and less reliable, depending on the question), as the fixed factors. In this way, the fixed structure was always composed by a three-way interaction (group x phase x condition). The variable school was included as a random slope of Items in the knowledge application question[3] to control for school error variability. All trials were included in the analyses. Chi-squares (N = 137) and p values of significant effects were provided by *Anova* function of the car package (Fox, et al., 2016). In cases where post-hoc comparisons were necessary, we used the *testInteractions* function of the phia R package (De Rosario-Martínez, 2015), with Bonferroni correction. Finally, effects sizes for LME were informed by the explained deviance (*dv*), extracted by the pamer.fnc function of the LMERConvenienceFunctions Rpackage (Tremblay & Ransijn, 2015). This statistic serves as a generalisation of R^2 because it measures the marginal improvement or reduction in unexplained variability in the fixed component after accounting for a given predictor. Effects size for MELR were reported using odds ratio (OR) and the 95% confidence interval (CI), by the lsmeans function of the lsmeans R package (Lenth, 2017). OR indicates the constant effect of a specific predictor on the likelihood that one outcome will occur, whereas CI estimates the precision of the OR (i.e., small CI signals a higher precision).

3. Results

Our results are organized into three sections. We first explored the general effects of the reading comprehension and working memory measures, to check the equivalence between groups. Second, the normalized rating score in the knowledge application task was analyzed to clarify whether trained students improved their ability to discriminate more reliable from

[2] The near transfer question only included Participants, because responses were based on a single item.
[3] Models including the random slope of school in Participants caused convergence problems, probably due to limited sample size (Barr, Levy, Scheepers, & Tily, 2013).

less reliable source information. Finally, we examined proportion of accuracy and justification responses to the near transfer question, to understand whether trained students did significantly more reference to the more reliable source after the training (see correlations within and between the three assessment questions in Appendix 1). Taking into account the large number of results, we focused on the fixed effects of each LME or MELR analysis. The summary details (lmerTest package) of each model are provided in Appendix 2.

3.1. Comprehension and memory skills

All schools performed above chance, with more than 21 correct questions (84%) in the reading comprehension task, and at least 37 items correctly recalled (span of 4.5) in the working memory task. Additionally, to understand if there were general differences between the control and trained groups, we ran *t*-test comparisons on both measures. No significant differences were found between groups either in reading comprehension, $t(135) = .38$, $p = .71$ (control: $M = 21.30$, $SE = 0.21$, and trained: $M = 21.41$, $SE = 0.18$), or working memory, $t(135) = 1.05$, $p = .30$ (control: $M = 38.64$, $SE = 1.67$, and trained: $M = 35.89$, $SE = 2.06$), confirming similar text-level comprehension and processing ability across groups.

3.2. Knowledge application task

A LME model with group (trained vs control), test phase (pre-test, post-test 1, post-test 2) and condition (good, fair, poor, bad) was run on the mean normalized rating score (see Table 1 for means and standard errors).

Table 1. Adjusted means (and standard errors) of the normalized rating scores in the knowledge application task, as a function of group, phase and condition.

			Knowledge application		
			Pre-test	*Post-test 1*	*Post-test 2*
More reliable	*Good*	Control	0.66 (0.039)	0.63 (0.040)	0.63 (0.039)
		Trained	0.64 (0.041)	0.56 (0.042)	0.60 (0.041)

Less reliable	*Fair*	Control	0.57 (0.039)	0.47 (0.040)	0.47 (0.039)
		Trained	0.58 (0.041)	0.51 (0.042)	0.45 (0.041)
	Poor	Control	0.42 (0.039)	0.40 (0.040)	0.38 (0.039)
		Trained	0.37 (0.041)	0.31 (0.042)	0.36 (0.041)
	Bad	Control	0.30 (0.039)	0.38 (0.040)	0.36 (0.039)
		Trained	0.33 (0.041)	0.31 (0.042)	0.23 (0.041)

Note. The distinction between "More reliable" (good and fair) and "Less reliable" (poor and bad) is merely indicative, as condition (type of link) always referred to good, fair, poor and bad links. Ratings represent the self-reported likelihood for a student to consult the link and could take values from 0 (certainly not) to 1 (certainly).

