How Good Is This Page? Benefits and Limits of Prompting on Adolescents' Evaluation of Web Information Quality

Mônica Macedo-rouet, Anna Potocki, Lisa Scharrer, Christine Ros, Marc Stadtler, Ladislao Salmerón, Jean-François Rouet

How good is this page?

Benefits and limits of prompting on adolescents' evaluation of Web information quality

Mônica Macedo-Rouet[1], Anna Potocki[2], Lisa Scharrer[3], Christine Ros[2],

Marc Stadtler[3], Ladislao Salmerón[4], & Jean-François Rouet[2]

[1]University of Paris 8 Vincennes-Saint-Denis, Paris, France

[2]CNRS-University of Poitiers, Poitiers, France

[3]Ruhr University Bochum, Bochum, Germany

[4]University of Valencia, Valencia, Spain

Citation:

Macedo-Rouet, M., Potocki, A., Scharrer, L., Ros, C., Stadtler, M., Salmerón, L., & Rouet, J.F. (2019). How good is this page? Benefits and limits of prompting on teenagers' assessment of Web information quality. *Reading Research Quarterly*, *54*(3), 299-321. DOI: 10.1002/rrq.241

Note : This is a preprint version of a published manuscript. The published paper may slightly differ from this version.

Abstract

The present study examined adolescents' detection of features that affect the quality of Web information. In Experiment 1, participants (12-16 years old) rated the goodness/usefulness of four Web-like documents for a simulated study assignment. Each document came with an issue that potentially undermined its quality. Two documents had source-related issues (i.e., non-competent author, outdated) and two had content-related issues (i.e., topic mismatch, poor readability). Most students failed to notice the issues, including topic mismatch. The participants also produced inconsistent evaluations of topic-match, readability, author competence and currency. In Experiment 2, students were prompted to assess each criterion separately. The participants distinguished poorer from better documents in relation to each criterion, except for author competence. We discuss these results in light of previous research on adolescents' evaluation behavior. We propose further avenues for reading research, and we articulate a few recommendations for educational practice.

Keywords: adolescents, information quality and reliability, evaluation criteria, online reading

How good is this page?

Benefits and limits of prompting on adolescents' evaluation of Web information quality

Current discussions about the spread of fake news through social media and other online channels have attracted public attention to the importance and complexity of information evaluation. Mainstream media and academic organizations have taken initiatives to debunk fake news and point them out to the users (e.g., https://crosscheck.firstdraftnews.com). However, these initiatives remain sparse and their impact is still unknown (Chokshi, 2017; Lazer et al., 2017). In addition, "determining who's behind information and whether it's worthy of our trust is more complex than a true/false dichotomy" (McGrew, Ortega, Breakstone, & Wineburg, 2017, p. 4). Since the burden of information evaluation is put on users, it is important to understand how users evaluate online information and what can be done to improve their evaluation skills.

More than any other group of users, adolescents (i.e. young people between the ages of 10 and 19 years, according to the World Health Organization [WHO], 2017) experience the challenges of information evaluation, because they are "almost constantly" connected to the Internet via mobile devices (Lenhardt, 2015). Research on adolescents' evaluation skills has been growing over the past few years, as concerns about their vulnerability to Internet misinformation arise, particularly in areas such as health, history and science (Metzger, Flanagin, Markov, Grossman, & Bulger, 2015; Sbaffi & Rowley, 2017; Wineburg & Reisman, 2015). There is a growing interest in understanding how adolescents evaluate online information and what types of intervention may foster their evaluation skills (Brante & Strømsø, 2017; Castek, Coiro, Henry, Leu, & Hartman, 2015; Flanagin & Metzger, 2010; Gasser, Cortesi, Malik, & Lee, 2012; Stanford History Education Group, 2016).

Evaluation can be defined as "the stage of the information-seeking process when an information seeker decides to use (or not use) a piece of information she or he has found"

(Gasser et al., 2012, p. 58). The decision to use a piece of information may be informed by a wide range of dimensions related to the appearance (e.g., is the information visually salient or in a prominent position), content (e.g., does the title of the document match the search query) and source of the information (e.g., is the author a competent person; Hilligoss & Rieh, 2008; Rieh, 2002).

Research has examined adolescents' evaluation behavior using methods ranging from surveys, interviews or focus groups (e.g. Connaway, Dickey, & Radford, 2011; Flanagin & Metzger, 2010; Hargittai, 2010; Hargittai & Hinnant, 2008; Paul, Macedo-Rouet, Rouet, & Stadtler, 2017) to direct observation (e.g. Barzilai & Zohar, 2012; Brand-Gruwel, Wopereis, & Walraven, 2009; Cho, 2014; Cho & Afflerbach, 2015; Watson, 2014). Only a few experimental studies have precisely investigated how adolescents implement evaluation criteria after receiving prompts to do so (Coiro, Coscarelli, Maykel, & Forzani, 2015; Goldman, Braasch, Wiley, Graesser, & Brodowinska, 2014; Kammerer, Meier, & Stahl, 2016; Mason, Scrimin, Tornatora, Suitner, & Moè, 2018; Mason, Junyent, & Tornatora, 2014; Stadtler, & Bromme, 2007; Stanford History Education Group, 2016; Wiley et al., 2009). Whereas some studies found that adolescents are well aware of the need to pay attention to quality, others suggest that many students fail to detect specific issues. Moreover, their selections are often guided by shallow cues, such as the position or visual salience of information.

The two experiments presented in this paper aimed to identify conditions that may support adolescents' detection of quality issues in a set of Web-like documents. In the first experiment, we assessed adolescents' holistic evaluation of information quality (Rieh, 2002) in the absence of specific prompts. In the second experiment, participants were provided with specific evaluation criteria. Based on goal-focusing theories of reading (Britt, Rouet, & Durik, 2018; McCrudden & Schraw, 2007), we expected that the provision of criteria would prompt

the participants to notice quality issues. Following Stadtler and Bromme (2008), we define prompting as "support measures which are embedded in the learning context and prompt the learner to execute specific metacognitive processes [such as evaluation]" (p. 718). Moreover, in Experiment 2 we compared two age groups corresponding to different educational levels (middle school and high school) in order to account for possible differences between younger and older adolescents, as suggested by previous research (Eastin, 2008; Metzger et al., 2015).

In the next sections, we review the research literature on adolescents' awareness of information quality issues on the Web, their actual use of content and source evaluation criteria to assess information quality, and the role of task contexts and evaluation prompts in this process.

Adolescents' awareness of information quality

Research based on self-reports, such as surveys and focus groups, suggests that adolescents are aware of issues regarding the quality and reliability of information they acquire on the Web. In a large-scale survey involving 2747 11 to 18 year-old participants, Flanagin and Metzger (2010) found that 79% of the participants "think about whether they should believe information they find online" and 71% agree with the idea that "people should be 'somewhat' to 'very' concerned with the believability of online information" (p. 32). Similarly, in a survey with 1060 18 to 19 year-old undergraduates, Hargittai, Fullerton, Menchen-Trevino, and Thomas (2010) found that "being able to identify the sources of information on the site" was the most important criterion when deciding to visit a Web site. Gray, Klein, Noyce, Sesselberg, and Cantrill (2005), who conducted focus groups with 157 11 to 19 year-old participants, observed that participants expressed "widespread mistrust" of health advice released by individuals through personal Web sites. In a review about youth media practices, Palfrey (2016) also reports that adolescents do care about privacy and credibility on the Web, contrary to some popular beliefs about young users of the Internet.

In addition, adolescents appear generally confident in their ability to identify "good" information from the Web. As a high-school student in Julien and Barkers' (2009) study put it: "I guess just basically from years of experience I can tell whether or not something is reliable or not reliable" (p. 15). Paul et al. (2017) interviewed 44 15-year-old students from two countries and found that all participants were able to cite reasons for evaluating source parameters (e.g. author, venue, date) when searching information on the Web. The participants were also aware of benefits of sourcing, such as determining whether an author is competent about a topic. Confidence in their own search strategies contributes to increasing adolescents' trust in information (Sbaffi & Rowley, 2017). Since most adolescents spend a great deal of time online, they may feel that their evaluation skills develop accordingly, which is not always true (Salmerón, García, & Vidal-Abarca, 2018).

In contrast with their reported awareness and concerns about information quality, adolescents' actual evaluation behavior seems to be unsystematic and sometimes inconsistent. A number of studies show that when they search for information online, adolescents tend to skim rather than read Web pages, to rely on "known" sources, and to prioritize easiness of use over content quality (Brand-Gruwel et al., 2009; Foss et al., 2013; Gasser et al., 2012; Julien & Barker, 2009; Kiili, Laurinen, & Marttunen, 2008; Zhang, 2013a; Watson, 2014). Watson (2014) has characterized this type of behavior as "convenience and pragmatism" (p. 1401). By observing and interviewing 37 Australian adolescents of 14 to 17 years of age as they searched information for school tasks, Watson found that they: (1) find the search process unchallenging and rely on easy solutions to find information; (2) are strongly motivated to find overviews or introductions to their topic of interest (this is one of the reasons why many rely heavily on Wikipedia as a source); (3) establish reliability incidentally, by considering as reliable sites that corroborate information from previous sources; (4) are influenced by prior knowledge and preconceptions about a source; (5) carry out relevance and reliability

judgments simultaneously in the process of filtering, and seldom assess reliability explicitly and independently (for similar findings, see Brand-Gruwel et al., 2009; Walraven, Brand-Gruwel, & Boshuizen, 2008).

Convenience and pragmatism, as well as evaluation heuristics (Metzger & Flanagin, 2013), allow adolescents to cope with the complex demands of information search and evaluation. However, even though they might be useful, simple heuristics may also lead to suboptimal decisions (Hilligoss & Rieh, 2008; Van der Heide & Lim, 2016).

In sum, studies based on self-reports suggest that adolescents are aware of the potential lack of quality and reliability of information on the Web, and they do express some concern about it. They are also knowledgeable about the benefits of evaluating information. However, concrete observations of adolescents' actual search behavior indicates that they do not evaluate information as systematically and accurately as their responses to surveys and interviews would suggest. Instead, they tend to use general heuristics and to rely on superficial cues when selecting online information.

Adolescents' use of content and source evaluation criteria

Studies of expert users' evaluation behavior indicate that they use a wide range of criteria to support their assessment of information quality (Kim, Park, & Bozeman, 2011; Rieh, 2002; Tabatabai & Shore, 2005; Wineburg, 1991). Those criteria may be roughly categorized as content- and source-based criteria, respectively. Content-based criteria include the topical relevance of the material (does the page deal with my search topic?) as well as its readability (is the text easy to read and understand?). Source criteria include an assessment of the author's competence and intentions (Britt, Perfetti, Sandak & Rouet, 1999), as well as other features such as whether the information is up to date, or whether it has been reviewed prior to publication (Pérez et al., 2018). Indeed, a source can be defined as information about the origin of a document and the circumstances of its production, such as who the author is,

when and where the document was published, and so forth (Bromme, Stadtler & Scharrer, 2018; Rouet & Britt, 2014).

Adolescents' use of content-based criteria varies as a function of age and education. In their review of adolescents' media use, Gasser et al. (2012) found that topic-match is associated with finding documents that "contain the given words" (p. 60). Indeed, Julien and Barker (2009) observed that 41% of the 15 to 16 year-old students who completed a biology assignment (finding information about the major world biomes) skimmed information for relevant key terms in order to assess topicality. Similarly, in an experiment by Foss et al. (2013) 65% of the 14 to 17 year-old participants used text snippets to evaluate topical relevance of the results provided by a search engine in response to self-generated queries. Finally, Rouet, Ros, Goumi, Macedo-Rouet, & Dinet (2011) showed that students from fifth to ninth grades were misguided by typographical cues when selecting document headers from a search results page. The younger participants were more likely to select links that contained keywords from the search phrase if those were displayed in capital letters, even though the links did not match the topic of the query.

