Received: 28 June 2022 | Revised: 9 November 2022 | Accepted: 1 January 2023

DOI: 10.1111/jcal.12779

ARTICLE

Journal of Computer Assisted Learning WILEY

# Examining the structure of credibility evaluation when sixth graders read online texts

Carita Kiili[1] | Eija Räikkönen[2] | Ivar Bråten[3] | Helge I. Strømsø[3] | Michelle Schira Hagerman[4]

[1]Faculty of Education and Culture, Tampere University, Tampere, Finland

[2]Faculty of Education and Psychology, University of Jyväskylä, Jyväskylä, Finland

[3]Department of Education, University of Oslo, Oslo, Norway

[4]Faculty of Education, University of Ottawa, Ottawa, Ontario, Canada

**Correspondence**
Carita Kiili, Faculty of Education and Culture, Tampere University, P.O. Box 700, 33014 Tampere, Finland.
Email: carita.kiili@tuni.fi

**Funding information**
Academy of Finland, Grant/Award Number: 324524

## Abstract

**Background:** Previous research indicates that students lack sufficient online credibility evaluation skills. However, the results are fragmented and difficult to compare as they are based on different types of measures and indicators. Consequently, there is no clear understanding of the structure of credibility evaluation.

**Objectives:** The present study sought to establish the structure of credibility evaluation of online texts among 265 sixth graders.

**Methods:** Students' credibility evaluation skills were measured with a task in which they read four online texts, two more credible (a popular science text and a newspaper article) and two less credible (a layperson's blog text and a commercial text). Students read one text at a time and evaluated the author's expertise, the author's benevolence and the quality of the evidence before ranking the texts according to credibility. Four competing measurement models of students' credibility evaluations were assessed.

**Results:** The model termed the Genre-based Confirming-Questioning Model reflected the structure of credibility evaluation best. The results suggest that credibility evaluation reflects the source texts and requires two latent skills: confirming the more credible texts and questioning the less credible texts. These latent skills of credibility evaluation were positively associated with students' abilities to rank the texts according to credibility.

**Implications:** The study revealed that the structure of credibility evaluation might be more complex than previously conceptualized. Consequently, students would benefit from activities that ask them to carefully analyse different credibility aspects of more and less credible texts, as well as the connections between these aspects.

KEYWORDS
adolescents, credibility evaluation, information literacy, internet, sourcing

## 1 | INTRODUCTION

As we write this paper, the world is still in the middle of the COVID-19-pandemic. In Finland, the context of this study, vaccinations for all citizens between 12 and 15 years of age are about to begin. When medically warranted, adolescents may decide for themselves whether to take a COVID-19-vaccine. Consider the context in which adolescents are supposed to make such a decision.

Coinciding with the pandemic, there is an 'infodemic' that refers to an overabundance of accurate and inaccurate information (Tangcharoensathien et al., 2020). The infodemic spreads in a manner similar to an epidemic through physical and digital interactions among people, making it hard to know what to trust. The results of a large survey conducted with over 1000 German adults illustrate the challenge: almost half of the participants reported having difficulties judging whether they could trust media information on COVID-19 (Okan et al., 2020). Given that adults struggle, the importance of examining 12–13-year-old students' credibility evaluation of health-related online information, as we did in the current study, is of fundamental importance. Because previous research has already documented what adolescents are capable (and not capable) of doing in this regard (e.g., Coiro et al., 2015; McGrew et al., 2018), we aimed to move the field forward by providing new understandings of the structure of credibility evaluation of online information among early adolescent readers. Establishing the structure of credibility evaluation in online contexts would offer future research a conceptual framework for designing measures, investigating the development of credibility evaluation skills and comparing results across cultural contexts.

To reach these aims, we designed an online inquiry task concerning common myths about sugar in which sixth graders were asked to evaluate four texts that represented different text genres and included either accurate or inaccurate information from different perspectives. Using this design, we compared four different credibility evaluation structures.

### 1.1 | Credibility evaluation as an online inquiry skill

In the post-truth era characterized by the spread of mis- and disinformation, the ability to evaluate the credibility of information is particularly important for successful use of online information (Barzilai & Chinn, 2020). Accordingly, evaluation of information is considered one of the five core skills of online inquiry, the others being formulating a driving question and locating, synthesizing and communicating information (Leu et al., 2019). Evaluation of information occurs during different phases of online inquiry (Kiili et al., 2021) and concerns the relevance and credibility of online information (Hahnel et al., 2020; McCrudden, 2018).

The theoretical model of online inquiry by Leu et al. (2004, 2019) represents the evaluation of information as a single construct. A study by Hahnel et al. (2020), which examined university students' ($N = 152$) evaluation of online information, supported a unidimensional structure of evaluation of information. In that study, students completed a computer-based test comprising eight tasks, which included three to five links to websites. Source and content features concerning the relevance and credibility of information on the websites were consistent. For example, claims presented by an expert were accurate. Students were asked to select the website that contained the most useful and trustworthy information. There were four types of items that differed in terms of the processing required to select a correct website. In the easiest item type, the correct selection could be determined by relying on predictive judgement based on the information in the links. The other three item types also required inspection of the website.

A study by Kiili et al. (2018) evaluated the structure of online inquiry when about 400 adolescents, 12–13 years of age, completed an online inquiry task in a closed web-based environment. The study confirmed the hypothesised four component skills of online inquiry: locating, evaluating, synthesizing and communicating information. However, a model with six factors fitted the data even better than the model with four factors. In the six-factor model, both evaluating and synthesizing information were separated into two different factors. The two evaluating factors were labelled the ability to confirm the credibility (of more credible text) and the ability to question the credibility (of less credible text). The two synthesizing factors were identifying main ideas from single texts and synthesizing information across multiple texts. However, in the study by Kiili et al. (2018), students evaluated only two online texts: one more credible text (a text written by an expert) and one less credible text (a commercial text). In the present study, we asked students to evaluate four texts: two more credible online texts and two less credible online texts. In doing so, we sought to determine whether these two constructs—confirming and questioning credibility—would also appear when students evaluate more than two texts.

Notably, Kiili et al. (2018) findings with younger students have been supported by several other studies. For example, Potocki et al. (2020) examined 5th, 7th, and 9th graders' and undergraduates' ($N = 245$) abilities to differentiate less competent authors from more competent authors (having expertise on the text topic) and less benevolent authors from more benevolent authors (having a good will). Students were asked to evaluate the competence and benevolence of authors of short, printed texts on a scale ranging from 0 to 10. Fifth graders had difficulties differentiating between more and less competent authors. Surprisingly, all age groups struggled with differentiating between more and less benevolent authors. Further, Pieschl and Sivyer (2021) examined how 7th, 9th, and 11th graders ($N = 218$) evaluated more and less credible blog posts, finding that 7th graders were not able to differentiate between the credibility of the blog posts. In contrast, older students were better able to do so. These results suggest that younger students may have particular or unique difficulties in questioning less credible online texts.