The three main effects were significant: group, χ^2 (1) = 5.49, $p < .05$, $dv = .09$, where the trained group ($M = 0.43$, $SE = 0.03$) provided lower ratings than the control group ($M = 0.46$, $SE = 0.03$); phase, χ^2 (2) = 9.17, $p < .05$, $dv = .08$, where students' ratings decreased in the post-test 1 ($M = 0.45$, $SE = 0.03$) and post-test 2 ($M = 0.42$, $SE = 0.03$) compared to the pre-test ($M = 0.48$, $SE = 0.03$); and condition, χ^2 (3) = 62.69, $p < .001$, $dv = .37$, where students provided lower ratings for bad and poor links ($Ms = 0.30$ and 0.37, $SE = 0.03$ and 0.03, respectively) compared to fair and good links ($Ms = 0.52$ and 0.64, $SE = 0.03$ and 0.03). The three two-way interactions were not significant: group and phase, χ^2 (2) = 5.13, $p = .08$; group and condition, χ^2 (3) = 7.29, $p = .06$; and phase and condition, χ^2 (6) = 11.73, $p = .07$.

More importantly, the three-way interaction of group, phase and condition was significant, χ^2 (6) = 12.64, $p < .05$, $dv = .04$ (see Figure 3). To identify the locus of this interaction, we performed additional analyses for each condition (type of link). Students' ratings for good and fair links did not differ across groups on any of the three phases (for good and fair links: $ps \simeq 1.00$, in the pre-test; $ps > .17$ in the post-test 1; $ps > .98$ in the post-test 2). In contrast, although trained and control students' ratings of poor links did not differ at the pre-test, χ^2 (1) = 1.69, $p = .58$, trained students gave lower ratings at the post-test 1, χ^2 (1) = 7.09, $p < .05$; no group differences were found at the post-test 2, χ^2 (1) = 0.45, $p \simeq 1.00$. For

bad links, control and trained students again did not differ at the pre-test, χ^2 (1) = 1.15, p = .85), but trained students tended to give lower ratings at the post-test 1, χ^2 (1) = 4.93, p = .08, and gave significantly lower ratings at the post-test 2, χ^2 (1) = 16.69, p < .001.

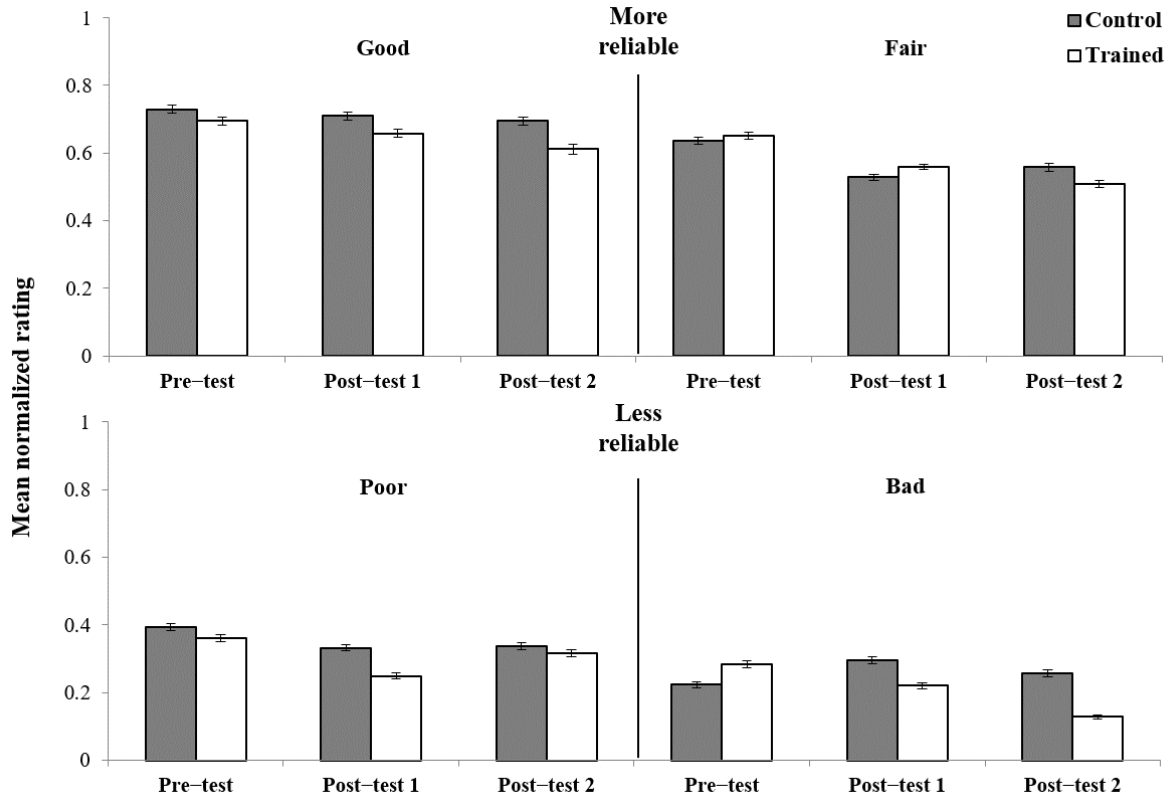


Figure 3. Mean normalized rating score in the knowledge application question, divided by group, phase and condition.