Readability may be defined as how simple and easy to understand a text is, based on its structure and vocabulary (Crossley, Allen, & McNamara, 2011). In a review of 73 journal articles on the perceived credibility of Web-based health information, Sbaffi and Rowley (2017) found that readability was the fourth most common evaluation criterion, and a particularly important criterion for adolescents. In Watson's (2014) study cited above, participants considered Wikipedia a major, relevant source because it is fast and easy to find, and easy to understand (see also Blikstad-Balas & Hvistendahl, 2013; Connaway et al., 2011; Menchen-Trevino & Hargittai, 2011). Furthermore, adolescents have expectations regarding the readability of documents that are directed to them. Grootens-Wiegers, De Vries, Vossen, and van den Broek (2015) conducted focus groups with 77 11 to 12 year-old adolescents to

discuss their perceptions of a pediatric research information form that had a Flesch score of 55.43 (i.e. fairly difficult to read). Participants found the form hard to read, and expressed the need for a better explanation of the meaning of scientific concepts. One of the participants stated: “If this is supposed to be for children, I would make it easier” (p. 105). For these reasons, reading scholars stress the importance of readability as a way to engage adolescents with scientific texts and help them develop reading literacy (McCormick & Segal, 2016).

So far, less is known about when and how adolescents evaluate source-related criteria such as author competence (i.e. whether the author has expertise on the topic) and currency (i.e. whether the document is up to date) (Hilligoss & Rieh, 2008; Metzger, 2007). Early studies of Web navigation in the context of school activities suggested that adolescents use simple heuristics and are frequently misled by surface features (e.g., a picture showing a medical doctor; Brem, Russell, & Weems, 2001). Coiro et al. (2015) asked seventh-graders to locate and evaluate two to four websites for a school project on one of two topics (energy drinks or heart health). Students could browse a Web-like database and select any site they wanted from a list of search results. For the selected sites, they had to summarize relevant details in a notepad, then evaluate each site with four questions: (1) Who is the author? (2) Is he/she an expert and how do you know? (3) What is his/her point of view and how does it affect the text and images on the site? (4) Is the information reliable and how do you know? Students could perform the evaluation at their own pace and browse the websites while answering the questions. Most students were able to correctly identify the author of the website. However, only 31% provided a clear yes/no answer to whether the author was an expert, and 51% failed to cite any specific criteria for expertise. Rather, students’ answers were vague, e.g. “The author knows what he is talking about”. Moreover, 10% percent of the students said the author was not an expert when he/she actually was. Regarding information reliability, only 25% of responses displayed a clear decision and a correct and sufficient

explanation of the students' reasoning. Students provided vague or contradictory answers, such as judging a website reliable and at the same time noting that the information was outdated.

In summary, the evidence regarding adolescents' use of evaluation criteria is mixed and inconsistent. On the one hand, constructs such as topical relevance, readability, author competence or currency of the information seem to make sense to most of them, as evidenced in the studies reviewed above. What is less clear is how they come to an assessment of these dimensions of information. Furthermore, the studies conducted so far have not independently manipulated these criteria in their materials, therefore it is difficult to separate participants' understanding of a criterion from other uncontrolled features of the documents.

The role of task contexts and evaluation prompts

Another dimension that makes it difficult to compare adolescents' use of evaluation criteria across studies is the amount of cueing involved in the research procedures used to collect self-reports or to observe their actual behavior. Surveys and focus groups provide the strongest level of cueing as they directly ask participants about the dimensions of interest (e.g., reliability of information). The behavioral studies reviewed above have used a diversity of procedures ranging from open Internet search (Barzilai & Zohar, 2012) to worksheets asking explicit questions (Coiro et al., 2015). As suggested by the discrepant findings of these studies, adolescents' evaluation behavior may partly depend on the context and framing of the task.

In fact, some studies have directly evidenced the role of contexts and prompts in adolescents' evaluation behavior. In the study by Paul et al. (2017) cited above, the 15 year-old participants cited "conditions for sourcing", such as receiving explicit prompts and reinforcement from the teachers. Zhang (2013a, 2013b) used a digital notepad to prompt sixth-graders to answer questions while skimming, reading and summarizing websites for a

school assignment. The questions included criteria such as topic-match (e.g. “Is this website related to your question?”) as well as source credibility (e.g., “Is the author biased?”). As compared to a control group with no prompts (“Google only”), the prompted students spent more time reading the websites and were more selective of the passages to read (Zhang, 2013b; see also Kammerer, Meier, & Stahl, 2016). However, at the skimming phase, students from both groups assessed the websites quickly and dichotomously (Zhang, 2013a). Students spent fewer than 30 seconds answering the evaluation questions and did not know how to determine whether an author was biased or not. Therefore, they made arbitrary decisions regarding this criterion.

Adolescents' sensitivity to contextual cues and prompts when dealing with information problems is consistent with recent theories of purposeful reading (Britt et al., 2018; McCrudden & Schraw, 2007) and information problem-solving (Brand-Gruwel et al., 2009). In particular, the RESOLV model put forward by Britt et al. (2018) proposes that information users generate goals based on their understanding of the task context and instructions. More precisely, readers are assumed to construct a mental model of the reading context (or "Context model") that includes relevant features of the social and physical environment such as who is assigning the reading task, what are the expectations and consequences at stake, how much time and resources are available and so forth. Those contextual cues are turned into reading goals and decisions that unfold iteratively until the activity outcomes match the perceived expectations (or the expectations are revised). Thus, the various types of questions and prompts used in the research procedures (as well as in the classroom and in other real-life situations) would lead adolescents to generate various reading goals and standards (van den Broek, Bohn-Gettler, Kendeou, Carlson, & White, 2011), including the amount of resources they dedicate to assessing information quality. However, because document features and amount of prompting have not been manipulated systematically in past studies, adolescents'

reliance on contextual prompts when evaluating online documents is still to be determined. The present study attempts to shed further light on the role of prompting on students' evaluation of document features.

Rationale

To summarize some core lessons from research to date, when explicitly asked about the importance of evaluating, for instance in surveys or in focus groups, adolescents generally acknowledge that this is an important part of Web navigation. They also feel rather confident in their ability to detect quality issues. Observational studies of adolescents' search behavior, however, show a rather mixed picture. Although adolescents seem to spontaneously address issues of text relevance and difficulty at least to some extent, their ability to evaluate information sources is uncertain. Source evaluation might be intrinsically difficult for students in the elementary grades, perhaps due to their lack of familiarity with constructs such as "author competence" (see Macedo-Rouet, Braasch, Britt, & Rouet, 2013), but the evidence regarding adolescents between the ages of 11 and 15 is still sparse and inconclusive. In the present study, we asked the following research questions:

Q1. When asked to assess the quality of a Web document, to what extent do adolescents spontaneously notice issues with content (e.g., topic mismatch, difficult text) and source (incompetent author, outdated text)? Based on the extant research, we hypothesized that content-related issues would be easier to detect than source-related issues because they rely on available world knowledge and a default comprehension strategy.

Q2. Does the provision of explicit evaluation prompts (e.g., "Does this document contain information related to your search topic?") help adolescents detect content and source issues in Web documents? Based on empirical (Coiro et al., 2015; Zhang, 2013a) and theoretical research (Britt et al., 2018), we hypothesized that such prompts would facilitate the detection of content issues, but not necessarily of source issues, which depend on students'

knowledge of constructs such as competence and benevolence and the language used to convey those constructs.

To investigate these questions, we conducted two experiments. The procedures and materials were defined so as to be applicable to a quasi-naturalistic classroom setting. Experiment 1 addressed research question #1, by assessing adolescents' spontaneous use of cues when asked to evaluate the quality of a set of web-based documents which all included a specific content- or a source-related issue. Our main expectations were that students would possess the skills needed to identify content-related issues to some extent, but they may have more trouble assessing source-related cues such as author competence or the currency of information. Experiment 2 addressed research question #2. It implemented materials and tasks similar to Experiment 1, but provided the participants with prompts referring to specific content and source criteria. We expected that explicit prompts would improve adolescents' detection of content-related issues more than source-related issues.

Experiment 1

Experiment 1 investigated middle school students' criteria when evaluating the quality of Web-like documents (Rieh, 2002). More specifically, we wanted to know whether students would spontaneously identify some typical quality issues that may be found in Web documents, such as topical mismatch, overly difficult texts, outdated information or incompetent authors. We hypothesized that (1) students would attribute lower ratings to a topically-mismatched text, as compared to the ratings attributed to topically-relevant texts, even though the latter may come with other quality issues. Our rationale was that students' perception of topical relevance is mostly based on basic comprehension processes, which most middle school students are expected to achieve at least with simple texts (Organisation for Economic Co-operation and Development, 2016). In comparison, the detection of source-related issues such as author competence or currency of information would require one's

awareness of source features, which middle school students seem to be typically lacking. The status of text difficulty is less clear. On the one hand, adolescents seem to be able to bring up issues of readability. On the other hand, reader's perception of text difficulty requires some level of comprehension monitoring, which some adolescents may still lack. (2) Students would generate a variety of criteria to justify their ratings, with most criteria being content-related as opposed to source-related. (3) Topic-mismatch and poor readability (i.e. content-related criteria) would be easier to detect in students' justifications as compared to the non-competent author and the outdated text (i.e. source-related criteria). Finally, (4) we anticipated that most students would adequately mention topic match and text difficulty more often than author competence and information currency.

Since one goal of the experiment was to collect and analyze adolescents' explanations about the quality of information, and given the scarcity of data regarding the acquisition of these skills between the ages of 12 and 15, we decided to recruit a panel of students spanning across three grade levels, namely 7th, 8th and 9th grades. We did not seek to test any specific hypotheses regarding skill acquisition, but rather to broaden the range of perceptions and explanations of information quality.

Method

Participants. Participants were 57 students from three intact classes (7^{th}, 8^{th} and 9^{th} grades, respectively) of a public middle school located in a large urban area in France. Their mean age was 14 years (SD=1.1, range 12.5-16), and 24 of them (42%) were female. All participants were French native speakers or had attended a French school for at least seven years. At the time of the study, all students had participated in one of several educational activities related to "sustainable development", which is part of the curriculum since 2004. Despite some familiarity with the topic, students did not seem to know much about climate

change as stated by themselves when asked informally by the researchers. Moreover, they had not received any formal instruction on source evaluation prior to the study. School principal and teachers' consent were obtained prior to the study (see procedure below). All students from the three classrooms were included in the experiment.

Materials. This section presents the materials that were developed for both experiments. The materials consisted of two sets of four documents plus an assessment form, all printed on A4 sheets. In Experiment 1, only the first set of texts was used.

Texts. We developed two sets of four texts each, on two different social-scientific topics: “The causes of climate change: what do we know nowadays?” and “Young people and obesity nowadays: should we stop eating fast-food?” These topics are part of the French curriculum for secondary education and they are likely to be taught in French middle schools (French Ministry of Education, 2015; French National Institute for Health Education, 2012).

The texts were inspired by real websites (e.g., the Wikipedia entry on “climate”, in French), but they were rewritten and shortened to ensure that all texts were equivalent in all but four target criteria: readability, currency, author competence, and topic-match. Each text featured a title (3-7 words), a source description including the name and professional status of the author together with a publication date (month, year), and two to three paragraphs for a total of 145 to 150 words. The content and source features of each text were manipulated so as to create four types of issues with respect to the text's quality as an information resource:

- Text 1 (the "Irrelevant text") was topically mismatched, although the text header shared a keyword with the search topic. In the Climate change document set, Text 1 described the climate zones on Earth. In the Obesity set, it described a video-game called “Fast-food Panick”, in which the goal is to manage a fast-food restaurant quickly and efficiently.

- Text 2 (the "Difficult text") was relevant with respect to the topic, but difficult to read for middle-school students. It was adapted from online presentations at scientific conferences, whereas the other texts were all based on educational or news websites . It contained complex technical terms and long sentences. The Fog index (Gunning, 1952) for the Difficult text was 21.3 (Climate change) and 25.8 (Obesity), as opposed to a range from 13 to 13.8[1] for the other texts.
- Text 3 (the "Unreliable text") was attributed to an author with no explicit credentials in the domain ("non-competent" author). Author competence was manipulated by attributing a profession that was unrelated to the topic to one of the authors (e.g. “a sales-manager computer engineer”, for the topic of Climate change), whereas all the others authors were qualified professionals in the domain (e.g. “a university professor in Atmospheric sciences”).
- Text 4 (the "Outdated text") was outdated as regards current social and scientific status of the topics. It came with a publication year of “1989”, whereas the other texts were all dated from “2014”, i.e., the year before the experiment was run.