Readers can employ various strategies when evaluating more or less credible online texts. Some researchers have classified these strategies as first- and second-hand evaluation strategies (Barzilai et al., 2020; Stadtler & Bromme, 2014). Readers can use first-hand evaluation strategies to judge the validity of knowledge claims (i.e., text content) presented in a text and second-hand evaluation strategies to judge the trustworthiness of the source of the information (i.e., the author). Barzilai et al. (2020) also emphasized that first- and second-hand judgements are reciprocal, influencing each other.

First-hand evaluation strategies include validating the content against one's prior knowledge or beliefs, corroborating content using information from other resources, and evaluating the quality of arguments (Barzilai et al., 2020). If readers have prior knowledge or beliefs about the text topic, they may routinely judge the validity of

knowledge claims (Richter & Maier, 2017). Readers tend to evaluate content consistent with their prior beliefs higher than content inconsistent with their beliefs (e.g., Abendroth & Richter, 2021). When evaluating the credibility of the content, readers can also evaluate the quality of the arguments presented in the text by assessing the coherence of the arguments (Stadtler & Bromme, 2014) or the strength of the evidence (Nussbaum, 2020). For example, readers can consider whether the author relies on research-based evidence, expert statements, or anecdotal evidence, such as personal experiences (Zarefsky, 2019).

Second-hand evaluation strategies focus on assessing the expertise and benevolence of the author (Barzilai et al., 2020; Stadtler & Bromme, 2014). Expertise refers to the author's competence and experience in sharing accurate information about the topic (Stadtler & Bromme, 2014; Thomm & Bromme, 2016). Readers can use different cues to infer the level of the author's expertise by paying attention to credentials, experience, and affiliation (Bråten et al., 2018; Hendriks et al., 2015). Authors' benevolence refers to their intention to act in the interest of the readers without pursuing any personal aims or benefits (Hendriks et al., 2016; Thomm & Bromme, 2016). For example, readers can question authors' benevolence by noticing their commercial or political interests. The benevolence of authors with persuasive intentions may also be questioned because they may be assumed to provide one-sided information that serves their personal goals.

In this study, we prompted students to evaluate the credibility of the source by asking them to evaluate the expertise and benevolence of the authors of the online texts. To evaluate the credibility of the content, we prompted students to evaluate the quality of the evidence that the authors used to support their claims.

### 1.2 | Credibility evaluation of online texts representing different genres

In addition to the evaluation of content and source features of the text, readers may use their knowledge about genres when evaluating online texts (Flanagin & Metzger, 2007; Forzani, 2020; Sundin & Francke, 2009). Genres are socially situated practices that reflect certain formal text features, social norms and rhetorical purposes of texts (Duke & Roberts, 2010; Purcell-Gates et al., 2007). According to Berkenkotter and Huckin (1995), readers' genre knowledge includes knowledge about forms, conventions and contents of texts that are appropriate in a particular situation. Importantly, this knowledge evolves when readers participate in various communicative activities.

Genres are not stable but change over time, echoing historical and contextual changes that are reflected in readers' generic expectations (Fisher, 2019). The advent of the internet is an example of a historical turn that has changed communicative practices (Leu et al., 2019; Tierney & Pearson, 2021). Broadly speaking, internet genres both create and reflect the networked essence of the internet. As such, internet genres change over time and come to include new forms of collaboration, new forms of multimodal expression and new organizational structures (Bauman, 1999). More specifically, the rise of the internet has evoked specific genres, such as online encyclopaedias, discussion forums and blogs (Crowston, 2010). However, some genres may include various kinds of texts with various conventions and purposes. Blogs, for example, can be journal blogs, travel blogs, or science blogs (Crowston, 2010). Additionally, the boundaries of genres on the internet are sometimes blurred (Flanagin & Metzger, 2007), which may complicate the identification of genres (Leeder, 2016).

Several studies have manipulated the text genre or document type when examining readers' perceptions and reasoning about the credibility of multiple texts (Bråten et al., 2015; Flanagin & Metzger, 2007; List et al., 2017). List et al. (2017) presented university students ($N = 197$) with six texts about the Arab Spring in Egypt that students could use when composing a short essay on the topic. They were also asked to rate the credibility of the six texts, which represented the following genres: blog post, analysis essay, newspaper article, public opinion survey, Tweet and Wikipedia article. Of these texts, the newspaper article and the Wikipedia article were written in a neutral tone, whereas the others leaned towards one side of the issue. The newspaper article was accessed by the majority of the students and was thus the most accessed resource. In contrast, less than two-thirds of the students accessed the blog post. Text genre may have played a role in students' decisions to access (or not access) each resource. In the same study, students rated the newspaper article as the most credible, whereas the blog post was rated as least credible. Students may consider blogs as forums of opinion, maybe because their experiences of blogs mainly include the personal journal type of blogs (Sundin & Francke, 2009).

Similarly, Flanagin and Metzger (2007) studied the role of genre in the perceived credibility of online texts among adults ($N = 574$). In their study, four genres were used: a news site, an e-commerce site, a website of a special interest group and a personal webpage. The findings suggested that the online text genre, in addition to text attributes, may be an important factor when readers evaluate the credibility of online texts. Together, these findings from previous studies suggest that genre might signal particular intentions and evoke pre-existing expectations for credibility.

## 2 | PRESENT STUDY

Informed by previous research, the aim of this study is to examine the role of text credibility and text genre in students' credibility evaluation and, further, the relationship between their credibility evaluation and their ability to compare the credibility of online texts. To examine the structure of credibility evaluation when sixth graders read and evaluate the credibility of online texts, we asked them to read and evaluate four texts that varied in quality along three criteria: author's expertise, author's benevolence and the provided evidence. The four online texts represented different genres: a popular science text, a newspaper article, a layperson's blog text and a commercial text. In creating these texts, we manipulated the three credibility aspects—the author's expertise, the author's benevolence and the quality of the evidence—such that these aspects were present but different across the four