Overall, these results indicate that compared to controls, trained students became more critical toward less reliable sources after the training sessions, with the effect still holding after a period of three weeks for the least reliable (bad) links.

3.3. Transfer task

Students used very different amount of source information in their responses to the near and far transfer questions (see Table 2 for means and standard errors). In the near transfer question ("*Which of the two texts is better and why?*") students cited sources quite frequently

(0.36 control and 0.48 trained), whereas in the far transfer question ("*What can be concluded*") only a small minority referred to source (0.06 control and 0.07 trained). In light of these data, statistical analyses were conducted only in the near transfer question.

Table 2. Mean proportion (and standard errors) of accuracy and source reference in the near transfer and far transfer questions of the transfer task, as a function of group, phase and condition.

			Near transfer		
			Pre-test	*Post-test 1*	*Post-test 2*
Accuracy	*More reliable*	Control	0.52 (0.06)	0.48 (0.06)	0.59 (0.06)
		Trained	0.52 (0.06)	0.81 (0.05)	0.81 (0.05)
Source	*More reliable*	Control	0.33 (0.06)	0.37 (0.06)	0.49 (0.06)
		Trained	0.30 (0.06)	0.75 (0.05)	0.75 (0.05)
	Less reliable	Control	0.29 (0.05)	0.33 (0.06)	0.33 (0.06)
		Trained	0.25 (0.05)	0.50 (0.06)	0.34 (0.06)

			Far transfer		
			Pre-test	*Post-test 1*	*Post-test 2*
Source	*More reliable*	Control	0.10 (0.03)	0.03 (0.02)	0.04 (0.02)
		Trained	0.08 (0.03)	0.06 (0.03)	0.05 (0.03)
	Less reliable	Control	0.12 (0.04)	0.03 (0.02)	0.03 (0.02)
		Trained	0.09 (0.04)	0.08 (0.03)	0.06 (0.03)

Note. "Source" referred to any mention of author, author position, or media quality. Scores could take values from 0 to 1.

3.3.1. Near transfer question

Accuracy. A binomial MELR model with group (trained vs control) and test phase (pre-test, post-test 1, post-test 2) was run on students' selection of the more reliable text[4]. The model demonstrated a significant main effect of group, χ^2 (1) = 10.68, $p < .01$, with more correct responses in the trained (M = 0.72, SE = 0.18) than in the control group (M = 0.53, SE

[4] Because condition (more reliable vs. less reliable) was already in the answer (Text A, Text B, or both texts are at the same level), the model for accuracy only included the fixed factors of group and test phase.

= 0.15; OR = 0.39, CI 0.23-0.64); and phase, χ^2 (2) = 8.13, $p < .05$, with more correct responses in the post-test 1 (M = 0.65, SE = 0.20) and post-test 2 (M = 0.71, SE = 0.21) than in the pre-test (M = 0.52, SE = 0.19; OR = 0.52, CI 0.27-0.99 and OR = 0.42, CI 0.22-0.82 for the post-test 1 and post-test 2, respectively). More importantly, the interaction between these two factors was also significant, χ^2 (2) = 9.73, $p < .01$ (see Figure 4). Post-hoc comparisons showed no group differences in the pre-test phase, χ^2 (1) = 0.0001, $p \simeq 1.00$ (OR = 0.99, CI 0.34-2.95); whereas the trained group outperformed the control group in the post-test 1, χ^2 (1) = 14.90, $p < .001$ (OR = 0.19, CI 0.05-0.64), and post-test 2, χ^2 (1) = 7.19, $p < .05$ (OR = 0.31, CI 0.09-1.08). Therefore, consistent with our hypothesis, trained students were better than controls at detecting more reliable source information after the intervention. Importantly, this finding was also true three weeks later, indicating the intervention had a rather robust impact.