Insert Table 1 about here

Half of the attributed author names were male, the other half female. The names and surnames were chosen from among the most common French names, according to the French national institute of statistics and economic studies (Institut national de la statistique et des études économiques, 2014). The texts were embedded in webpage-like, internet browser window. A sample text is featured in Figure 1.

Insert Figure 1 about here

[1] There are no published norms linking readability to reading grade level in French. However, according to Blanger (2007) in French a Fog index of 10 to 14 corresponds to popular magazines whereas 15 to 24 to university and legal texts. https://documix.wordpress.com/2007/06/23/evaluer-la-lisibilite-dun-texte-avec-lindice-de-confusion-de-gunning/

The texts were printed and displayed in a booklet with instructions on the front page. Texts were labeled by number (e.g., “Text 1”) in order to facilitate the completion of the assessment grids in the rating and justification task (see below). However, text labels and presentation order were counterbalanced across students in the same classroom. Half of the students saw texts 1 and 2 first, the other half saw texts 3 and 4 first (relabeled as 1 and 2).

Pilot study. Prior to the experiment, a pilot study was conducted with a panel of 16 education professionals including teachers of different disciplines and school librarians. The purpose of the pilot study was to verify that (a) educated adults with experience in education could easily detect the quality issues that were introduced in the texts and (b) that both sets of texts would yield comparable ratings on the four quality criteria. Participants were recruited through email and face-to-face invitations in various schools and institutions. They were asked to read the instructions, the document set, and to rate the documents according to the four target criteria. They were told that the documents were intended for middle-school students and that they should rate the adequacy of each document for that type of audience. Table 2a presents the average ratings for the document of poorer quality vs. the ratings collapsed across the other three documents ("good"). The results of a 2-way ANOVA with document type (poor, good) and topics as within-participant independent variables and ratings as a dependent measure are summarized in Table 2b.

The participants gave lower ratings to the "poorer" documents on the relevant criterion as compared with their average rating of the other three documents on the same criterion (all ps <.05, Table 2b). For instance, they gave an average topic match rating of 1.1. out of 7 to the document dealing with "Climate zones" compared to an average of 5.1 for the other three documents on the topic of climate change (Table 2a, upper left). For the most part, the overall ratings did not differ across topics. The one exception was that the documents about climate change received lower ratings on "currency" overall compared to those on obesity. This may

be easily explained by the prevalence of the issue in the media and the emphasis on incoming new facts at the time of the study, which may have made documents published a year before look rather old, compared to documents about fast food and obesity. No other effect or interaction was significant.

In short, the pilot study confirmed that education professionals would notice all four types of quality issues, that the topic would have no effect on ratings (save for an overall effect on currency ratings) and would not interact with the type of quality issue implemented in the documents.

Insert Tables 2a and b about here

Based on the teachers' majority opinion about the intrinsic relevance of the two topics with respect to the contents being taught at the time, and because Experiment 1 was to be run in three classrooms at different grades, only the set of documents dealing with climate change was used in Experiment 1. The two sets of documents were used in Experiment 2.

Assessment form. An assessment form was used to collect students' ratings and justifications. Students were asked to assess the goodness/usefulness of each text by answering the following question: "Do you think this is a good text to prepare for your presentation? Would you use it?" For each text, students had to write down the title of the text, rate it in a scale from 0 (not at all) to 7 (yes, certainly), and provide a short written justification for the rating.

Procedure. Several weeks before the experiment, the researchers made contact with the school principal and staff in order to explain the objectives of the experiments and to discuss the details of the procedure. Based on the shared opinion that the proposed experimental materials and tasks had an intrinsic educational value, the decision was made to run the experiment as part of a science class and to offer the tasks to the students as non-mandatory practice activities. The teachers then informed the students about the upcoming

special class session. The experiment was run in a single session of 55 minutes. The students arrived at the classroom with their science teacher, sat at their usual places, and met two researchers. The researchers introduced themselves and explained that they were conducting a research project to find out how middle-school students read and evaluate information from the Internet. The researchers explained that they would ask the students to read and evaluate four short texts on the topic of "the causes of climate change", as if they were preparing for a classroom presentation. The exact instructions were as follows: "Imagine you have to give a 10-minute presentation in your classroom. The title of your presentation is 'The causes of climate change: what do we know nowadays?' In the following pages, you will see four texts retrieved from the Internet through a search engine. Since we do not have much time, you will only see excerpts of these texts. For each one, you will have to say if the text is useful to prepare for your presentation and why." These instructions were printed in the front page of a booklet and were available for students during the experiment. Students were informed that they did not have to answer the questions if they did not want to, and that their participation would not have an impact on their school grades. However, no student declined to participate.

Students were further told that they would have to rate each text on a scale of 0 (not at all) to 7 (yes certainly), and they were given a practice rating task using examples related to food preferences. One researcher drew the 8-point scale (i.e., 0 to 7) on a large white board and asked students "how much they liked chocolate and why". As most students responded "pretty much" and "a lot", the researcher explained that these answers corresponded to six or seven points on the scale. She checked number 7 on the scale and wrote down a few justifications given by the students (e.g. "I like sweet stuff"). Then, the researcher asked 'how much they liked spinach with cream' (NB. a prototypically distasteful meal for many kids in France). As expected, most students responded "not at all", therefore the researcher checked 0 (zero) on the scale and wrote down some justifications (e.g., "the texture is horrible"). These

examples allowed for a demonstration of the rating scale in a relaxed and cheerful atmosphere.

After the practice phase, the researchers distributed the booklet containing task instructions and the four texts. Students were asked to read the four texts attentively, at their own pace, and to call the teacher or one of the experimenters when they felt they were done. Upon calling an experimenter, they were given the assessment form, which they also completed at their own pace, with the texts available. If a student finished the rating and justification task before the end of the session, he/she was given a filler task (e.g. cross-words or Sudoku). On the contrary, if a student was going too slowly, the researchers gently encouraged him/her to speed up. At the end of the session, students were invited to ask questions and make comments about the procedure. They were then thanked and dismissed.

Data analysis. We computed average quality ratings for each text. An ANOVA with texts as within-subjects variable was performed to compare the scores attributed to the four texts. We analyzed the participants' justifications through qualitative and quantitative approaches. We defined a “justification” as the sentence(s) written by the student to justify the usefulness score attributed to each text. Each student provided four justifications in total (one per text). All justifications were transcribed in a separate file and content-analyzed.

Table 3 presents the coding scheme that was developed in order to categorize the justifications[2]. Based on the informal reading of some of the students' responses, we identified seven main categories of justifications (Kvale & Brinkmann, 2009). The first four categories corresponded to the target variables (i.e. topic-match, readability, currency, and author competence). Category 5 captured justifications related to the amount and precision of the information (e.g., "there is a lot of information"). Category 6 reflected students' satisfaction with the information (e.g., helpful, vs. not helpful to complete assignment), whereas Category

[2] The full coding guide (in French) is available upon request to the first author.

7 included statements of opinion with respect to the text contents (e.g., agree/disagree with what the text says). An eighth category ("Other") was created to account for responses that could not be matched with any of the seven other categories.

Insert Table 3 about here

The justifications were coded by assigning each student's response to one or several categories. Most responses included a single propositional clause and thus were assigned to one category. However, when the responses included several clauses separated by a period or a connective (and, in addition, however), the clauses were analyzed separately. Thus, a single response including several clauses could be coded as matching two or - more rarely - three different categories (see examples in the Results section below).

The first author trained a second coder to learn the criteria and apply them to the justifications. Then, the two coders independently coded 15% of the justifications (randomly selected). In cases of disagreement, the coding was discussed until all conflicts were resolved and the coding scheme modified accordingly. Once the coding scheme was finalized, another random sample of 15% the justifications were double-coded. Following Coiro et al. (2015), we calculated inter-rater percent agreement, resulting in a reliability level of 91% or more, according to the category (see Table 3). The remaining justifications where coded by the second coder.

We then calculated the percentage of each category by type of text (i.e. topically-mismatched, least readable, outdated, author not-competent).

Results

Ratings. All texts from the document set received similar scores on average. The topically mismatched text received a mean score of $M = 4.28$ ($SD = 2.28$), the outdated text $M = 4.49$ ($SD = 2.09$), the text by a non-competent author $M = 5.18$ ($SD = 1.80$), and the least

readable text $M = 4.61$ ($SD = 1.98$). These scores were not significantly different ($F(3, 56) = 1.831$, $p = .143$, partial $\eta^2 = .032$).

In sum, the topically mismatched text was not rated as less useful than the other texts. Instead, all four texts were considered to be rather good/useful according to students' ratings.

Justifications. A total of 223 justifications were provided for usefulness ratings (two students did not provide justifications for three of the texts and were excluded from the analysis). Of these, 180 (81%) contained one category from the coding scheme, 38 (17%) two categories (e.g., "Because it tells us what's the cause [1) Topic match], but it is not well explained [2) Readability]", S40) and 5 (2%) three or more categories (e.g. "The text doesn't cite the greenhouse effect gases and says it's the industry who is responsible for warming up the Earth, but why? [1) Topic match] Incomplete text [5) Amount and precision] and too old [6) Currency]", S22). All in all, there were 271 codes from the coding scheme attributed to the whole sample.

As shown in Table 4, students cited a variety of criteria when justifying their ratings. The most frequent criteria overall were topic-match (36.2% of the codes), readability (27.3%), amount and precision of information (14.4%) and satisfaction (10.3%). Source-related criteria were rare, with author being mentioned in only 2.2% of the justifications and currency in 0.7% of the justifications. Other criteria occurred in 6.3% of the justifications.

Insert Table 4 about here

The content analysis further revealed that students sometimes disagreed on the direction (positive or negative) of their evaluation. For instance, 19 students (correctly) stated that the topically-mismatched text did not match the topic of the query, but 7 students actually stated the opposite (e.g. "I find this one useful [on] the types of climate, I think with this one I can make my presentation", S17). Moreover, 9 students found that this text was readable (e.g., "I gave five because I can understand this text", S1), whereas 5 students found it difficult to

understand (e.g., “It’s hard to understand”, S30). Seven students expressed general satisfaction (e.g. “I liked this text”, S23) and 2 were dissatisfied (e.g. “It didn’t help me”, S33).

Regarding the least readable text, 18 students considered that the text was difficult to understand, but 7 students actually found it easy to understand (e.g. “It explains well the topic of the presentation, it’s easy to understand”, S6). Moreover, 19 students considered that the text was topically relevant or matched with the task, whereas 4 considered it unmatched or not relevant, and 6 students considered that the text provided a fair amount of information, whereas 2 students said the opposite.

For the outdated text, 19 students thought the topic was relevant and 9 not relevant; 11 that it was easy to understand and 5 difficult to understand; 4 that it had a fair amount of information and 7 the opposite. Only two students noted the date of the text and justified their ratings as follows: “Because it’s in 1989” (S55), “(…) too old” (S22).

Concerning the text authored by a non-competent author, 18 students found it topically relevant, 5 topically irrelevant; 13 stated that it was easy to understand, and 6 that it was difficult to understand; 7 found the document helpful whereas 2 found it unhelpful. Four students mentioned the author in their justifications: two of them mentioned characteristics of the author (“Because it’s an informatics engineer”, S55; “Good text because an engineer wrote it…”, S28), one simply mentioned the name of the author (“Sylvie Renaud”, S48) and the fourth one was vague (“It gives us who is responsible [for climate change] and the person who wrote is part of it”, S3). A single student (S55) noticed that the author was not a professional in the domain of climate sciences.

In sum, there were contradictory assessments for each of the texts. Taking into account these assessments, we calculated the percentage of students who identified the target criterion

(e.g., readability for the poorly readable text), in the expected and in the opposite direction, for each type of text (see Figure 2).