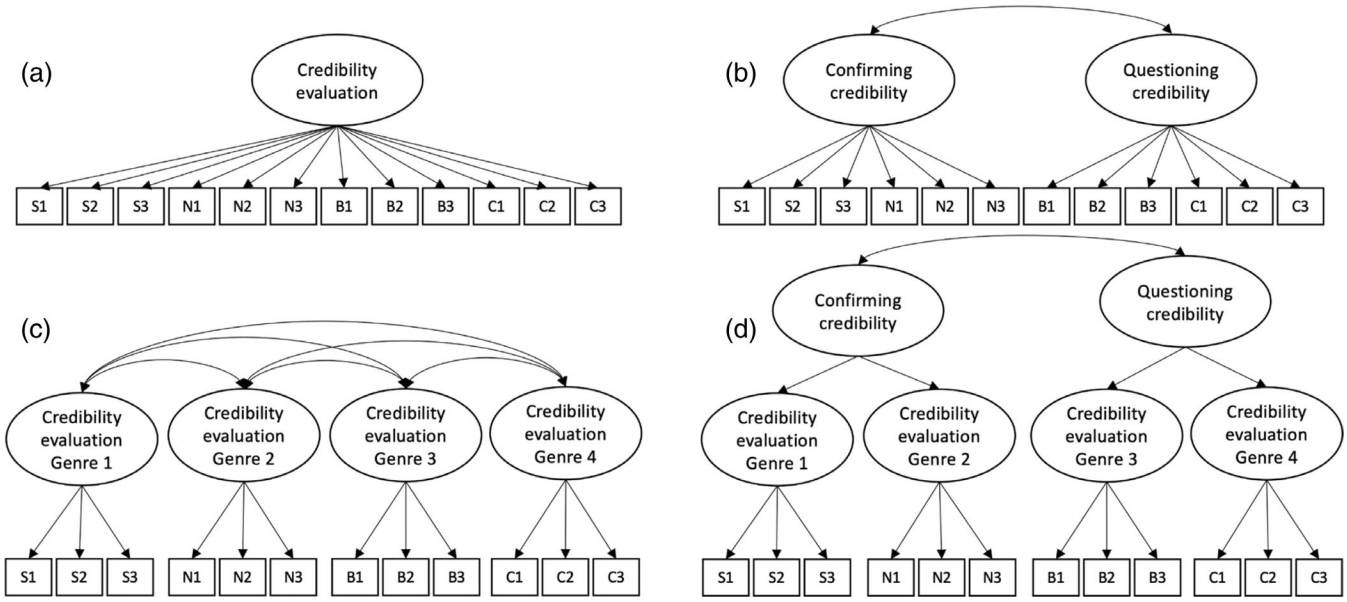


**FIGURE 1** The four competing measurement models: (a) General Credibility Evaluation Model, (b) Confirming-Questioning Model, (c) Genre-based Model, and (d) Genre-based Confirming-Questioning Model. S = Popular Science text, N = Newspaper article, B = Blog text, and C = Commercial text. 1 = evaluation of the author's expertise, 2 = evaluation of the author's benevolence, and 3 = evaluation of the quality of evidence

texts. More specifically, the popular science text and the newspaper article included indicators of expertise, benevolence and quality of evidence consistent with their genres and were written to be of higher quality than the layperson's blog text and the commercial text, as signalled by the indicators included in the latter texts. As designed, the popular science text and the newspaper article were more credible texts, whereas the layperson's blog text and the commercial text were less credible texts.

With this text design, the structure of students' credibility evaluation was examined by comparing four competing models that are described below and graphically depicted in Figure 1.

*The one-factor model* titled the *General Credibility Evaluation Model* (A) assumes that credibility evaluation reflects a general skill that a reader can employ across various kinds of texts regardless of their genre, level of credibility or other features. This assumption is based on the online inquiry model (Leu et al., 2019), which describes evaluation of online information as one online inquiry component skill and on empirical support for the unidimensional structure of evaluation found among university students (Hahnel et al., 2020).

*The two-factor model* titled the *Confirming-Questioning Model* (B) assumes that credibility evaluation requires two different skills: confirming the credibility of the more credible texts and questioning the credibility of the less credible texts. This presumption is grounded on previous findings suggesting that confirming the credibility requires overlapping, yet somewhat different skills than does questioning the credibility among early adolescent readers (Kiili et al., 2018).

*The four-factor model* titled the *Genre-based Evaluation Model* (C) assumes four latent factors reflecting that the online texts represent four different genres (popular science text, newspaper article, layperson's blog text, commercial text) where the source (expertise and intentions) and content features (the quality of the evidence) are consistent with the general expectations for the genres. This assumption is based on the idea that readers' genre knowledge, developed through diverse communicative activities, create expectations about text features and rhetorical purposes and, thus, about the credibility of texts (Flanagin & Metzger, 2007; List et al., 2017).

Based on empirical evidence derived from younger readers, we hypothesised that the two- and four-factor models would better capture the structure of students' credibility evaluation than the one-factor model. This is to say that credibility evaluation is not one general skill but constitutes a more fine-grained set of skills. We did not formulate any hypotheses about the superiority of the two- and four-factor models because previous studies have found empirical evidence for both.

In the case that the *Genre-based Evaluation Model* (C) would turn out to be the best of the Models A–C (cf. Rindskopf & Rose, 1988), we also planned to test a fourth model that could explain the relations between evaluation of online texts representing different genres. *This was a second-order factor model* titled the *Genre-based Confirming-Questioning Model* (D), which includes four first-order factors reflecting the genres of the online texts that map onto the two second-order factors reflecting the skills of questioning and confirming the credibility. This model was tested because there is preliminary empirical evidence supporting both genre-based evaluation and the need for two latent skills to, respectively, confirm the credibility of more credible texts and question the credibility of less credible texts.

In summary, our specific research questions were:

1. Which of the four hypothesised measurement models (A–D, see Figure 1) describes the structure of credibility evaluation of online texts among sixth graders best?
2. How is students' credibility evaluation associated with their ability to compare the credibility of online texts, specifically with their ability to rank the four online texts?

## 3 | METHODS

### 3.1 | Participants and context

Participants were recruited from 15 classrooms in five Finnish elementary schools. The participants were 265 sixth graders ($M = 12.45$ years; $SD = 0.32$). Of the participating students, 53.6% were girls, 44.5% were boys, and 1.9% selected the option other. According to students' guardians, most of the students' (90.9%) home language was Finnish and 7.6% of the students' homes were bilingual, with Finnish being one of the spoken languages. Only 1.5% spoke a language other than Finnish at home. Of students' guardians ($N = 482$), 58% reported having a degree either from a university or a university of applied sciences. The guardians with higher education degree were over-represented compared to the Finnish population (44% with higher education degrees) (Official Statistics of Finland, 2020).

In Finland, all schools follow the national curriculum (Finnish National Core Curriculum for Basic Education, 2014). The curriculum includes seven areas of transversal competencies that should be developed in every subject, multiliteracy being one of these competencies. To support students' multiliteracy, teachers guide students in interpreting, producing and evaluating various types of texts in different contexts. In addition, the learning objectives of language arts include guidance in information seeking, the use of diverse information resources and evaluation of the credibility of information.

### 3.2 | Instrument used

To assess students' credibility evaluation of online texts, we used the Critical Online Reading Research Environment (CORRE). It is a web-based environment where researchers can create critical online reading tasks (see Figure 2). In the task created for this study, students were asked to read, evaluate and rank the credibility of four online texts about the health effects of sugar. During task completion, students were instructed by a fact-checker, Max. The task instruction is presented in Appendix A. The task interface was split into two areas. The left-hand side of the interface presented the online texts and Max's instruction. On the right-hand side, students responded to the items (see Figure 2).