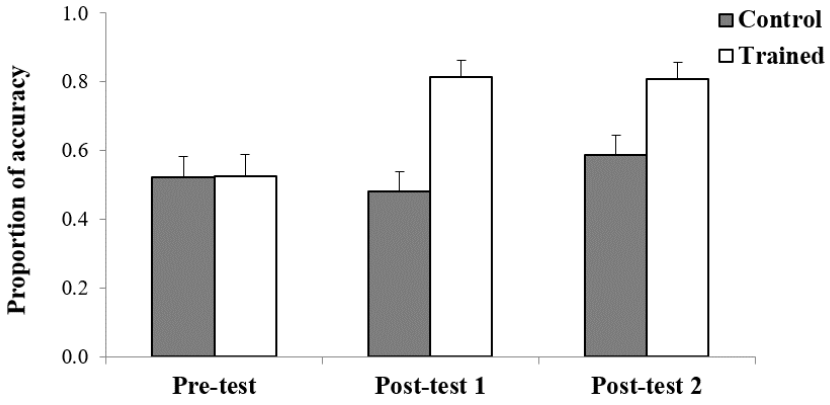


Figure 4. Proportion of times the more reliable source was selected (accuracy) in the near transfer question, divided by group and phase.

References to source in justifications. A second binomial MELR model with group (trained vs control), test phase (pre-test, post-test 1, post-test 2) and condition (more reliable vs less reliable) was run on references to source information in their justifications. This model showed the significant main effects of group, χ^2 (1) = 6.66, $p < .05$, with more references to source in the trained (M = 0.48, SE = 0.18) than in the control group (M = 0.32, SE = 0.18; OR = 0.47, CI 0.27-0.84); phase, χ^2 (2) = 28.90, $p < .001$, with more references to source in

the post-test 1 (M = 0.47, SE = 0.17) and post-test 2 (M = 0.47, SE = 0.17) compared to the pre-test (M = 0.24, SE = 0.19; OR = 0.33, CI 0.20-0.55 and OR = 0.34, CI 0.20-0.57 for the post-test 1 and post-test 2, respectively); and condition, χ^2 (1) = 24.53, p < .001, with less source reference to the less reliable (M = 0.29, SE = 0.16) compared to the more reliable source (M = 0.49, SE = 0.15; OR = 2.48, CI 1.75-3.51).

Importantly, the three two-way interactions were also significant. Post-hoc comparisons in the interaction of group and phase, χ^2 (2) = 15.67, p < .001, demonstrated that trained students made more references to source in the post-test 1, χ^2 (1) = 16.99, p < .001 (OR = 0.21, CI 0.07-0.61). However, there were no significant differences in the pre-test, χ^2 (1) = 0.14, $p \simeq 1.00$ (OR = 1.15, CI 0.38-3.49), and post-test 2, χ^2 (1) = 4.53, p = .10 (OR = 0.45, CI 0.15-1.31). In addition, the interaction of group and condition, χ^2 (1) = 5.76, p = .02, demonstrated the two groups did not differ in their references to the less reliable source, χ^2 (1) = 0.91, p = .68 (OR = 0.72, CI 0.30-1.73), whereas trained students made more references to the features of the more reliable source, χ^2 (1) = 11.73, p < .01 (OR = 0.31, CI 0.13-0.75). Finally, the interaction between phase and condition, χ^2 (2) = 9.04, p < .05, indicated no differences between the two sources in the pre-test, χ^2 (1) = 0.99, p = .96 (OR = 1.36, CI 0.56-3.27), but differences were evident in both the post-test 1, χ^2 (1) = 7.62, p < .05 (OR = 2.28, CI 0.97-5.32), and post-test 2, χ^2 (1) = 27.39, p < .001 (OR = 4.91, CI 2.06-11.68), with less reference to the less reliable source after the intervention.

Although the three-way interaction between group, phase and condition was not significant, χ^2 (2) = 2.56, p = .28, the pattern of results suggests that trained students increased their use of source information after the intervention and they did so specifically with the more reliable as opposed to the less reliable source (see Figure 5). The increased reference to source information in trained students was maintained three weeks later, showing again that the intervention produced rather long-lasting effects.