Insert Figure 2 about here

As shown in figure 2, the percentage of students who cited the target criterion in the expected direction was 32.7% for the low readable text, 33.3% for the topic-mismatched text, 3.6% for the text by a non-competent author and 3.6% for the outdated text. The other students cited either the target criterion in the opposite direction, or other criteria. Chi-squared analysis shows that the distribution of answers was not independent of type of text ($X^2(6) = 53.17$, $p < .0001$). The results suggest that the proportion of students who identified the target criterion in the expected direction was lower for the outdated text and for the text by a non-competent author, as compared to the least readable and the topic-mismatched text.

Discussion and limitations

Consistent with the literature, Experiment 1 showed that identifying quality issues in a set of multiple documents is a challenging task for adolescents. A mere third of the students in our sample (33%) detected the topically-mismatched text (about "climate zones", not "climate change") from among a set of four texts. Not only did students attribute fairly high goodness ratings to this text, but most of them did not mention the topic mismatch when justifying their ratings. Seven students even stated explicitly that the text was topically relevant for their assignment.

The same pattern holds true for the other documents and evaluation criteria. Sixty-eight percent of the students did not detect that the text issued from a scientific conference on climate change had complex vocabulary and syntax, although readability was frequently mentioned in the justifications (indicating that students were aware of this criterion). As predicted, the most undetected issues were those that belonged to source parameters: author and date. Only 3.6% of the students noticed that the author of Text 3 was a sales-manager

computer engineer, thus not competent *a priori* on matters of climate change (which is outside of his/her "sphere of authority"; Wilson, 1983), and that Text 4 dated back to 1989, which is outdated regarding the topic of the assignment: "The causes of climate change: what do we know *nowadays*?" (our emphasis).

Interestingly, students did not always agree on the direction of the assessment for each criterion and type of text. There were contradictory evaluations of topic-match, readability, and author competence for each text. This suggests that students generated various understandings of the task at hand (Britt et al., 2018), and that the standards for validating information in the context of a school assignment were not shared by all of them. The topic of climate change may have prompted such disagreements because it is a scientific topic, somewhat familiar to adolescents, but also notoriously controversial and difficult to understand (Strømsø, Bråten & Britt, 2011).

Beyond the target criteria, students cited several other criteria in their justifications. Most of them were based on the content of the texts, such as finding a text useful because it provides "a lot of information", or on personal opinion, as when the student "agrees with the author" or simply "likes" the text. Justifications based on source parameters were much less frequent, and students often relied on what the author said to assess author competence, an approach previously reported in a study with younger students (Macedo-Rouet et al., 2013). It should be noted, however, that the difference between the competent and less competent authors may have been too small for students to provide contrasted assessments, as the phrase "computer engineer" suggested a person with a high level of education.

These results point to the complexity of multiple text evaluation for adolescents and corroborate the results of previous studies that used similar materials (Coiro et al., 2015). When asked "Do you think this is a good text to prepare for your presentation? Would you use it?" based on a set of available documents, many adolescents are unable to produce

accurate judgments of information usefulness. They rely mostly on the content of the text and seldom refer to source cues such as author's credentials or publication date. In addition, different students may come to contradictory conclusions regarding the same criterion about a given document.

On the other hand, it can be argued that the task was difficult for the students because there was little guidance on "what to assess" (Paul et al., 2017; Zhang, 2013b). The assessment question was broad and did not provide students with evaluation criteria that might help them focus on specific aspects of the texts (Britt et al., 2018). Moreover, students had to write down their justifications, a requirement that may lead to poorer performance as compared to oral answers because of the intrinsic difficulty of articulating written answers (Huxham, Campbell, & Westwood, 2012). Had students been asked to explain their answers with explicit guidance on "what to assess", they may have produced more accurate evaluations of each document. Experiment 2 was conducted to investigate this additional hypothesis.

Experiment 2

Experiment 2 used a mixed methods approach associating an interview procedure together with a ratings questionnaire (McCrudden, Stenseth, Bråten, & Strømsø, 2016). The main goals of Experiment 2 were (a) to examine the possible influence of the task context on students' qualitative evaluation of a set of documents; (b) to examine whether criterial prompts, would help the participants produce more accurate ratings distinguishing the "better" and the "poorer" documents on each criterion.

We anticipated that a spoken response mode would help some students focus their attention on the text and notice the information issues to a greater extent than in Experiment 1. Furthermore, we expected students to rate the "poor" document significantly lower than the other documents on the specific criterion matching the quality issue, and to cite target criteria

more frequently than in Experiment 1. Additionally, Experiment 2 sought to explore the effects of age and educational level on students' evaluation of information quality. We tested the broad assumption that high-school students would perform better than middle-school students, particularly as regards the evaluation of source criteria (Livingstone, Haddon, Görzig, & Ólafsson, 2010; Metzger et al., 2015). Finally, in order to broaden the scope of the findings, a second set of texts dealing with a different topic was included in the materials.

Method

Participants. Participants were 36 students from nine public schools located in a midsize urban area in France. We distinguished two subgroups as a function of students' age and school level[3]: (a) 21 students from middle school (6th and 8th grades, mean age 12.5 years, *SD*=0.95, 67% female); (b) 15 students from college-bound track high school (10th grade, mean age 15.8 years, *SD*=0.53, 33% female). All participants were native speakers of French. Some students were invited to participate by the school librarian, others were recruited "on the fly" during our visits to the library. Volunteers participated during "study hours", i.e. periods in between two classes. During study hours, students either work independently in a study room or attend the school library. These periods allowed us to meet a variety of students (i.e., not just regular library users), at the library on an individual basis. As in experiment 1, students had participated in one or several educational activities related to "sustainable development", which includes climate change, but they had not received any formal instruction on source evaluation, prior to the study. Head-teacher and the school librarian's consent was obtained prior to the study. Since the procedure was run during school hours and in the presence of the school librarian, parental consent was not requested.

[3] In the French school system middle and high schools differ substantially. Middle-schools (*college unique*) are non-selective and mandatory until 16 years of age, whereas high schools (*lycée*), are selective and track-based. The general track emphasizes academic disciplines and skills, including information literacy; the vocational track involves a mix of academic and vocational training. Therefore, college-bound high-schoolers have received more opportunities to acquire advanced information skills as compared to middle-schoolers on average, in addition to the increased maturity and world knowledge, among other differences.

Materials. As in Experiment 1, the materials consisted of a set of instructions, a document set and an assessment sheet. However, there were a number of differences meant to disentangle competing interpretations of students' performance as drawn from Experiment 1.

Texts. The two sets of documents initially developed for the study (see Materials section of Experiment 1) were used in Experiment 2 to check whether the results obtained would replicate across topics. Each topic was assigned to every other student in each school.

Moreover, the non-competent author in Text 3 (both topics) was changed to "a secondary school student" in order to create a larger contrast with the other authors in the document set, and thus increase the chances that students may detect the author competence issue. We assumed that if students are aware of issues related to author competence they should easily notice the contrast between a peer and a qualified professional.

Assessment form. The assessment form used in Experiment 2 comprised four specific questions on the target evaluation criteria (topic-match, readability, author competence, currency) instead of a general question on the goodness/usefulness of the documents. The exact question wordings were as follows:

- Is this document really about the topic of your assignment? (Topic-match)
- Is this document easy to understand for a student like you? (Readability)
- Does the author of the document have a lot of knowledge on the topic? (Author competence)
- Does this document present recent information about the topic? (Currency)

After each question, students had to rate the four texts on a scale of 0 (not at all) to 7 (yes, certainly).

Procedure. The experiment was run individually in a single session of about 50 minutes, at the school library. Participants were met by a researcher, and they sat at a table in a quiet area of the library, while the librarian was attending to other students. The researcher

explained that the interview was part of a research project whose goal was to find out how adolescents search and evaluate information from the Internet. The interview began with questions about students' experience in searching information on the internet and at the school library. The researcher followed a semi-structured interview protocol, in which a pre-defined set of questions were asked, with other, clarification questions being added if students' answers prompted it (e.g., after vague or incomplete answers). Therefore, there was a back-and-forth process in which the researcher and the student discussed the points raised by the later. The interview lasted about 20 minutes on average.

Once the interview was over, the researcher introduced the document evaluation task. The researcher read out loud the instructions for the student, made sure they were clear, and asked the student to read the four documents attentively. The students read the documents at their own pace without taking notes. Then, the interviewer asked about each document: "Do you think this is a good text to prepare for your presentation? Would you use it?" If a student answered the question simply by "yes/no", the researcher prompted him/her to elaborate. During the procedure, students were allowed to get back to the documents and add to their response concerning previous documents. No additional questions were asked by researcher to avoid influencing students' answers to the subsequent tasks. All interviews were conducted individually and audio-taped.

Once the student had commented orally on each text, the researcher handled him/her the assessment form and explained: "To help you specify <u>why</u> a document is good/not good, please rate the document on a scale from 0 to 7 on each of these questions". The student completed the form individually and silently, and handled it back to the researcher. Finally, the student was thanked and dismissed. The time spent on the written document evaluation task was about 30 minutes.

As in Experiment 1, students were informed that they did not have to answer the questions if they did not want to, and that their participation would not have any impact in their school grades.

It should be noted that, in Experiment 2, students justified their evaluations before rating the documents. This was done in order to ensure that students' ability to detect quality issues orally would not be influenced by the provision of evaluation questions. In sum, justifications were formulated without prompts for evaluating specific criteria, whereas ratings were guided by such prompts.

Data analysis. Justifications were transcribed and content analyzed using the same coding scheme as in Experiment 1.

Ratings were analyzed in the same way as in the pilot study. As a first step, we conducted t-tests on each criterion to check for a potential main effect of topic on the ratings attributed to the documents. No significant effects were found (all $p > .10$). Therefore the two topics (fast food & obesity, and climate change) were merged in the subsequent analyses. Next, mean scores for each evaluation criterion were computed and compared using two-way ANOVA with school level (middle school vs. high school) as between-subject factor and document type (poor vs. good) as a within-subject factor. As shown in Table 1, each document was "poor" in one criterion (topic-match, readability, author competence, or currency), and "good" in the remaining criteria. Therefore, all students read the four document types, each one being "poor" in one criterion.

Results

We present the results of Experiment 2 in an order consistent with the procedure: First, we present students' justifications when answering the holistic evaluation question about each document as part of the interview. Then, we present students' ratings for each document on each criterion based on the written evaluation form.

Justifications. A total of 144 justifications were provided by students as answers to the question "Is it a good document for your assignment? Would you use it?" Of these, 71 (49%) contained one category from the coding scheme, 56 (39%) two categories (e.g., "I don't know if I would use it because I don't understand some words very well, otherwise it's very complete", S33), and 17 (12%) three or more categories (e.g. "This text seemed interesting to me, they give a lot of numbers, it's about fast food… but then I remembered that before I had checked if the text was recent, then I saw '1989' so… the text is interesting but I don't think I would use it", S37). All in all, there were 244 codes from the coding scheme attributed to the whole sample. The proportion of justifications with two or more categories from the coding scheme was greater than in Experiment 1, where most of the justifications contained only one code. This provides initial evidence that response modality may affect adolescents' communication of their rationale for evaluating documents.

Students cited a variety of criteria in their justifications. As shown in Table 5, the most frequent criteria overall were topic match (29.6% of the codes), readability (15.6%), satisfaction (15.6%), and amount and precision of information (12.3%). Source-related criteria were scarce, although they were more frequently cited than in Experiment 1. Author competence accounted for 11.1% of the codes (against 2.2% in Experiment 1) and currency for 5.8% of the codes (against 0.7%).

Insert Table 5 about here

The accuracy with which students assessed the target criteria varied according to the type of criteria (figure 3). Whereas most of the students (75%) mentioned the topic mismatch for the topically-mismatched text, and 55.6% the poor readability for the poorly readable text, only 13.9% detected the outdated text. Indeed, when justifying their evaluations of the outdated text, 86.1% of the students cited "other" criteria.

Insert Figure 3 about here

Regarding the text authored by a non-competent author, 36.1% of the students cited the target criterion in the expected direction. For instance, they explained that "[the author] is a high-school student [publishing in] a personal blog so… it's not sure, not official" (S15, high schooler), or that "it's a student who 'speaks', she has done less studies" (S28, high schooler), and "I see that the author is a student, we don't know where she got her information from" (S11, high schooler). One student (2.8%) assessed author competence in the opposite direction: "It's good because it's a student, we understand it better" (S09, middle schooler).