#### 3.2.1 | Text materials

The online texts, designed by the research team for this study, are described in Table 1. Two texts focused on sugar and hyperactivity in children, and two texts focused on sugar and its effects on memory. For both sub-topics, one text was more credible and one text was less credible. For the credible texts, the authors of the texts had expertise on the topic or were professional information seekers with sincere (benevolent) intentions. The authors also provided evidence in the form of research-based knowledge. In contrast, the authors of the less credible texts did not have expertise on the topic and they had either persuasive or commercial intentions. In presenting their arguments, these authors relied on evidence, such as their own observations, or on a customer survey that was not informed by research. The online texts represented four different genres that we expected students of this age to be familiar with: a blog text, a newspaper article, a popular science text and a commercial text.

The texts were written so that they were of approximately the same length (110–119 words) and followed the same organizational structure. Each text had a title formulated as a question, three paragraphs and one to three pictures. In all texts, the main claim appeared in the same position, as the last sentence in the first paragraph. To counterbalance possible effects of reading order, students were randomly assigned to two groups. Group 1 (coded as 0) first read the less credible text and then the more credible text (sub-topic 1), after which they read the more credible text followed by the less credible text (sub-topic 2). Group 2 read the texts within the sub-topics in reverse order (coded as 1).

#### 3.2.2 | Item types

Table 2 presents the item types included in the assessment. Students were asked to read one text at a time and respond to three types of items for each text: (1) identification items, (2) evaluation items, and (3) justification items. The items appeared on the screen one by one.

Identification items required students to locate the author, the main claim and the evidence supporting the main claim by choosing the correct answer from three options. After responding to the identification items, students were asked to check whether their response was aligned with Max's view. If the students' response was not correct, they were provided with the right answer. This was done to ensure that all students would evaluate the same authors and evidence.

Credibility evaluation items, using a six-point scale, measured students' abilities to evaluate the author and the evidence. Students evaluated the author from the perspectives of expertise and benevolence. In evidence evaluation, students were asked to consider how well the evidence was able to support the main claim. Students were asked to confirm their response, after which it was locked. Students' responses were scored from 0 to 2 points. For the items representing the more credible texts, students were credited with 2 points for ratings of 5 or

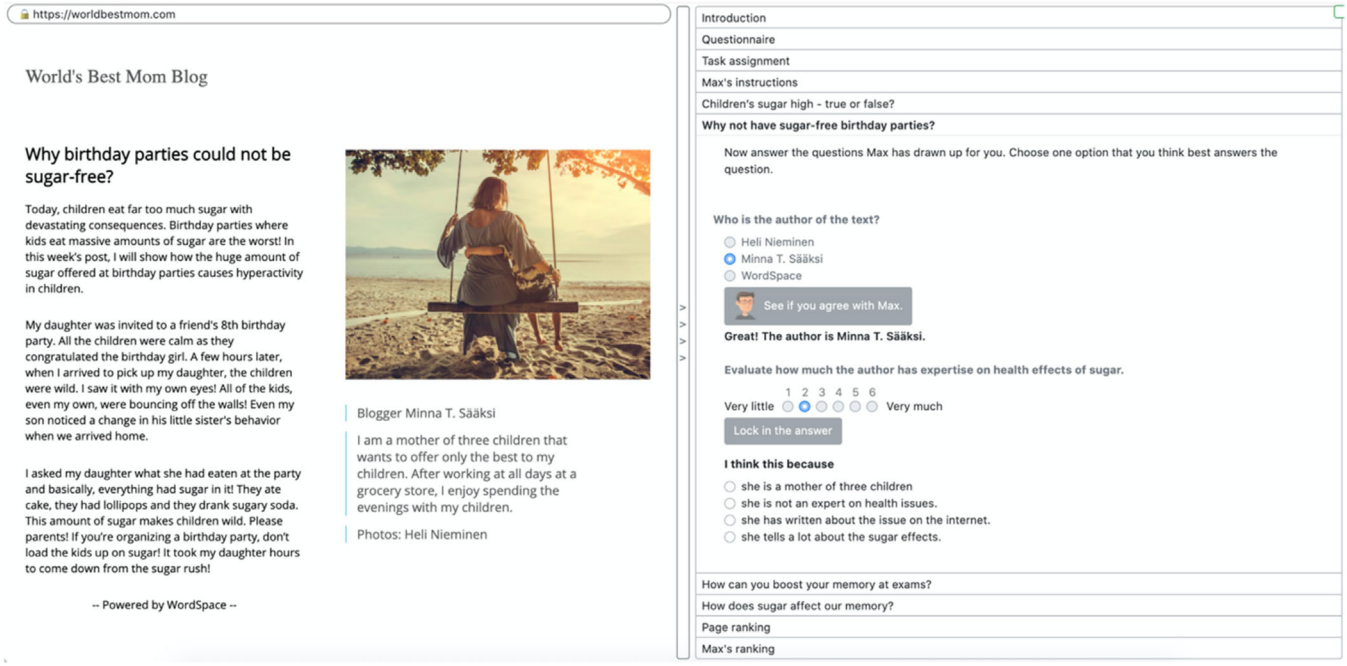


**FIGURE 2** Screenshot of the task interface. All item types with examples are presented in Table 2

**TABLE 1** The four online texts

| Sub-topic | Credibility | Text genre | Description |
|---|---|---|---|
| 1. Sugar and its effects on children's hyperactivity | More credible online text | Newspaper article | The article titled 'Children's sugar high – true or false?' is written by a journalist who is specialized in health and well-being. The journalist has interviewed a medical doctor who states that according to current knowledge, sugar does not cause hyperactivity in children. The doctor also shares the results of one study to back up her claim. |
| | Less credible online text | Layperson's blog text | The personal blog titled 'Why birthday parties could not be sugar free?' is written by a mother who works as a cashier. She claims that sugar causes hyperactivity in children. She uses her own observations of her daughter's wild behaviour after a sugary birthday party as evidence for her claim. She appeals to parents not to offer sugary treats at birthday parties. |
| 2. Sugar and its effects on memory | More credible online text | Popular science text | A researcher (Ph.D.) specialized in human memory has written a text titled 'How does sugar affect our memory?' The text appears on the research centre's website in a section on researchers' insights. The author states that sugar is essential for memory functions, but excessive use of sugar is harmful to memory. She provides research evidence for her two-sided claim. |
| | Less credible online text | Commercial text | A chief executive officer of a candy company is the author of the text titled 'How can you boost your memory at exams?' The text appears on the company website that includes some commercial slogans. The author claims that sugar intake improves memory, and presents the results of a customer survey as evidence. |

6, 1 point for ratings of 3 or 4 and 0 points for ratings of 1 or 2. The reverse scale was used for scoring the less credible texts. This scoring system was used to decrease the effect of some participants' tendency to favour or avoid answering in extremes (Greenleaf, 1992). Justification items asked students to justify their evaluations by selecting one of four justification options. These items were excluded from the analysis because the focus of this study was to understand the structure of credibility evaluation.