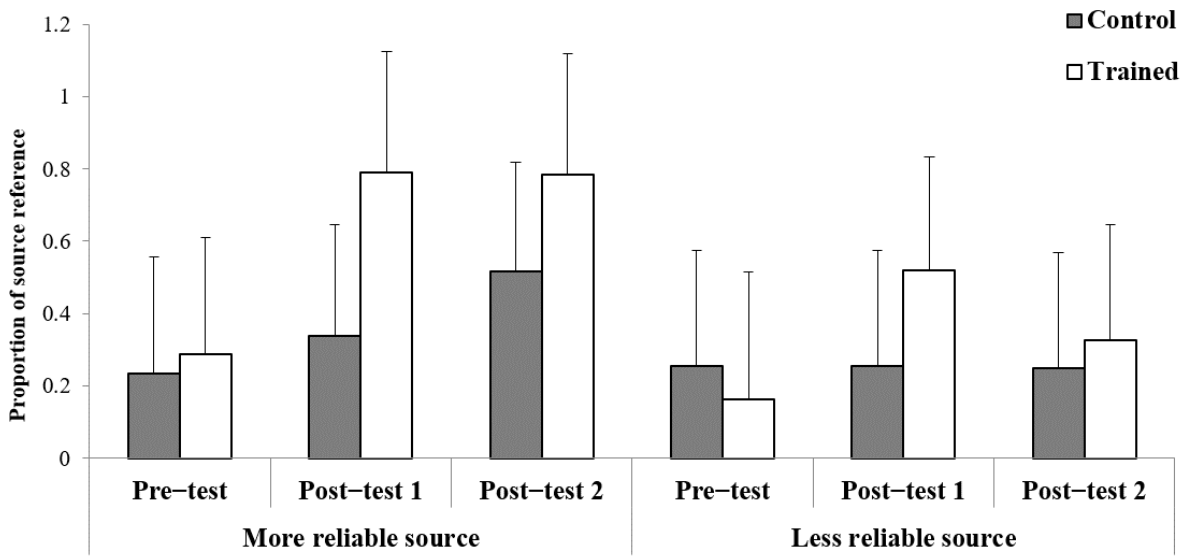


Figure 5. Proportion of students who referred to source information in the in the near transfer question, divided by group, phase and condition.

Overall, our findings signalled short and longer beneficial effects of the intervention, with trained students providing lower ratings to less reliable links (knowledge application question) as well as being better at detecting and using more reliable source information to justify the best text (near transfer question) than control students. In contrast, no intervention benefits were found on students' use of source information in their conclusion from the two texts (far transfer question).

4. Discussion

The main goal of our study was to assess the effects of an instructional intervention grounded in multiple text comprehension theories, and aimed at fostering teenagers' evaluation of information reliability. A related goal was to understand whether the skilled learned as part of the training could transfer to different source evaluation tasks. Finally, we also investigated whether the effectiveness of the intervention would carry over a few weeks' delay, which none of the studies conducted thus far had provided evidence for. In the following sections, we first discuss the general and specific findings obtained on the sourcing assessment tasks, and the impact of the intervention in the longer term. Subsequently, we

suggest some conclusions about how and why our instructional intervention was effective in training teenagers' sourcing skills. Finally, we discuss some of the limitations of our study and perspectives for future research.

Table 3. Summary of the hypotheses assumed in our intervention and the results obtained in the three sourcing skills assessment questions.

Task	**Question**	**Hypothesis**	**Results**	
			Post-test 1	***Post-test 2***
Knowledge application 	***1) Knowledge application*** Rate from 0 (certainly not) to 4 (certainly) if you would consult each link.	More contrasting ratings of trained students in the post-test 1 and post-test 2. Especially in the less clear-cut (fair and poor) links.	**Partially confirmed.** Compared to controls, trained students provided more contrasted ratings. However, this was true only for the less reliable (poor and bad) links, and not for the fair links.	**Partially confirmed.** Trained students also provided more contrasting ratings in longer term, but only in the case of bad links.
Knowledge transfer 	***2) Near transfer*** Which of the two texts is better and why?	Few references to sources in the pre-test, and more use of source information for trained students in the post-test 1 and post-test 2.	**Confirmed.** Trained students a) selected the more reliable text (accuracy) and b) increased the use of the more reliable source in their justifications more frequently than controls.	**Confirmed.** Trained students also a) selected the more reliable text (accuracy), and b) increased the use of the more reliable source in their justifications more frequently than controls.
	3) Far transfer What can be concluded about the topic?	No references to sources in the pre-test, and use of source information for trained students in the post-test 1 and post-test 2.	**Could not be assessed.** Students made almost no reference to source information before (pre-test) and after the training (post-test 1 and post-test 2).	

4.1. Assessing the effectiveness of the intervention

During the training workshops, the constructs of author position, author motivation and media quality were introduced, exemplified and discussed. To assess students' ability to apply this type of knowledge, we designed a task that required students to rate how likely they would be to use links in which all three dimensions were varied (knowledge application task). We found that compared to controls, trained students assigned lower rating scores to poor and bad links, that is, links containing less reliable features on two or three of the dimensions. In other words, our intervention made trained students more critical about less reliable source information. This finding suggests that the evaluation of untrustworthy information (e.g., information linked to some vested commercial interest, or information published in a non-validated media) is more complex and requires more source knowledge than the evaluation of reliable information. Our findings are consistent with previous intervention studies, which have shown students to become more critical about untrustworthy websites by realising that information on the Internet is not always accurate or true (e.g., Zhang & Duke, 2011).