In sum, when asked for evaluations as part of an interview, most of the students were able to identify issues with the content (topic-mismatch) and language (readability) of the text even before they received prompts to evaluate these criteria. However, only a minority was able to detect the issue in source parameters although the percentages were higher than in Experiment 1.

Ratings. The mean scores attributed to the "poor" and the "good" texts, for each evaluation question, are shown in table 6.

Insert Table 6 about here

Two-way ANOVAs with school level as a between-subject variable and type of document (poor, good) as a within-subject variable were conducted for each of the four evaluation questions. For topic-match, there was no significant main effect of school level (middle vs. high school) ($F(1,34) = 0.001$, $p = .972$, partial $\eta^2 = .000$), a significant main effect of type of document (poor vs. good) ($F(1,34) = 157.90$, $p < .001$, partial $\eta^2 = .823$), and no significant interaction ($F(1,34) = 0.06$, $p = .81$, partial $\eta^2 = .002$). For readability, there was no significant main effect of school level ($F(1,34) = 2.66$, $p = 0.112$, partial $\eta^2 = .073$), a significant main effect of document type ($F(1,34) = 67.68$, $p < .001$, partial $\eta^2 = .666$), and no significant interaction ($F(1,34) = 1.95$, $p = .172$, partial $\eta^2 = .054$). For author competence, there was a marginally significant effect of school level ($F(1,34) = 3.36$, $p = .076$, partial $\eta^2 =$

.090), a significant main effect of type of document ($F(1,34) = 13.08$, $p = .001$, partial $\eta^2 = .278$), and a significant interaction ($F(1,34) = 11.58$, $p = .002$, partial $\eta^2 = .254$). For currency, there was no significant main effect of school level ($F(1,34) = 0.01$, $p = .914$, partial $\eta^2 = .000$), a significant main effect of type of document ($F(1,34) = 33.68$, $p < .001$, partial $\eta^2 = .498$), and no significant interaction ($F(1,34) = 2.82$, $p = .102$, partial $\eta^2 = .077$).

In sum, students at both educational levels attributed significantly lower scores to the "poor" document as compared to the good documents regarding three criteria: topic-match, readability and currency. Concerning the criterion of author competence, there was a significant type of text per educational level interaction as shown in Figure 4.

Insert Figure 4 about here

Only high-school students attributed significantly lower scores to the non-competent author as compared to the average score attributed to the competent authors. Thus, the evaluation question on author competence was effective in prompting high-schoolers, but not middle-schoolers, to differentiate between poor and good documents on this specific criterion.

Discussion

As in Experiment 1, most of the evaluation criteria spontaneously cited by the students in Experiment 2 referred to the content of the texts. These included not only topic-match and readability, but also the amount and precision of information provided by the text (e.g. "[the text contains] a lot of numbers"). In contrast, references to the source of the information were still sparse, although more frequent than in Experiment 1. The data support the view that when asked if a document is "good/useful" for a school-type of assignment, teenage students tend to focus what the text "says", not to "who is the author" or "when" the text was published (see Paul et al., 2017, for convergent evidence).

Most students were able to distinguish "poor" from "good" texts on the basis of content and source criteria when they were provided with evaluation questions targeting these

criteria (e.g. by asking students if the author has a lot of knowledge about the topic). The evaluation questions led students to attribute lower ratings to the "poor" text as compared to the "good" texts for criteria of topical match, readability, and currency. However, evaluating author competence remained challenging for the middle school students (6th and 8th grades). Contrary to their high-school counterparts, those students were not able to distinguish a competent from a less competent author even when prompted with specific questions. As we further argue in the general discussion section below, our results suggest that younger students need more than simple prompts to show an ability to evaluate the reliability of information sources. Because the comparison of middle school and high school students involves a large number of confounded factors, no specific interpretation can be derived from our data. However, intervention studies (e.g., Britt & Aglinskas, 2002; Pérez et al., 2018; Wiley et al., 2018) suggest that specific instruction into what makes a good source may be required for students to develop that type of skill.

Experiment 2 also suggests that allowing students to discuss information quality in an interview as opposed to through written responses increases their detection of quality issues that pertain to the content of the text, and to some extent to source parameters. Students used a variety of criteria to justify their evaluations of each text from the document set, with most of them citing two or more criteria in their justifications. Several explanations may be put forward. On the one hand, the oral modality may promote students' articulation of a broader set of criteria, because students could invest more effort in analyzing the documents, not in constructing written answers constrained by a small response box as in Experiment 1. Participants in Experiment 2 could speak as much as they wanted, even though most answers turned out to be relatively short. Another, compatible explanation is that the face-to-face interaction with the interviewer added some pressure for the students to scrutinize the information more closely. To put it in the terms of the RESOLV model (Britt et al., 2018), the

interviewer may become part of the students' Context model, and the perspective of communicating responses to an adult may raise their standards of coherence and quality (van den Broek et al., 2011). Be that as it may, our study further supports the view that the information evaluation skills demonstrated by adolescents depend in part on the setting of the experimental situation, including the presence of an addressee, the provision of questions and prompts, and the format of the responses.

An obvious limitation to Experiment 2 is the small number of students involved and the use of two different document sets across participants. Despite the rather clear pattern of results at a descriptive level, this obviously limits the scope of any statistical inference from the data. Therefore, the conclusions from Experiment 2 are to be considered tentative pending on replication with a larger group of students.

General discussion and conclusions

The main goal of the two experiments presented in this paper was to assess adolescents' ability to evaluate Web documents and detect quality issues, under different conditions. Quality issues can arise due to a topic mismatch between the document and the query, poor readability in relation to readers' profile (e.g. documents that are difficult to understand for a adolescent because of technical jargon), an author who has no or little knowledge of the topic, and a publication date that makes the document too old to accurately address the topic, among a long list of other criteria (Britt et al., 1999; Rieh, 2002). Past research has provided many examples of such issues on the Web, and ample evidence that adolescents run a high risk of missing them, partly due to their use of inconsistent and superficial cues (Rouet et al., 2011; Foss et al., 2013; Gasser et al., 2012; Julien & Baker, 2009), and their insufficient knowledge of documents as communication devices (Coiro et al., 2015; Mason et al., 2014; Wineburg & Mcgrew, 2017). However, by implementing a procedure in which a specific set of criteria and the provision of prompts and response format

were systematically varied, our study sheds additional light into the conditions that might promote criteria-based evaluation and criteria learnability.

In the first experiment, we asked middle-school students to read and evaluate a set of four texts from the Web in order to determine how good and useful each of them was to prepare for a school presentation on the topic of climate change. Each text contained a quality issue (topic-mismatch, poor readability, author not competent, outdated), although all but one (i.e., the topically-mismatched text) provided adequate information to answer the query. Students provided ratings and written justifications for each text. The results showed a surprisingly low detection rate of both content- and source-related quality issues. Most participants did not even distinguish the topically-mismatched text from the other texts in terms of ratings, and that they seldom mentioned the quality issue that had been built in each text when justifying their ratings. Most of the justifications were content-based and inconsistent across participants. For instance, some students mentioned that the topically-mismatched text was not about the topic of the query whereas other students stated the opposite. In line with research based on the actual observation of adolescent behavior under quasi-spontaneous Web search conditions, Experiment 1 confirmed that students have a low ability to spontaneously identify specific criteria that may affect the quality of information. This led us to the hypothesis that this kind of deliberate, purposeful evaluation (Britt, Richter & Rouet, 2014) needs to be supported by explicit guidance on "what to assess" and appropriate conditions on "how to assess" (Britt et al., 2018; Paul et al., 2017).

Based on the RESOLV theory of purposeful reading (Brit et al., 2018) and on evidence from prior research (e.g., Mason et al., 2014; Pérez et al., 2018; Wiley et al., 2009), we hypothesized that students would be better able to detect information issues if they were guided by evaluation prompts. In Experiment 2, students were asked to perform the same evaluation task in the context of a face-to-face interview. The participants rated the

documents based on specific questions on each of the four criteria (topic-match, readability, author competence, currency). We also included a second topic in the document set in order to diversify the materials and avoid the risk that evaluation difficulties be linked to a specific topic. Finally, we included participants from two school levels (middle school and high school) in order to account for potential differences in information skills between these two groups of adolescents.

Participants in Experiment 2 spontaneously cited more (and more accurate) criteria per justification than in Experiment 1, and generated less contradictory assessments of the same texts. Moreover, they were better able to distinguish “poor” and “good” texts based on explicit criteria. An exception concerns the criterion of author competence. Only high-schoolers were able to detect the non-competent author by attributing lower ratings to the “poor” text on this particular criterion. Middle-schoolers did not attribute significantly different ratings to the poor and the good texts on this criterion. These results are consistent with previous studies (Coiro et al., 2015; Coiro & Dobler, 2007; Zhang, 2013a, 2013b). They suggest that unlike other criteria such as topic-match or readability, evaluation prompts are not enough to enable middle school students' evaluation of source dimensions. Students at this level may need to receive additional training in how to read and interpret source descriptions (Britt & Aglinskas, 2002; Pérez et al., 2018; see also Brante & Strømsø, 2018, for a review). This conclusion is in line with Britt et al.'s (2018) assumption that the implementation of reading goals requires preexisting goal-specific knowledge and heuristics.

The interview format used in Experiment 2 may have prompted both a deeper processing of the materials and increased verbalization, since students could concentrate on the evaluation of texts, not on writing. Indeed, research suggests that the oral modality can be more useful than the written modality in order to foster student performance in exams (Huxham et al., 2012) and to explore students’ epistemic processing of online texts (Cho,

Woodward, & Li, 2018). In our study, it is not possible to disentangle the role of oral modality from other variables, such as students' age and school level. Knowing that students' ability to evaluate online information develops from 6th to 10th grades, but also depends on individual variables (Pérez et al., 2018; Sbaffi & Rowley, 2017), this hypothesis should be tested within the same school level in future experiments.

Overall, the present study shows that even basic criteria such as topic match are not uniformly assessed by adolescents when they evaluate Web documents by answering general questions such as "Is this text good for your assignment? Would you use it?" using a written response mode (e.g. via assessment sheets). For instance, adolescents may be misguided by keywords in the text and evaluate a topically mismatched document as useful for their school assignment. However, when provided with explicit questions to evaluate specific dimensions of information quality and credibility, adolescents' assessment of text content is improved, although the assessment of information sources is still problematic.

We believe that our findings have interesting instructional implications. They suggest that the provision of specific evaluation questions improves students' evaluation of multiple documents. Evaluation questions are an instructional approach that teachers can easily implement in their classroom activities, as well as for homework purposes. Our findings suggest that teachers should encourage students to express their evaluations orally and informally during document-based activities. The oral modality may facilitate the detection of quality issues in Web documents because students can concentrate in the analysis of the texts, not on the construction of their written answers.

This is not to state, however, that the provision of scoring sheets is all students need to develop effective evaluation skills. In fact, there is no evidence that prompts have an instructional effect *per se*. On the opposite, intervention studies conducted with secondary school students (e.g., Britt & Aglinskas, 2002; Mason et al., 2014; Pérez et al., 2018; Wiley et

al., 2018) suggest that students need both explanations, guided practice, application and feedback, perhaps over several sessions, in order to learn transferable skills.