After reading all texts and responding to the related items, students ranked the four texts in terms of their credibility (ranking item; cf. List & Alexander, 2018; Mason et al., 2014). After confirming their rank-order, students were able to compare their order to Max's order,

**TABLE 2** The item types of the task

| Item type | Description | Example item |
|---|---|---|
| Identification item | Students were asked to identify<br>a. the author<br>b. the main claim<br>c. the evidence supporting the main claim<br>by choosing the right answer from three options. | Who has written the text?<br>a. Sisä-Suomen Sanomat (newspaper)<br>b. Market Valtasalo (a doctor who was interviewed)<br>c. Reijo Kangaskorpi (journalist) |
| Evaluation item | Students were asked to evaluate<br>1. author's expertise<br>2. author's benevolence<br>3. the quality of the evidence<br>4. overall credibility<br>of each text on a six-point scale. | Evaluate how much the author knows about sugar effects.<br>Knows only a little 12 3 4 5 6 Knows a lot |
| Justification item | Students were asked to justify their evaluations of<br>1. author's expertise<br>2. author's benevolence<br>3. the quality of the evidence<br>by selecting one of the four justification options. | I think so because<br>a. there is a lot of information in his text.<br>b. he is a doctor.<br>c. he is a journalist who writes about health.<br>d. he is unsure because he asks, 'True or false'. |
| Ranking item | Students were asked to rank the four online texts according to their credibility. | |

with the rankings presented side-by-side. Students' rankings were scored from 0 to 5. Students earned 5 if they ranked the popular science text as first and the newspaper article as the second most credible text. Students earned 4 if they put the aforementioned texts in the reverse order. Three points were credited if students ranked the popular science text as the first and one of the less credible texts as the second, and 2 points if they ranked the newspaper article the first and one of the less credible texts as the second. Students were credited with 1 point if they ranked one of the less credible texts as the first and one of the more credible texts as the second. Students did not receive any points if they ranked the two less credible texts as the first and the second.

#### 3.2.3 | Prior beliefs about the topic

Before reading and evaluating the four texts, students responded to two prior belief items: (1) Sugar causes hyperactivity in children (Sub-topic 1) and (2) Sugar improves memory (Sub-topic 2) on a 7-point scale (1 = totally disagree; 7 = totally agree). The items were embedded in the assessment and used as two separate control variables in the subsequent analyses.

### 3.3 | Procedure

The data were collected through Microsoft Teams during the 2020–2021 school year when the COVID-19-pandemic hindered us from entering classrooms. The researcher appeared on screen in the classrooms and introduced the task briefly. Then, students watched a short video (1 min 49 s) that introduced the features of the task environment and how to access the task. After watching the video, students logged on with their computers and accessed the web-based task by a given code. The researcher monitored that all students were able to access the environment by using the administrative version of the environment. After ensuring that all students had started on the task, the researcher shut her camera and microphone. When all students had completed the task, the researcher opened her camera and microphone and thanked the students for their work. When students were working on the task, the teacher and researcher communicated using the chat if needed. The research was conducted during one class period (45 min). The mean time spent on task was 20 min and 11 s (SD = 5:31).

For ethical reasons, we offered feedback to the students and teachers. About 1 or 2 weeks after the research session, the class received a feedback letter from Max that teachers read aloud. In addition, students received individual positive feedback (that referred to two issues on which students succeeded) from Max. We provided class- and student-level feedback to teachers about their students' evaluation performance.

### 3.4 | Data analyses

Confirmatory factor analysis (CFA) was employed to assess the competing measurement models of students' credibility evaluation by using the MPlus software (Version 8.0; Muthén & Muthén, 1998–2017). Specifications of the hypothesised models A–D are shown in Figure 1. Since the variables were ordinal and some of them were skewed (see Table 3), we employed the weighted least square (WLSMV) estimator, as it has been shown to provide less biased estimates for factor loadings with categorical variables (Li, 2016). There were no missing values in the data.

In the analyses, we controlled for the text order and students' prior beliefs about the topic. In Model A (General Credibility Evaluation Model), all three control variables (text-order, prior beliefs about

**TABLE 3** Descriptive statistics of students' credibility evaluation scores, prior beliefs, and ranking score ($N = 265$)

| Evaluation score (max. score = 3) | *M* | SD | Skewness | Kurtosis |
|---|---|---|---|---|
| Popular science text—expertise | 1.67 | 0.51 | −1.14 | 0.19 |
| Popular science text—benevolence | 1.78 | 0.44 | −1.73 | 1.97 |
| Popular science text—evidence | 1.72 | 0.47 | −1.19 | −0.01 |
| Newspaper article—expertise | 1.60 | 0.53 | −0.87 | −0.37 |
| Newspaper article—benevolence | 1.70 | 0.47 | −0.98 | −0.74 |
| Newspaper article—evidence | 1.70 | 0.48 | −1.19 | 0.14 |
| Laypers blog—expertise | 0.98 | 0.71 | 0.03 | −0.99 |
| Layperson's blog—benevolence | 0.52 | 0.59 | 0.66 | −0.52 |
| Layperson's blog—evidence | 0.88 | 0.69 | 0.16 | −0.90 |
| Commercial text—expertise | 0.94 | 0.66 | 0.06 | −0.72 |
| Commercial text—benevolence | 0.87 | 0.72 | 0.20 | −1.06 |
| Commercial text—evidence | 0.93 | 0.65 | 0.07 | −0.65 |
| Prior beliefs (max. 7) | | | | |
| Sugar causes hyperactivity | 5.18 | 1.59 | −0.79 | −0.01 |
| Sugar improves memory | 2.79 | 1.38 | 0.69 | 0.56 |
| Ranking score (max. 5) | 3.34 | 1.55 | −0.50 | −1.00 |

sugar effects on hyperactivity and prior beliefs about sugar effects on memory) were set to explain the credibility evaluation factor. In Model B (Confirming-Questioning Model), the control variables were set to explain the Confirming and Questioning factors. In Model C (Genre-based Model) and Model D (Genre-based Confirming-Questioning Model), text-order was set to explain all first-order factors (i.e., factors based on online texts), whereas prior beliefs about sugar effects on hyperactivity were set to explain the online texts about sugar and hyperactivity (blog text, newspaper article) and prior beliefs about sugar effects on memory were set to explain the online texts about sugar and memory (popular science text, commercial text).

As students were nested within 15 different classrooms, we calculated intra-class correlations for each evaluation item included in the statistical analyses (12 variables). The analysis showed that 0%–1.7% of the variance was explained by classroom level differences. Although the intra-class correlations were small, we used class as a clustering variable and estimated unbiased standard errors with the COMPLEX option.

To evaluate whether the models adequately represented the data, the following fit indices and cutoff values were used: $\chi^2$ test (*ns*, $p > 0.05$), root mean square error of approximation (RMSEA) values ≤0.06, Tucker–Lewis index (TLI) and comparative fit index (CFI) values ≥0.95 and Standardized Root Mean Square Residual (SRMR) values ≤0.08 (Hu & Bentler, 1999). We evaluated the competing measurement models by testing the more restricted model against the less restricted one (i.e., Model A vs. Model B; Model B vs. Model C; and Model D vs. Model C, see Figure 1) with the Chi-square ($\chi^2$) difference test (Satorra & Bentler, 2001).