Our study also asked whether the effect of the intervention would transfer to a question involving an implicit request to pay attention to source features, such as deciding which of two texts was better and providing a short justification (near transfer question). At the pre-test, the two texts were equally likely to be selected (see Figure 4). However, at both the post-test 1 and post-test 2, trained students were more likely to be accurate by selecting the more reliable text (i.e., the one containing the more reliable source). Furthermore, their explanation of why the text was better made more reference to features of the more reliable source. Bearing in mind that the two texts provided clearly conflicting views about the topic, these effects can be interpreted within the CSI model (Stadtler & Bromme, 2014) which posits that when readers associate conflict with the existence of multiple sources, they may solve the conflict through either first-hand ("*what is true*") or second-hand ("*whom to believe*")

evaluations. First-hand evaluations are only possible when readers possess a high level of prior knowledge of the contents. Because the assessed topics were unfamiliar to students, it is unlikely that they would be in a position to perform first-hand evaluations, by discerning and integrating accurate vs. less accurate claims across texts. Indeed, trained students turned to second-hand evaluations, which led them to select the more reliable source. At this stage, it is important to note that performing a second-hand evaluation requires a deliberate decision on the part of the reader (see Wineburg, 1991, "sourcing heuristic"). In other words, to know "whom to believe" requires one to compare the different sources instead of trying to rely on one's own knowledge of the topic. In short, our intervention significantly impacted students' performance on a near transfer question involving the use of source evaluation to understand information reliability.

Finally, we also asked students to draw to a conclusion about the topic presented in the two contradictory texts. Prior studies have found that undergraduate university students rely on source information as a way to integrate conflicting claims (Braasch et al., 2012; Rouet et al., 2016). However, students in the present study made mostly reference to content information, with strikingly few reference to sources. This finding is similar to previous studies showing that students' epistemic evaluations were primarily based on content rather than on source information (Macedo-Rouet et al., 2013; Salmerón et al., 2016), and ruled out the possibility that students would be using a comprehension strategy based on source evaluation.

Altogether, our findings indicate that trained students benefited from the intervention in performing knowledge application and near transfer questions, but not a more demanding far transfer question. Differences between the three source evaluation questions are important, as they reflect distinct mechanisms underlying one's ability to evaluate source information. For instance, the knowledge application question explicitly required to rate source information,

whereas the near and far transfer questions only did so implicitly. The implicit request, and the fact that it was triggered by the contradiction among the texts, required students to be aware that source information was available and potentially helpful in resolving the conflict.

Moreover, trained students needed a minimum level of prompting ("*One of the two texts is better*") to pay attention to source information. However, considering that trained students were better than controls in detecting and using source reliability to make a decision, this prompting was not sufficient to engage in sourcing behavior. Then, in addition to specific prompting, students receiving the intervention benefitted from group discussions about conflicting situations and explicit declarative knowledge about source dimensions (see also section 4.3). This interpretation is congruent with the literature on collaborative discourse and argumentation, whereby in the presence of different viewpoints or (socio)cognitive conflict, content integration is enhanced by prompting (e.g., through explanations and discussion) elaborative and metacognitive skills (Nussbaum, 2008). In fact, research has also shown that discussions in a classroom environment are an optimal way to strength and preserve longer-term effects (e.g., Aulls, 1998), which is also consistent with some of our specific results (see section 4.2).

4.2. Short and longer term effects

Our study also aimed to understand whether the effects of the intervention would hold in the longer term. Although some of our findings only reflected short term effects (i.e., lower ratings of poor links and increased reference to source in student's justifications), some of them manifested both in the short and the longer term. Compared to controls, trained students gave lower ratings to bad links in the knowledge application question, as well as they selected the more reliable text (accuracy) more frequently in the near transfer question, three weeks after the intervention. Moreover, despite the three-way interaction was not significant in the near transfer question, the pattern of results indicated that trained students increased their use

of more reliable (but not less reliable) source information in the post-test 2 (see Figure 5), suggesting a better sourcing behavior in trained students in the longer term.

Because these results prove that information about how to detect and evaluate source's reliability in multiple documents can be relatively consolidated after three training sessions, taking into account the importance of promoting teenagers' critical thinking by providing sourcing strategies, this evidence points at the need to gradually integrate sourcing skills strategies in students' formal curriculum.