The present study has a number of limitations that would need to be addressed in future research. First, the samples were small and the results require confirmation using larger groups of participants. Second, we did not measure the participants' prior knowledge on the documents' topics, yet prior knowledge is known to influence readers' evaluation abilities (Wopereis & van Merriënboer, 2011). Third, we manipulated the provision of prompts (evaluation criteria) only across experiments. Although difficult to achieve in a naturalistic school setting, a manipulation of prompting and/or response format as part of the same procedure, either between or - better- within participants, would provide a stronger test of our hypothesis. Future research should also aim at disentangling the various dimensions that differentiate the procedures of Experiment 1 and 2 (e.g., participating individually as opposed to in a classroom, having an interviewer present, speaking as opposed to writing and so forth). Our comparison of middle school and high school students is also of limited value, because of the many factors, ranging from developmental to educational to experiential, that may explain the better performance of the latter students. For instance, students may have different levels of motivation and interest in the topics, which might influence their evaluation performance. Although all participants in our study accepted and seemed to enjoy the tasks and topics, we did not include an objective measure of their motivation and interest, which precludes any firm conclusion as to their potential effect. Finally, the texts used in the study were simplified to make them comparable on the basis of the same criteria and avoid distractors. However, authentic Web documents are usually much richer in textual and pictorial information, and we do not know whether these characteristics contribute negatively or positively to the evaluation of source parameters. Additionally, the fact that students read web-like documents on paper constitutes a limitation of our study, because print and screen reading imply a number of

differences in text legibility, which may influence information evaluation as well (Murphy, Long, Holleran, & Esterly, 2003). At the same time, other researchers have argued that students' ability to evaluate online sources can be measured offline with print materials that simulate web pages (Stanford History Group, 2016). Their results suggest that students pay little attention to sources, both offline and online. The study by Coiro et al. (2015) suggests that authentic materials do not lead to a more accurate evaluation of source criteria, but future studies should explore this issue further. In the same vein, it could also be interesting to include a document without any obvious "flaw" to obtain a baseline rating.

There is little doubt that generalized access to information technology provides a host of opportunities for students to learn and reflect on a v ariety of topics relevant to their school and out of school interests. This wealth of information, however, comes at the cost of increased demands on students' ability to critically evaluate the quality of information both content- and source- wise (Britt et al., 2018). This study adds to a growing body of evidence showing that mere exposure to digital media is not enough for students to develop these skills. Although explicit prompts improved participants' detection of content-related issues, they failed to support accurate assessments of author competence. The findings warrant an increased effort to design relevant instructional interventions. In particular, source evaluation skills (i.e. the ability to evaluate the source, defined as information about the origin of a document and the circumstances of its production; Bromme, Stadtler & Scharrer, 2018; Rouet & Britt, 2014) should be considered a central component of information literacy. Students need to learn how to read "laterally", such as information professionals, in order to evaluate sources and decide whether they deserve credibility (Wineburg & McGrew, 2017). For adolescents at the middle-school level, interventions should aim at fostering adolescents' evaluation skills. Recent studies (e.g. Pérez et al., 2018; see also Brante, & Strømsø, 2017, for a review) suggest that this is a rather realistic objective.

Acknowledgements

This study was supported by a joint grant from the Agence Nationale de la Recherche (grant ANR-12-FRAL-0015-01) and the Deutsche Forschungsgemeinschaft (grant STA 1291/1-41). Data collection was conducted in compliance with local regulations for educational research at the time of conducting the field studies. We thank all participating teachers, staff, and students for their kind cooperation. Special thanks to Isabelle Sarfati, Symphorienne Suaudeau, Agathe Cornil, and Abdoulaye Coulibaly for their assistance with data collection and coding.

References

Barzilai, S., & Zohar, A. (2012). Epistemic thinking in action: Evaluating and integrating online sources. *Cognition and Instruction*, *30*(1), 39-85.

Blikstad-Balas, M., & Hvistendahl, R. (2013). Students' digital strategies and shortcuts. *Nordic Journal of Digital Literacy*, *8*(01-02), 32-48.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education*, *53*(4), 1207-1217.

Brante, E. W., & Strømsø, H. I. (2018). Sourcing in Text Comprehension: a Review of Interventions Targeting Sourcing Skills. *Educational Psychology Review*, *30*, 773-779. https://doi.org/10.1007/s10648-017-9421-7

Brem, S.K., Russell, J., & Weems, L. (2001). Science on the Web: Students' evaluation of scientific arguments. *Discourse Processes*, *32*, 191-213.

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, MJ: Lawrence Erlbaum Associates.

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

Britt, M.A., Rouet, J.-F., & Durik, A. (2018). *Literacy beyond text comprehension: A theory of purposeful reading*. Routledge, UK.

Bromme, R., Stadtler, M., & Scharrer, L. (2018). The provenance of certainty: Multiple source use and the public engagement with science. In: J. L. G. Braasch, I. Bråten, & M.

T. McCrudden (Eds.), *Handbook of multiple source use* (pp. 269-284). New York, NY: Routledge.

Boubée N. (2007). L'image dans l'activité de recherche d'information des élèves du secondaire : Ce qu'ils en font et ce qu'ils en disent [Role of pictures in secondary school students' information search: what they do and what they say about them]. *Spirale*, *40*, 141-150.

Boubée, N. (2014). The cross self-confrontation method and challenges in researching the active information-seeking of young people. *Libraries in the Digital Age (LIDA) Proceedings*, 13. Retrieved from http://ozk.unizd.hr/proceedings/index.php/lida/article/viewFile/124/230

Castek, J., Coiro, J., Henry, L. A., Leu, D. J., & Hartman, D. K. (2015). Research on instruction and assessment in the new literacies of online research and comprehension. In S. R. Parris & K. Headley (Eds.), *Comprehension instruction: Research-based best practices* (3rd ed., pp. 324-344), New York: Guilford Publications.

Cho, B. Y. (2014). Competent adolescent readers' use of Internet reading strategies: A think-aloud study. *Cognition and Instruction*, *32*(3), 253-289.

Cho, B. Y., & Afflerbach, P. (2015). Reading on the Internet. *Journal of Adolescent & Adult Literacy*, *58*(6), 504-517.

Cho, B. Y., Woodward, L., & Li, D. (2018). Epistemic processing when adolescents read online: A verbal protocol analysis of more and less successful online readers. *Reading Research Quarterly*, *53*(2), 197-221.

Chokshi, N. (2017, September 18). How to Fight 'Fake News' (Warning: It Isn't Easy). *The New York Times*. Retrieved from http://www.nytimes.com

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.

Coiro, J., & Dobler, E. (2007). Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly*, *42*(2), 214-257.

Connaway, L. S., Dickey, T. J., & Radford, M. L. (2011). "If it is too inconvenient I'm not going after it:" Convenience as a critical factor in information-seeking behaviors. *Library & Information Science Research*, *33*(3), 179-190.

Cool, C., Belkin, N.J., Frieder, O., & Kantor, P. (1993). Characteristics of texts affecting relevance judgments. *Proceedings of the 14th National Online Meeting* (pp. 77–84).

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, *23*(1), 84-101.

Eastin, M. S. (2008). Toward a Cognitive Development Approach to Youth Perceptions of Credibility. In M. J. Metzger and A. J. Flanagin (Eds.). *Digital Media, Youth, and Credibility* (pp. 29–48). Cambridge, MA: The MIT Press. doi: 10.1162/dmal.9780262562324.029

Flanagin, A. J., & Metzger, M. J. (2010). *Kids and credibility: An empirical examination of youth, digital media use, and information credibility*. Cambridge: MIT Press.

Francke, H., Sundin, O., & Limberg, L. (2011). Debating credibility: the shaping of information literacies in upper secondary school. *Journal of Documentation*, *67*(4), 675-694.

French Ministry of Education (2015). *L'édu cation au développement durable*. Retrived from http://www.education.gouv.fr/cid205/l-education-au-developpement-durable.html

French National Institute for Health Education (2012). *Icaps, un dispositif « qui marche »… et qui fait « bouger » les pratiques*. Retrieved from http://inpes.santepubliquefrance.fr/transfert-connaissance/tc-activites/soutien-icaps.asp

Foss, E., Druin, A., Yip, J., Ford, W., Golub, E., & Hutchinson, H. (2013). Adolescent search roles. *Journal of the Association for Information Science and Technology*, *64*(1), 173-189.

Gasser, U., Cortesi, S., Malik, M. M., & Lee, A. (2012). *Youth and digital media: From credibility to information quality*. Berkman Center for Internet & Society. Retrieved from https://dmlcentral.net/wp-content/uploads/files/youthanddigitalmediacredibilityreport2.16.12.pdf

Gray, N. J., Klein, J. D., Noyce, P. R., Sesselberg, T. S., & Cantrill, J. A. (2005). Health information-seeking behaviour in adolescence: the place of the internet. *Social science & medicine*, *60*(7), 1467-1478.

Goldman, S. R., Braasch, J. L., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from Internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly*, *47*(4), 356-381.

Grootens-Wiegers, P., De Vries, M. C., Vossen, T. E., & van den Broek, J. M. (2015). Readability and visuals in medical research information forms for children and adolescents. *Science Communication*, *37*(1), 89-117.

Gunning, R. (1952). *The Technique of Clear Writing*. New York: McGraw-Hill.

Hargittai, E. (2010). Digital na (t) ives? Variation in internet skills and uses among members of the "net generation". *Sociological inquiry*, *80*(1), 92-113.

Hargittai, E., Fullerton, L., Menchen-Trevino, E., & Thomas, K. Y. (2010). Trust online: Young adults' evaluation of web content. *International journal of communication*, *4*, 468–494.

Hargittai, E., & Hinnant, A. (2008). Digital inequality: Differences in young adults' use of the Internet. *Communication research*, *35*(5), 602-621.

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, *44*(4), 1467-1484.

Huxham, M., Campbell, F., & Westwood, J. (2012). Oral versus written assessments: a test of student performance and attitudes. *Assessment & Evaluation in Higher Education*, *37*(1), 125-136.

Institut national de la statistique et des études économiques (2014). *Fichier des prénoms – État civil*. Retrieved from https://www.insee.fr/fr/statistiques/2540004#documentation

Julien, H., & Barker, S. (2009). How high-school students find and evaluate scientific information: A basis for information literacy skills development. *Library & Information Science Research*, *31*(1), 12-17.

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://dx.doi.org/10.1016/j.compedu.2016.07.001

Kiili, C., Laurinen, L., & Marttunen, M. (2008). Students evaluating Internet sources: From versatile evaluators to uncritical readers. *Journal of Educational Computing Research*, *39*(1), 75-95.

Kim, H., Park, S. Y., & Bozeman, I. (2011). Online health information search and evaluation: observations and semi-structured interviews with college students and maternal health experts. *Health Information & Libraries Journal*, *28*(3), 188-199.

Kuiper, E., Volman, M., & Terwel, J. (2005). The Web as an information resource in K–12 education: Strategies for supporting students in searching and processing information. *Review of Educational Research*, *75*(3), 285-328.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the Craft of Qualitative Research Interviewing* (2nd ed.). Thousand Oaks, CA: Sage.

Lazer, D., Baum, M. Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017, May). *Combating Fake News: An Agenda for Research and Action*. Retrieved from: https://shorensteincenter.org/wp-content/uploads/2017/05/Combating-Fake-News-Agenda-for-Research-1.pdf

Lenhardt, A. (2015, April). *Teens, Social Media and Technology Overview 2015* Pew Research Center, Retrieved from http://www.pewinternet.org/files/2015/04/PI_TeensandTech_Update2015_0409151.pdf

Livingstone, S., Haddon, L., Görzig, A., & Ólafsson, K. (2010). *Risks and safety on the internet: the perspective of European children: key findings from the EU Kids Online survey of 9-16 year olds and their parents in 25 countries*. Retrieved from http://eprints.lse.ac.uk/53058/1/__lse.ac.uk_storage_LIBRARY_Secondary_libfile_shared_repository_Content_EU%20Kids%20Online_EU_Kids_Online_Report_Risks_and_safety_for_children_on_the_internet_2010.pdf

Macedo-Rouet, M., Braasch, J. L., 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.

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.

Mason, L., Scrimin, S., Tornatora, M. C., Suitner, C., & Moè, A. (2018). Internet source evaluation: The role of implicit associations and psychophysiological self-regulation. *Computers & Education*, *119*, 59-75.

McCormick, M. K., & Segal, P. H. (2016). How to make science texts more accessible. *The Science Teacher*, *83*(4), 41.

McCrudden, M.T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review*, *19*, 113-139.

McCrudden, M. T., Stenseth, T., Bråten, I., & Strømsø, H. I. (2016). The effects of topic familiarity, author expertise, and content relevance on Norwegian students' document selection: A mixed methods study. *Journal of Educational Psychology*, *108*(2), 147-162.

McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). Bigger Than Fake News. *American Educator*, *41*(3), 4-9.

Menchen-Trevino, E. & Hargittai, E. (2011). Young Adults' Credibility Assessment of Wikipedia. *Information, Communication and Society*. *14*(1), 24-51.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the Association for Information Science and Technology*, *58*(13), 2078-2091.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*, 210-220.