After determining the best measurement model for students' credibility evaluation, we examined its association with students' ranking score (observed variable). In this model, the ranking score served as the dependent variable whereas the latent variables for credibility evaluation were set as independent variables predicting the ranking score.

The effect sizes were estimated with the correlation coefficient effect size *r* and interpreted according to the cut-offs suggested by Cohen (1988).

## 4 | RESULTS

### 4.1 | Descriptive statistics

Table 3 presents the descriptive statistics for students' credibility evaluations and prior beliefs. As shown, students scored higher when evaluating the credibility of the more credible online texts (popular science text and newspaper article) than when evaluating the less credible online texts (blog text and commercial text). The related test statistics (Wilcoxon Rank tests) comparing scores for more and less credible texts are presented in Table B1 in Appendix B. Table C1 in Appendix C shows the Spearman correlations among the students' credibility evaluations.

### 4.2 | The structure of credibility evaluation of online texts

Our first research question concerned the structure of credibility evaluation of online texts. Four competing models (see Figure 1) were examined. As shown in Table 4, our hypothesis that both two-factor (Confirming-Questioning Model) and four-factor (Genre-based Model) models fit the data better than the one-factor model (General Credibility Evaluation Model) was confirmed. Next, we compared the Confirming-Questioning Model with the Genre-based Model.

**TABLE 4** Model Fit Statistics of the competing measurement models for credibility evaluation of online texts and $\chi^2$ difference test

| Model fit statistic | General Evaluation Model (A) | Confirming-Questioning Model (B) | Genre-based Model (C) | Genre-based Confirming-Questioning Model (D) | $\chi^2$ difference test |
|---|---|---|---|---|---|
| $\chi^2$-test (df) | 479.99 (87) | 206.42 (83) | 105.39 (76) | 107.68 (79) | |
| | $p < 0.001$ | $p < 0.001$ | $p = 0.015$ | $p = 0.018$ | |
| $\Delta\chi^2$ (df) | | | | | |
| General Evaluation Model (A) versus Confirming-Questioning Model (B) | | | | | 273.57 (4); $p < 0.001$ |
| General Evaluation Model (A) versus Genre-based Model (C) | | | | | 374.60 (11); $p < 0.001$ |
| Confirming-Questioning Model (B) versus Genre-based Model (C) | | | | | 101.53 (7); $p < 0.001$ |
| Genre-based Confirming-Questioning Model (D) versus Genre-based Model (C)[a] | | | | | 2.29 (3); $p = 0.514$ |
| RMSEA | 0.131 | 0.075 | 0.038 | 0.037 | |
| CFI | 0.630 | 0.884 | 0.972 | 0.973 | |
| TLI | 0.566 | 0.857 | 0.963 | 0.965 | |
| SRMR | 0.195 | 0.110 | 0.078 | 0.082 | |

*Note*: $\Delta$ = difference.
[a]Because the $\chi^2$ difference-test showed that Model D fitted the data equally well as Model C, Model D was preferred on the grounds of parsimony (cf., Rindskopf & Rose, 1988).

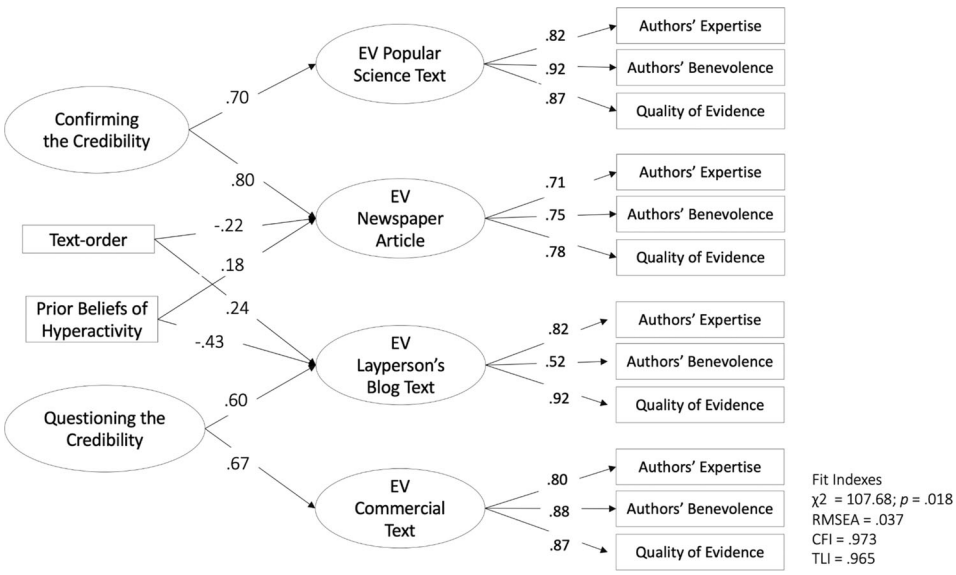


**FIGURE 3** The second-order factor model for students' credibility evaluation based on online texts and skills to confirm the credibility of more credible texts and skills to question the credibility of less credible texts. The coefficients are standardized estimates. All connections are at least $p < 0.05$

The Genre-based Model showed a good fit with the data and the $\chi^2$ difference test also supported its superiority over the Confirming-Questioning Model. Thus, of the hypothesised first-order models A–C (Figure 1), the model based on the four text genres seemed to be the best approximation to our data.

The results showed that evaluations of the two more credible online texts—the popular science text and the newspaper article—were positively associated with each other with a large effect size ($r = 0.59$). Thus, the more students were able to confirm the credibility of the popular science text, the better they also were at confirming the credibility of the newspaper article. A similar pattern was observed for credibility evaluation of the less credible texts. The more able students were at questioning the credibility of the layperson's blog text, the better they were at questioning the credibility of the commercial text. This association was also approaching a large effect size ($r = 0.48$). Notably, the association between the newspaper article and the commercial text was negative and rather weak ($r = -0.19$). The popular science text was not associated with either of the less credible texts.

Since the factors representing the two more credible online texts correlated with each other ($r = 0.59$), as did the two factors representing the less credible texts ($r = 0.48$), we further tested the hypothesised Genre-based Confirming-Questioning Model (model D) with two additional second-order factors: the first one of them was set to capture the common variance within the more credible texts and the second captured the common variance of the less credible texts. As shown in Table 4, the fit indexes of Models C (Genre-based Model) and D (Genre-based Confirming-Questioning Model) indicated that both models fit our data well. To determine the final model, the more restricted Model D (df = 79) was tested against the less restricted Model C (df = 76). The $\chi^2$ difference test showed that Model D fit the data as well as Model C. Therefore, we chose Model D as the final model on the grounds of parsimony (cf. Rindskopf & Rose, 1988). In addition, Model D is theoretically more plausible because it explains the high correlations between the two first-order, genre-based factors representing more credible online texts and between the two first-order, genre-based factors representing less credible online texts.