4.3. Intervention effectiveness: The role of instructional techniques

Our intervention was based on the systematic introduction and discussion of the key source dimensions of author position ("*who says what*"), author motivation ("*why the author says it*") and media quality ("*where it is published*"), and the use of combined several instructional techniques that prior research has found to be effective (see e.g., Braasch, et al., 2013; Macedo-Rouet et al., 2013; Wiley, et al., 2009). Specifically, we organized group discussions about pairs of texts presenting contradictory information, provided explicit declarative knowledge about each of the three source dimensions, and assigned practice exercises in which students had to rate contrasted items. The students were invited to share their responses and received feedback accordingly.

As suggested before, some of these strategies were essential in fostering students' sourcing behavior. For instance, group discussions helped students to become aware that the evaluation of source's reliability was a way to understand which text was better in the near transfer question, whereas our explanations about pre- and post-publication validation processes were fundamental to foster students' evaluation of media quality. Furthermore, trained students assigned lower ratings to the less reliable links of the knowledge application question (which assessed the interplay of the three dimensions), although none of the practice exercises provided in the sessions ever included all three dimensions in interaction. Besides

the specific knowledge given for each dimension, we believe that trained students' better ability to critically discriminate untrustworthy information was also due to additional oral discussions in which students reflected about the possible interactions between the three source dimensions. Overall, these findings shed some light on the importance of implementing a well-designed instructional intervention to promote short and longer term sourcing skills in ninth graders' students.

4.4. Limitations and future research

Clearly, our intervention study entails some limitations that should be addressed in future research. First, researchers visited the trained classes for both the assessment and training sessions, whereas the control classes only met the researchers at the assessment sessions. A better control condition would consist in implementing a different intervention in the control classes, by for example training control students to find information on the Internet without emphasizing the importance of source evaluation. Although in the present case it seems unlikely that the mere presence of the researchers in the classroom would explain trained students' increased performance at the post-tests, an alternative intervention would enable a better control of the potential researcher effect. This intervention would also guarantee the same level of fidelity and homogeneity in the trained and control classes, respectively. Second, the transfer task used in the assessment manipulated author position and media quality, whereas author motivation was only controlled. Future studies should try to assess the impact of author motivation in the near and far transfer questions, for example by providing the authors' affiliation (e.g., "*a public institution*" or "*a company with vested interests*"). Finally, the fact that our intervention did not lead to beneficial effects in the far transfer question could be due to our excessive use of the rating task as opposed to the writing task during the training workshops. Therefore, future studies should also try to prompt

sourcing skills by training students to write about multiple documents, with instructions prompting them more or less explicitly to make use of source information.

Despite these limitations, we believe that the present intervention study provides a realistic and innovative approach to train teenagers' critical thinking in educational contexts. More specifically, considering teenagers' heavy use of the Web and their reported difficulties when trying to comprehend multiple documents, the implementation of a school program focusing on the evaluation of author position, author motivation and media quality, by combining class discussions about conflicting scenarios, declarative knowledge about the source dimensions and appropriate practice, seems to be an efficient procedure to foster teenagers' skills on source reliability. In addition, the present study also provides evidence for the need to use different types of sourcing assessment questions to disentangle the real learning that students can reached with a particular intervention.

6. References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, *85*, 275-314.

Aulls, M. W. (1998). Contributions of classroom discourse to what content students learn during curriculum enactment. *Journal of Educational Psychology*, *90*(1), 56.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014.* Retrieved from https://cran.r-project.org/

Braasch, J. L. G., & Bråten, I. (2017). The discrepancy-induced source comprehension (D-

ISC) model: Basic assumptions and preliminary evidence. *Educational Psychologist*, 1-15.

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*, 180-195.

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

Bråten, I., Stadtler, M., & Salmerón, L. (in press). The role of sourcing in discourse comprehension. In M.F. Schober, M.A. Britt, & D.N. Rapp (Eds.), *Handbook of discourse processes* (2nd. ed.). New York: Routledge.

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

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

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

De Rosario-Martínez, H. (2015). *Phia: post-hoc interaction analysis. R package version 0.1-3*. Retrieved from https://cran.r-project.org/

Duchêne, A. (2010). *Emilie - Protocole d'évaluation de la compréhension de textes chez les collégiens avec 1 Cédérom* [*Emily – Text comprehension assessment protocol for secondary school students with 1 CDROM*]. Isbergues: Ortho-Edition.

Eastin, M. S., Yang, M. S., & Nathanson, A. I. (2006). Children of the net: An empirical exploration into the evaluation of Internet content. *Journal of Broadcasting & Electronic Media, 50*, 211-230.