Metzger, M. J., Flanagin, A. J., Markov, A., Grossman, R., & Bulger, M. (2015). Believing the unbelievable: understanding young people's information literacy beliefs and practices in the United States. *Journal of Children and Media*, *9*(3), 325-348.

Murphy, P. K., Long, J. F., Holleran, T. A., & Esterly, E. (2003). Persuasion online or on paper: A new take on an old issue. *Learning and Instruction*, *13*, 511–532.

Organisation for Economic Co-operation and Development (2016), « PISA 2015 Reading Framework », In: OECD *PISA 2015 Assessment and Analytical Framework: Science,*

Reading, Mathematic and Financial Literacy, Éditions OCDE, Paris, http://dx.doi.org/10.1787/9789264255425-4-en.

Palfrey, J. (2016). Reframing Privacy and Youth Media Practices. In C. Greenhow, J. Sonnevend and C. Agur (Eds.) *Education and Social Media: Toward a Digital Future* (pp. 113-130), Cambridge: MIT Press.

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.

Pérez, A., Potocki, A., Stadtler, M., Macedo-Rouet, M., Paul, J., Salmerón, L., & Rouet, J. F. (2018). Fostering teenagers' assessment of information reliability: Effects of a classroom intervention focused on critical source dimensions. *Learning and Instruction*, *58*, 53-64.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication*, 12(3), 801-823.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the Association for Information Science and Technology*, *53*(2), 145-161.

Rouet, J.-F., & Britt, M.A. (2014). Learning from Multiple Documents. In Mayer, R.E. (Ed.) *Cambridge Handbook of Multimedia Learning* (2nd Edition, pp. 813-841). Cambridge, MA: Cambridge University Press.

Rouet, J.-F., Ros, C., Goumi, A., Macedo-Rouet, A., & Dinet, J. (2011). The influence of surface and deep cues on grade school students' assessment of relevance in Web menus. *Learning and Instruction*, *21*, 205-219.

Salmerón, L., García, A., & Vidal-Abarca, E. (2018). The development of adolescents' comprehension-based Internet reading skills. *Learning and Individual Differences*, *61*, 31-39.

Sbaffi, L., & Rowley, J. (2017). Trust and Credibility in Web-Based Health Information: A Review and Agenda for Future Research. *Journal of Medical Internet Research*, *19*(6), e218. http://doi.org/10.2196/jmir.7579

Smith, A. G. (1997). Testing the surf: criteria for evaluating Internet information resources. *Public Access-Computer Systems Review*, *8*(3), 5-23.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in read-23ing comprehension*. Rand Corporation.

Stanford History Education Group (2016). *Evaluating information: the cornerstone of civic online reasoning*. Retrieved from https://sheg.stanford.edu/upload/V3LessonPlans/Executive%20Summary%2011.21.16.pdf

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

Stadtler, M., & Bromme, R. (2007). Dealing with multiple documents on the WWW: The role of metacognition in the formation of documents models. *International Journal of Computer Supported Collaborative Learning*, *2*, 191–210.

Stadtler, M. & Bromme, R. (2008). Effects of the metacognitive computer-tool met.a.ware on the web search of laypersons. *Computers in Human Behavior*, *24*, 716-737. doi:10.1016/j.chb.2007.01.023

Sundin, O., & Francke, H. (2009). In search of credibility: pupils' information practices in learning environments. *Information Research: An International Electronic Journal*, *14*(4), Retrieved from http://InformationR.net/ir/14-4/paper418.html

Tabatabai, D., & Shore, B. M. (2005). How experts and novices search the Web. *Library & information science research*, *27*(2), 222-248.

van den Broek, P., Bohn-Gettler, C.M., Kendeou, P., Carlson, S., & White, M.J. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension. In M.T. McCrudden, J.P. Magliano, & G. Schraw (Eds.), Text *Relevance and Learning from Text* (pp. 123-140). Greenwich, CT: Information Age Publishing.

Van Der Heide, B., & Lim, Y. S. (2016). On the conditional cueing of credibility heuristics: The case of online influence. *Communication Research*, *43*(5), 672-693.

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. (2008). Information-problem solving: A review of problems students encounter and instructional solutions. *Computers in Human Behavior*, *24*(3), 623-648.

Watson, C. (2014). An exploratory study of secondary students' judgments of the relevance and reliability of information. *Journal of the Association for Information Science and Technology*, *65*(7), 1385-1408.

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*(4), 1060-1106.

Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Westport, CT: Greenwood Press.

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

Wineburg, S., & McGrew, S. (2017). *Lateral Reading: Reading Less and Learning More When Evaluating Digital Information*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3048994

Wineburg, S., & Reisman, A. (2015). Disciplinary literacy in history. *Journal of Adolescent & Adult Literacy*, *58*(8), 636-639.

Wopereis, I. G., & van Merriënboer, J. J. (2011). Evaluating text-based information on the World Wide Web. *Learning and Instruction*, *21*(2), 232-237.

World Health Organization (2017). *Adolescent Health*. Retrieved from http://www.who.int/topics/adolescent_health/en/

Zhang, M. (2013a). Prompts-based Scaffolding for Online Inquiry: Design Intentions and Classroom Realities. *Educational Technology & Society*, *16*(3), 140-151.

Zhang, M. (2013b). Supporting middle school students' online reading of scientific resources: moving beyond cursory, fragmented, and opportunistic reading. *Journal of Computer Assisted Learning*, *29*(2), 138-152.

Table 1.

Characteristics of the texts used in Experiments 1 and 2. In Experiment 1, only the topic of climate change was used (texts 1a, 2a, 3a, 4a). In Experiment 2, both texts were used.

Text title	Topic-match	Readability (Fog index)	Author competence	Currency
Text 1 ("Irrelevant text") 1a. The climate zones 1b. Fast-food Pannic	Poor ✗	Good ✓	Good ✓	Good ✓
Text 2 ("Difficult text") 2a. Climate change: the CO2 in question 2b. Cities confronted to obesity	Good ✓	Poor ✗	Good ✓	Good ✓
Text 3 ("Unreliable text") 3a. Climate change: who is responsible? 3b. Does fast-food make you gain weight?	Good ✓	Good ✓	Poor ✗	Good ✓
Text 4 ("Outdated text") 4a. The causes of climate change 4b. Fast-food: accountable but not guilty?	Good ✓	Good ✓	Good ✓	Poor ✗

Table 2a.

Means scores attributed to the "poor" and "good" documents with respect to each of the four criteria by teachers and librarians in the pilot study (scale 0-7, n=57).

	Fast food & obesity		Climate change	
	Poor	Good (avg)	Poor	Good (avg)
1. Is this document really about the topic of the assignment? (Topic-match)	1.1 (1.9)	5.1 (1.3)	0.4 (0.7)	5.7 (0.7)
2. Is this document easy to understand for a student? (Readability)	3.8 (1.5)	6.1 (1.1)	4.0 (2.7)	6.3 (0.5)
3. Does the author of the document have a lot of knowledge on the topic? (Author competence)	1.4 (1.3)	4.8 (0.7)	3.3 (2.5)	5.1 (1.1)
4. Does this document present recent information about the topic? (Currency)	3.6 (3.0)	5.0 (1.6)	0.8 (1.4)	3.8 (1.5)

Table 2b.

Summary of the two-way ANOVA analysis of results from the pilot study, per criterion, with topic and type of document (flawed, control) as within-participant variables and ratings as an average measure.

	Topic-match			Readability			Author competence			Currency		
	***F*(1,14)**	***p***	**η^2**	***F*(1,14)**	***p***	**η^2**	***F*(1,14)**	***p***	**η^2**	***F*(1,14)**	***p***	**η^2**
Type of document	145.65	.001	.912	21.964	.001	.611	25.503	.001	.646	6.916	.020	.331
Topic	0.016	.902	.001	0.117	.737	.008	0.242	0.63	.017	14.273	.002	.505
Type of document x Topic	3.196	.095	.186	0.002	.967	.000	1.399	.257	.091	1.086	.315	.072

Table 3.

Coding scheme for justifications for goodness/usefulness of each text (Experiments 1 and 2).

Category (% agreement)	Description	Example
1) Topic-match (91%)	The topic of the text matches / does not match the topic of the assignment (i.e. "causes of climate change"). Includes justifications that describe the content of the text and/or its relevance for the task.	"Because it is about climate zones, not about the topic we are interested in" (S19) "It explains why the Earth 'warms up'" (S13)
2) Readability (97%)	The text is considered easy / difficult to understand, well / not-well written, accessible / not accessible, either for the student himself/herself or for other people.	"This text explains well the climate change, it's easy to understand" (S6) "There are words that my classmates will not understand" (S4)
3) Author competence (94%)	The author of the text and/or its qualifications are cited.	"She is an environmental engineer" (S55)
4) Currency (100%)	The publication date and/or the currency of the text are cited.	"Too old" (S22)
5) Amount and precision of information (97%)	The text provides a certain amount of information, with a certain degree of precision about the topic: a lot / not a lot, enough / not enough, complete / incomplete, precise / vague, accurate / inaccurate…	"There is a lot of information, it can be useful for a presentation" (S24) "Although it's vague we can learn something from it" (S52)
6) Satisfaction (94%)	The student expresses a general satisfaction with the text without further explanation. The text is considered to be good / bad, helpful / not helpful, liked / not liked…	"I pretty much liked it" (S17) "It doesn't help to prepare for the presentation" (S54)
7) Opinion (100%)	The student agrees / does not agree with what is written in the text or with the author.	"I agree because I also think that humans are responsible for global warming" (S45) "I gave a 2 because I don't totally agree with [author name]" (S48)
8) Other (97%)	Vague responses, attributions of truthfulness (e.g. "it says the truth", S21) and "don't know" answers.	"It has some downsides as well as some benefits" (S38) "It says the truth" (S21) "I don't know" (S48)

Table 4.

Evaluation criteria cited in students' justifications by type of text and overall (Experiment 1, % of total coded segments, n=271).

	Text 1 (irrelevant)	Text 2 (difficult)	Text 3 (unreliable)	Text 4 (outdated)	Overall
Topic-match	43.3	30.9	31.9	39.4	36.2
Readability	23.3	36.8	26.4	22.5	27.3
Amount and precision of information	11.1	17.6	18.3	10.0	14.4
Satisfaction	15.0	4.4	12.5	9.9	10.3
Opinion	0.0	2.9	5.6	1.4	2.6
Author competence	1.7	1.5	5.6	0.0	2.2
Currency	0.0	0.0	0.0	2.8	0.7
Other	6.7	5.9	6.9	5.6	6.3

Table 5.

Evaluation criteria cited in students' oral justifications by type of text and overall (Experiment 2, % of total codes, n=244)

	Text 1 (irrelevant)	Text 2 (difficult)	Text 3 (unreliable)	Text 4 (outdated)	Overall
Topic-match	59.2	21.9	19.7	25.0	29.6
Readability	6.1	31.3	13.6	9.4	15.6
Satisfaction	16.3	12.5	15.2	18.8	15.6
Amount and precision of information	4.1	15.6	12.1	15.6	12.3
Author competence	2.0	10.9	21.2	7.8	11.1
Currency	6.1	4.7	4.5	7.8	5.8
Opinion	0.0	0.0	0.0	3.1	0.8
Other	8.0	3.1	13.6	12.5	9.4

Table 6.

Mean scores attributed to the poor and good documents per criterion and school level (Experiment 2, collapsed across domains).