Examination of the final measurement model (see Figure 3) shows that all parameter estimates of the four first-order factors related to text genres were statistically significant (*p*s < 0.001). The factor loadings were all positive and above 0.70 with one exception, that is, the evaluation of the benevolence of the mother for the blog text. Thus, participants may tend to believe the mother acted in good faith.

Figure 3 also shows that the two first-order factors representing the evaluation of the more credible online texts loaded onto a higher-order latent credibility evaluation factor labelled confirming the credibility. The two first-order factors representing the evaluation of the less credible online texts reflected a higher-order latent credibility evaluation factor labelled questioning the credibility. To sum up, these factor loadings suggest that for sixth graders, the credibility evaluation of online texts representing various genres may require particular skills (viz., confirming and questioning credibility) that enable them to differentiate more credible texts from less credible texts.

### 4.3 | Associations between credibility evaluation and ranking of online texts

To address the second research question, we examined how confirming the credibility and questioning the credibility were associated with students' performance in ranking the online texts according to their credibility. Both second-order dimensions of credibility evaluation were positively associated with the ranking score. The better students were at confirming the credibility ($\beta = 0.41$, $p < 0.001$) or at questioning the credibility ($\beta = 0.54$, $p < 0.001$), the better they performed on the ranking task. It is worth noting that the correlation between confirming the credibility and questioning the credibility was −0.093 ($p = 0.358$), indicating that confirming the credibility and questioning the credibility independently predict students' ranking score.

## 5 | DISCUSSION

The main aim of the current study was to examine the structure of credibility evaluation of online texts among young adolescent students. The novelty of this study lies in the careful task design that involved students in evaluating four texts representing different genres, two of which were more credible and two of which were less credible. The credibility of the texts was based on manipulation of the author's expertise, the author's benevolence, and the quality of the evidence in accordance with general expectations for the selected text genres. This task design allowed us to examine the structure of credibility evaluation by comparing four different factor structures. An additional strength of this study is that the validity of the latent credibility evaluation factors (confirming and questioning the credibility) was verified by examining the association of the factors based on credibility evaluations with the students' ranking scores. In the following, we discuss our main findings, limitations of the study, future directions, and the theoretical, methodological and instructional implications of the study.

### 5.1 | The structure of credibility evaluation

As expected, the results did not support the unidimensional structure of credibility evaluation of online texts among sixth graders. This finding suggests that credibility evaluation among younger students may be more dependent on the source texts than has been found for more mature students (Hahnel et al., 2020). Instead of a unidimensional structure, we identified a multidimensional structure in which both texts representing different genres and text credibility seemed to play a role in students' credibility evaluation.

First, students' credibility evaluations were found to load onto four factors representing different text genres. This suggests that credibility evaluation is contextualized, supporting previous studies that have shown that genre matters in online evaluation (Flanagin & Metzger, 2007; List et al., 2017). So far, the importance of genre in credibility evaluation of online texts has mainly been observed among adult readers, however. Our study indicates that adolescent readers' genre knowledge may also guide their reading and evaluation online. It is worth restating that the text genres corresponded to differences in credibility aspects (expertise, benevolence and quality of evidence) that align with general genre expectations (see also, Flanagin & Metzger, 2007). In this study, the contribution of genre and credibility aspects cannot be separated. It may be that students just tended to evaluate each text as an entity and thereby all credibility aspects in a parallel manner. This would be in line with the assumption that first- and second-hand judgements are reciprocal, thus influencing each other (Barzilai et al., 2020). Keeping these issues in mind, it would be interesting to clarify how better and poorer evaluators utilize genre knowledge in their credibility evaluations. It also remains for future research to examine the role of genres in situations where the genre expectations are not fulfilled.

Second, we found two higher-order factors representing two distinct latent skills: confirming the credibility of more credible texts and questioning the credibility of less credible texts. This finding captures the difficulties many students had in evaluating the credibility of the less credible texts. Students' credibility evaluation scores for the less credible texts were considerably lower than the scores for the more credible online texts (see Table B1 in Appendix B), consistent with previous findings among younger readers (Kiili et al., 2018; Pieschl & Sivyer, 2021; Potocki et al., 2020). In addition, both latent credibility evaluation skills uniquely predicted students' ability to accurately rank the four online texts according to their credibility. This highlights the importance of instructing adolescent readers to understand criteria that make texts more credible as well as features that make texts less credible. Without balanced attention to both as part of classroom instruction, we may compromise younger students' critical online reading skill development.

Our findings also indicate that in early adolescence, critical online reading skills are often still emerging. The abilities needed to differentiate the credibility of the more and less credible texts seem to develop during secondary schooling (Barzilai et al., 2021; Pieschl & Sivyer, 2021). This may relate to maturation but also increasing attention to (and instruction of) critical evaluation at higher educational levels. It might also reflect broader sets of online experiences that are part of growing up. By 17, students are likely to have interacted with a broader set of internet texts and to have had many more experiences where they have needed to question the trustworthiness of online information (List, 2019). However, adolescents' active consumption of online information at an increasingly early age in conjunction with poor critical online reading skills may increase their susceptibility to false information on the internet. Future research is urgently needed to determine factors and methods that support critical online reading development during primary education.

## 5.2 | Limitations

This study comes with some limitations. First, although the coherent text design can be regarded as a strength of this study, the real online world is messier. Blogs can be authored by experts relying on scientific evidence (Pieschl & Sivyer, 2021), or experts may have commercial intentions (Potocki et al., 2020). Thus, the structure found in this study captures credibility evaluation of online texts whose attributes signal either higher or lower credibility. For these reasons, it is an open question to what extent the findings of this study can be generalized to authentic online settings. Second, students evaluated only four online texts, one representing each genre. However, reading four texts and responding to the related items were already quite challenging for the sixth graders. It remains for future studies to examine the structure of credibility evaluation with more texts and more complex text designs.

Third, the structure of credibility evaluation found in this study is restricted to a certain age group and cannot be generalized to other age groups. Future studies could examine the structure of credibility evaluation among students at secondary level and beyond.

Fourth, students were scaffolded to identify the correct author, claim and evidence before they were asked to evaluate the author's expertise, the author's benevolence and the quality of the evidence. Students were given feedback about whether or not they identified the author, claim, or evidence correctly, and if they failed, they were shown the correct option. The received feedback may have altered students' confidence and engagement in subsequent items (Bandura, 1997). On the flip side of the coin, however, the scaffolds allowed us to ensure that students actually evaluated the author and evidence in question.

Finally, one of the items, concerning journalist expertise, could have been scored differently. We credited students with two points if they evaluated the journalist's expertise by choosing one of the two highest numbers on the scale (5 or 6). We did this even though the journalist had a medium level of expertise (compared to the other texts in the task) because we assumed that students at this age generally do not have specific knowledge about journalists' expertise, when compared with more mature students. However, this decision may have underestimated some students' skills.