Eurostat (2015): Being young in Europe today. Retrieved from: http://ec.europa.eu/eurostat/documents/3217494/6776245/KS-05-14-031-EN-N.pdf/18bee6f0-c181-457d-ba82-d77b314456b9

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., … Heiberger, R. (2016). *Package 'car'*. Retrieved from https://cran.r-project.org/

Goldman, S. R., & Scardamalia, M. (2013). Preface for the special issue multiple document comprehension. *Cognition and Instruction*, *31*, 121-121.

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, *44*, 1467-1484.

Kammerer, Y., Amann, D. G., & Gerjets, P. (2015). When adults without university education search the Internet for health information: The roles of Internet-specific epistemic beliefs and a source evaluation intervention. *Computers in Human Behavior*, *48*, 297-309.

Kammerer, Y., & Gerjets, P. (2014). The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *International Journal of Human-Computer Interaction*, *30*, 177-191.

Kammerer, Y., Kalbfell, E., & Gerjets, P. (2016). Is this information source commercially biased? How contradictions between web pages stimulate the consideration of source information. *Discourse Processes, 53*, 430-456.

Kennedy, M., Fisher, M. B., & Ennis, R. H. (1991). Critical thinking: Literature review and needed research. *Educational Values and Cognitive Instruction: Implications for Reform*, *2*, 11-40.

Lenth, R. V. (2017). Package "lsmeans". Retrieved from https://cran.r-project.org/

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*, 204-226.

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.

Nordheim, L. V., Gundersen, M. W., Espehaug, B., Guttersrud, Ø., & Flottorp, S. (2016). Effects of school-based educational interventions for enhancing adolescents' abilities in critical appraisal of health claims: A systematic review. *PloS One*, *11*, e0161485.

Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology*, *33*(3), 345-359.

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Towards 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: Lawrence Erlbaum Associates.

Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research*, *42*, 237-249.

Porsch, T., & Bromme, R. (2011). Effects of epistemological sensitization on source choices. *Instructional Science*, *39*, 805-819.

Pourcin, L., & Colé, P. (2016). L'évaluation cognitive de la lecture au collège: synthèse des principaux outils de dépistage et de diagnostic des troubles spécifiques de la lecture et cognitifs associés [*Cognitive assessment of reading in secondary schools: Summary of the main screening tools and diagnosis of specific reading and cognitive related disorders*]. *Psychologie Française*, 1-24.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, *53*, 145-161.

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

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

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

Rouet, J.-F., Le Bigot, L., de Pereyra, G., & Britt, M.A. (2016). Whose story is this? Discrepancy triggers readers' attention to source information in short narratives. Reading and Writing, *29*, 1549–1570.

Salmerón, L., Macedo-Rouet, M., & Rouet, J.-F. (2016). Multiple viewpoints increase students' attention to source features in social question and answer forum messages. *Journal of the Association for Information Science and Technology*, *67*, 2404–2419.

Scharrer, L., & Salmerón, L. (2016). Sourcing in the reading process: Introduction to the special issue. *Reading and Writing*, *29*, 1539-1548.

Stadtler, M., & Bromme, R. (2014). The content–source integration model: A taxonomic description of how readers comprehend conflicting scientific information. D. N. Rapp, & J. L. G. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 379-402), MIT Press, Cambridge, MA.

Stadtler, M., Scharrer, L., Macedo-Rouet, M., Rouet, J.-F., & Bromme, R. (2016). Improving vocational students' consideration of source information when deciding about science controversies. *Reading and Writing*, *29*, 705-729.

Tremblay, A., & Ransijn, J. (2015). *Package 'LMERConvenienceFunctions'*. Retrieved from https://cran.r-project.org/

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Computers & Education*, *52*, 234-246.

Wartella, E., Rideout, V., Zupancic, H., Beaudoin-Ryan, L., & Lauricella, A. R. (2015). *Teens, health, and technology: A national study*. Evanston, IL: Center on Media and Human Development, School of Communication, Northwestern University.

Wechsler, D. (2005). *WISC-IV: Echelle d'Intelligence de Wechsler pour Enfants* (4ème Edition) [*WISC-IV: Wechsler Intelligence Scale for Children (4th Edition)*]. Paris: Editions du Centre de Psychologie Appliquée.

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

Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*, 73-87.

Winter, S., Krämer, N. C., Appel, J., & Schielke, K. (2010). Information Selection in the Blogosphere – The Effect of Expertise, Community Rating, and Age. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 802–807). Austin, TX: Cognitive Science Society.

Zhang, S., & Duke, N. K. (2011). The impact of instruction in the WWWDOT framework on students' disposition and ability to evaluate web sites as sources of information. *The Elementary School Journal, 112*, 132-154.