	Poor	Good (avg)
Topic-match		
Middle schoolers	1.1 (1.8)	5.2 (1.2)
High schoolers	1.0 (1.5)	5.3 (0.8)
Readability		
Middle schoolers	2.5 (2.2)	6.0 (0.8)
High schoolers	3.7 (2.2)	6.2 (1.0)
Author competence		
Middle schoolers	4.8 (1.7)	4.9 (1.3)
High schoolers	3.0 (1.6)	5.6 (0.8)
Currency		
Middle schoolers	3.3 (2.2)	5.0 (1.3)
High schoolers	2.5 (2.0)	5.7 (1.1)

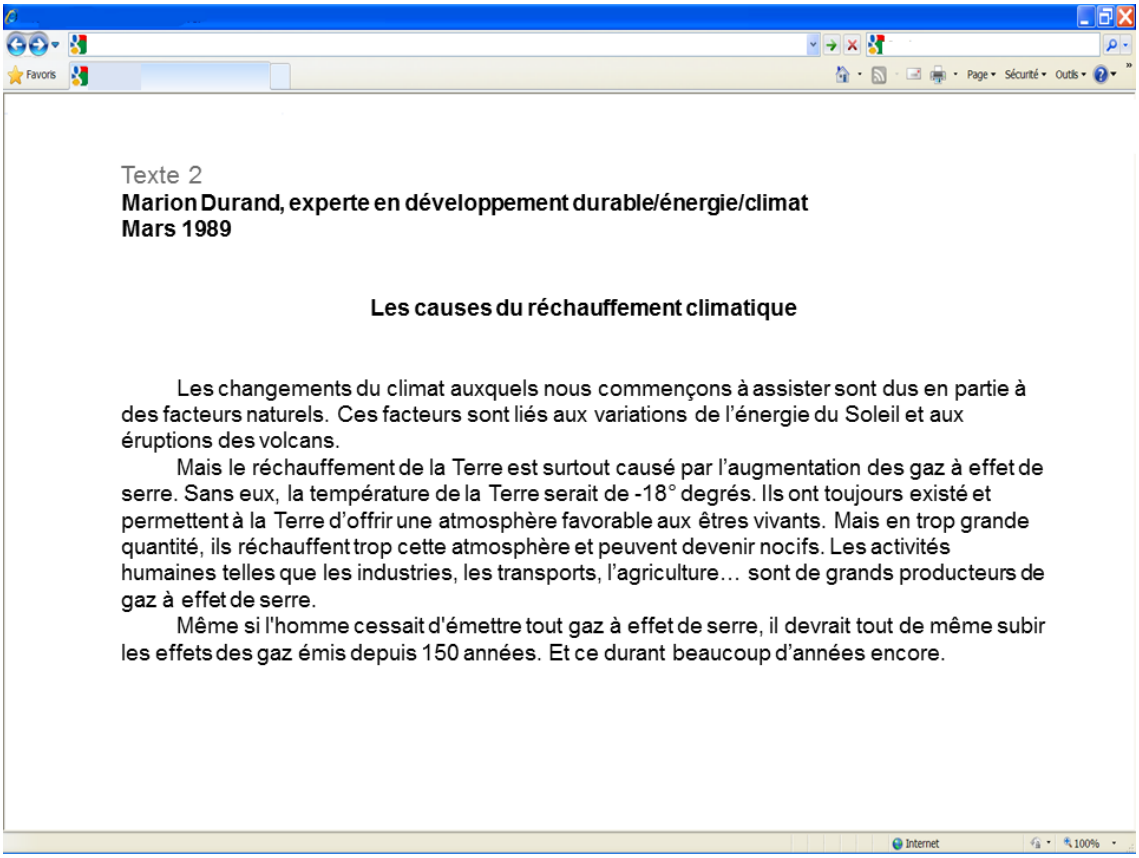

Figure 1. One of the texts presented to participants in the pilot study. In this example, the text is presented as published in 1989, that is, outdated in comparison to the other three texts that were dated of 2014.

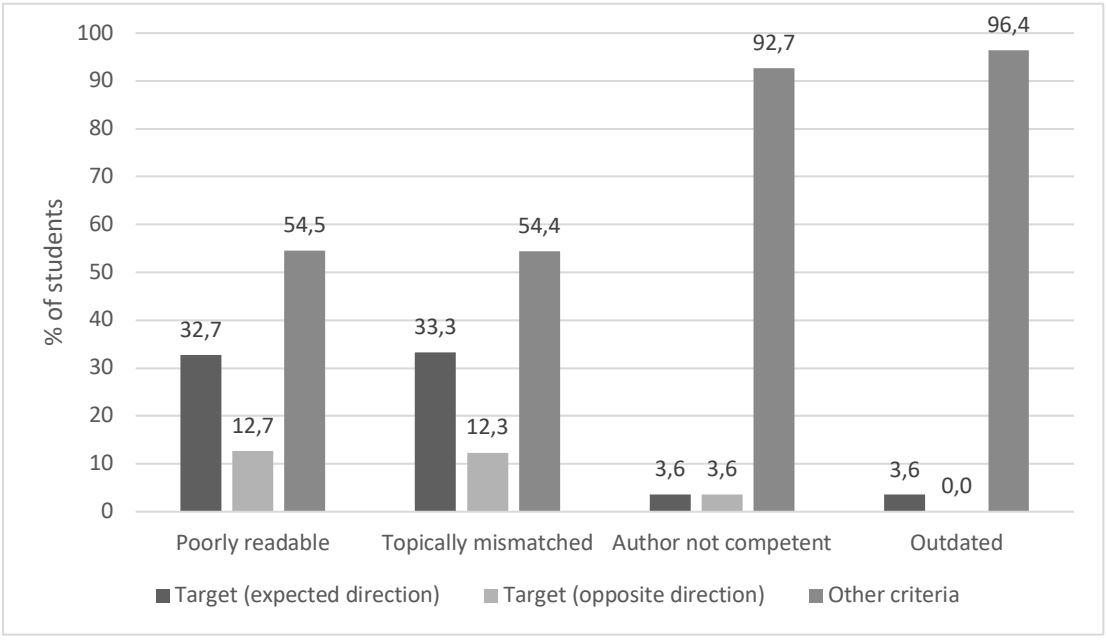

Figure 2. Percentage of students who cited the target criterion per type of text in Experiment 1

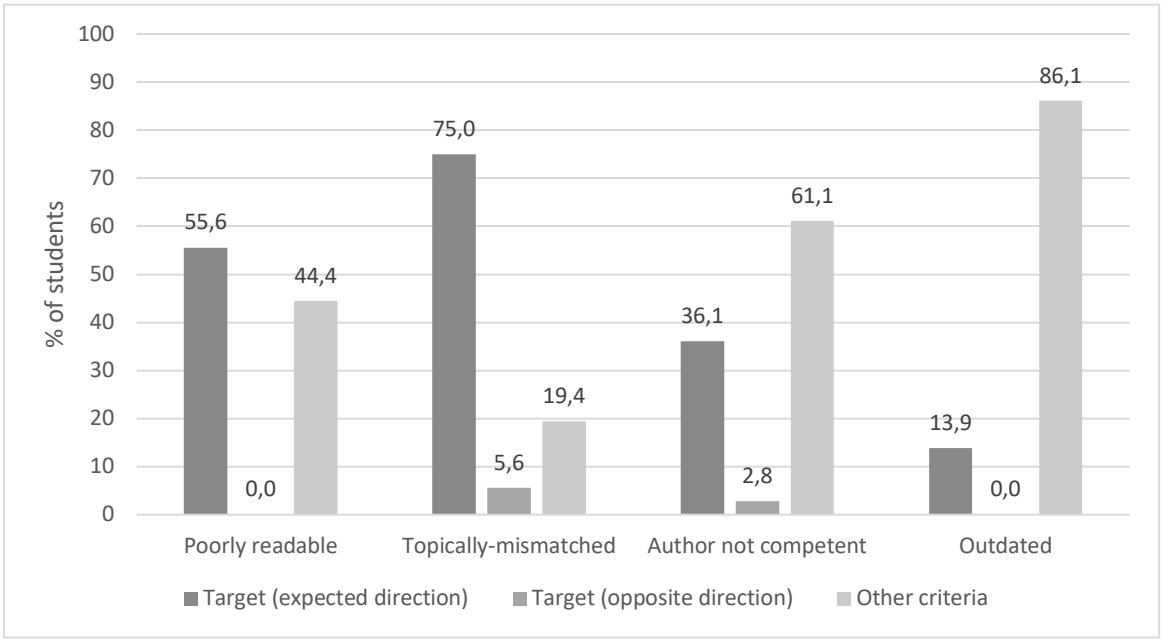

Figure 3. Percentage of students who cited the target criteria per type of text in Experiment 2

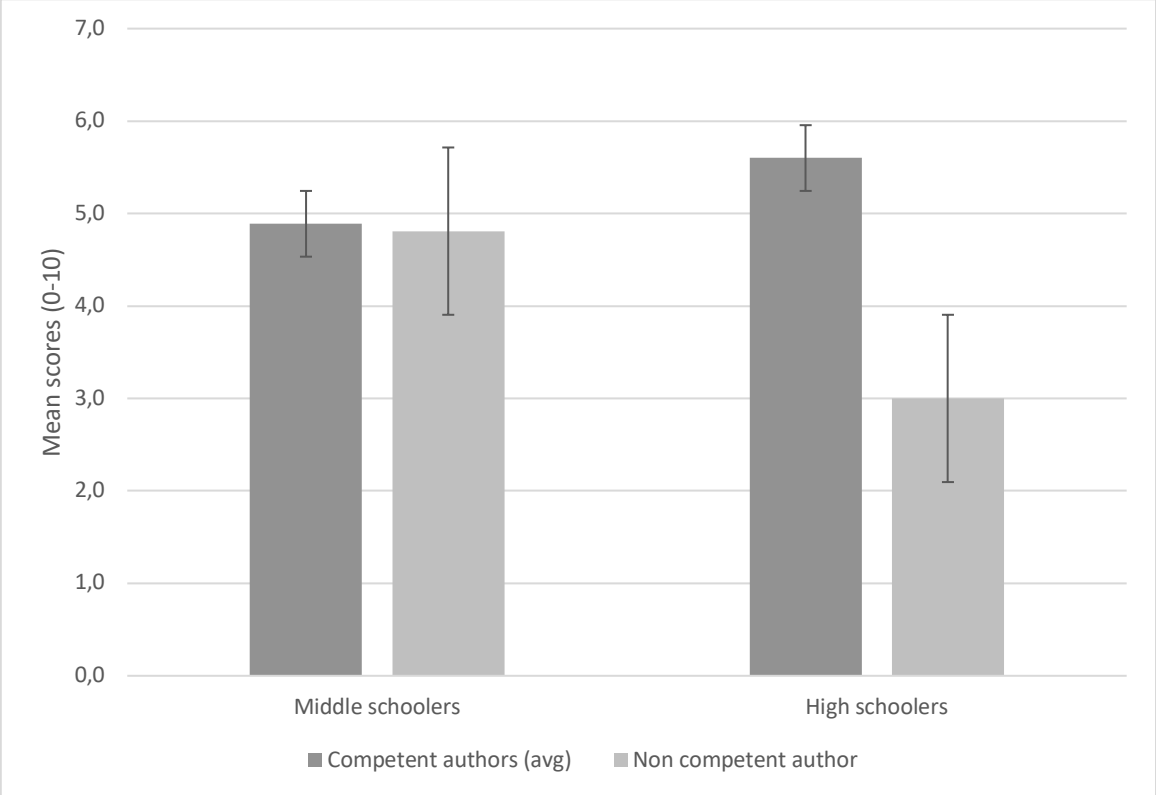

Figure 4. Mean scores attributed to the non-competent and the competent authors by middle schoolers and high schoolers in Experiment 2

Author statements:

Mônica Macedo-Rouet is an associate professor of Educational Sciences at the University of Paris 8 Vincennes-Saint-Denis, France, e-mail mgoncalves-macedo@univ-paris8.fr (corresponding author)

Anna Potocki is an associate professor of Psychology at the University of Poitiers, France, e-mail anna-potocki@univ-poitiers.fr

Lisa Scharrer is a researcher in the Institute of Educational Sciences at Ruhr University Bochum, Germany, e-mail lisa.scharrer@rub.de

Christine Ros is a research engineer at CNRS-University of Poitiers, France, e-mail christine.ros@univ-poitiers.fr

Marc Stadtler is a professor of Psychology in the Institute of Educational Sciences at Ruhr University Bochum, Germany, e-mail marc.stadtler@rub.de

Ladislao Salmerón is an associate professor of Psychology at the University of Valencia, Spain, e-mail ladislao.salmeron@valencia.edu

Jean-François Rouet is a research director at CNRS-University of Poitiers, France, e-mail jean-francois.rouet@univ-poitiers.fr