## 5.3 | Theoretical and methodological implications

Theoretically, this study extends our understanding of the structure of credibility evaluation of online texts, which often is portrayed as a single construct in models of reading on the internet (Brand-Gruwel et al., 2005; Leu et al., 2004). Importantly, these theoretical models were developed when the internet was still emerging, and before the online landscape included persuasive and manipulative choice architectures and the rapid spread of false information (Kozyreva et al., 2020). The present study revealed that in certain circumstances, the structure of credibility evaluation might be more complex than previously conceptualized. As such, it supports the earlier finding (Kiili et al., 2018) that confirming the credibility of more credible texts and questioning the credibility of less credible texts may be two separate credibility evaluation skills, using more texts and items in the assessment. In light of this evidence, there is a need for further theoretical considerations that address the complex structure of credibility evaluation.

Methodologically, the CORRE (Critical Online Reading Research Environment) developed for research purposes, provides a promising way of measuring students' credibility evaluation of online texts, at least among readers in their early adolescence. First, the study showed that the credibility evaluation items and the ranking item measured students' credibility evaluation skills consistently. Second, there is a need for more consistency in measuring students' credibility evaluation skills across different cultures and educational levels. This was shown by a recent review (Anmarkrud et al., 2022) examining individual differences in sourcing. The review included 72 studies, many of which investigated evaluation of source credibility. The review revealed that there are many various measures to assess

sourcing, which complicates the comparison of research findings. The CORRE may provide new avenues for valid and more consistent assessment in the field. However, additional research is needed to assess the usefulness of the tasks created for the CORRE with culturally diverse groups of students.

Third, the task created with the CORRE is well aligned with many nations' curricula and standards that recognize the need for educating critical online readers (e.g., Common Core State Standards Initiative, 2010; Finnish National Core Curriculum for Basic Education, 2014). Pellegrino et al. (2016) termed this type of alignment instructional validity of an assessment. Instructionally valid assessments provide students with learning opportunities and support teachers' practices. If used during instruction, the tasks created with the CORRE could provide feedback to students and information to teachers on how well their students can differentiate between more and less credible online texts, and what kinds of texts students struggle to evaluate critically. Actually, this is how we operated in this study. We gave individual, positive feedback to all students and information to teachers on how well their students performed in different areas of credibility evaluation. Future classroom-based research with the CORRE could provide new insights into the value of this type of feedback for students' learning to evaluate the text credibility over time.

### 5.4 | Instructional implications

Imagine a frequently assigned online inquiry task in schools: Students are instructed to use the internet to find information for a presentation or an essay in social studies or science. Teachers emphasize that students must search for credible resources, justify their resource selections and record resources used. This type of task underscores the importance of credibility evaluation of online texts and, when students are well prepared, it creates an opportunity to practice identifying more and less credible information sources. However, without adequate preparation, this activity may encourage students to use superficial heuristics to exclude less credible texts without offering opportunities to think thoroughly about features that make the excluded texts less credible.

Given the multifaceted structure of credibility evaluation shown in this analysis, we think that students would benefit from activities that ask them to carefully analyse different credibility aspects of less credible texts and connections between these aspects. For example, during an online inquiry task, students could be asked to point out two less credible texts they did not select and justify why not. The use of contrasting cases (Bråten et al., 2019), that is, asking students to compare more and less credible texts is another way to support students' skills in questioning the credibility of less credible online texts.

Further, closed task environments, such as the CORRE used in this study, may provide valuable instructional support, including prompts helping students focus their attention on various features relevant to the credibility of online texts and feedback given via an avatar. In this study, however, the feedback provided was simple and could presumably be developed to be better adapted to students' individual needs.

Our results also suggest that students' credibility evaluation skills could be supported by practicing evaluation of texts representing different genres. This could provide students with genre knowledge that they can use when evaluating online texts. In order to cope with the complexities of the internet and avoid superficial credibility evaluation solely based on genres, students should also be supported in evaluating online texts that do not match the general genre expectations. These data offer new evidence to support the use of such instructional practices with younger adolescents who can benefit from tailored supports that align with their unique developmental needs as emergent online readers and evaluators of internet information.

## ACKNOWLEDGEMENTS

The study was funded by the Academy of Finland (No. 324524). The authors would like to thank Jari Hämäläinen, Symcode Oy for the development of the CORRE and Riikka Anttonen for helping in the data collection.

## FUNDING INFORMATION

The study was funded by the Academy of Finland (No. 324524).

## CONFLICT OF INTEREST

The authors report no conflict of interest.

## PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12779.

## DATA AVAILABILITY STATEMENT

Data available on request from the authors.

## ETHICS STATEMENT

The ethical statement has been given by the Ethics Committee of the Tampere Region.

## ORCID

*Carita Kiili* https://orcid.org/0000-0001-9189-4094
*Eija Räikkönen* https://orcid.org/0000-0003-4450-9178
*Ivar Bråten* https://orcid.org/0000-0002-9242-9087
*Helge I. Strømsø* https://orcid.org/0000-0003-1836-3339
*Michelle Schira Hagerman* https://orcid.org/0000-0002-7743-6334

## TWITTER

*Carita Kiili* @ckiili

## REFERENCES

Abendroth, J., & Richter, T. (2021). Mere plausibility enhances comprehension: The role of plausibility in comprehending an unfamiliar scientific debate. *Journal of Educational Psychology*, *113*(7), 1304–1322. https://doi.org/10.1037/edu0000651

Anmarkrud, Ø., Bråten, I., Florit, E., & Mason, L. (2022). The role of individual differences in sourcing: A systematic review. *Educational Psychology Review*, *34*, 749–792. https://doi.org/10.1007/s10648-021-09640-7

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.

Barzilai, S., & Chinn, C. A. (2020). A review of educational responses to the "post-truth" condition: Four lenses on "post-truth" problems. *Educational Psychologist*, *55*(3), 107–119. https://doi.org/10.1080/00461520.2020.1786388

Barzilai, S., Tal-Savir, D., Abed, F., Mor-Hagani, S., & Zohar, A. R. (2021). Mapping multiple documents: From constructing multiple document models to argumentative writing. *Reading and Writing*. https://doi.org/10.1007/s11145-021-10208-8. Online ahead of print.

Barzilai, S., Thomm, E., & Shlomi-Elooz, T. (2020). Dealing with disagreement: The roles of topic familiarity and disagreement explanation in evaluation of conflicting expert claims and sources. *Learning and Instruction*, *69*, e101367. https://doi.org/10.1016/j.learninstruc.2020.101367

Bauman, M. L. (1999). The evolution of internet genres. *Computers and Composition*, *16*(2), 269–282. https://doi.org/10.1016/S8755-4615(99)00007-9

Berkenkotter, C., & Huckin, T. (1995). *Genre knowledge in disciplinary communication: Cognition, culture, power*. Erlbaum.

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving: Analysis of a complex cognitive skill. *Computers in Human Behavior*, *21*, 487–508. https://doi.org/10.1016/j.chb.2004.10.005

Bråten, I., Braasch, J. L., Strømsø, H. I., & Ferguson, L. E. (2015). Establishing trustworthiness when students read multiple documents containing conflicting scientific evidence. *Reading Psychology*, *36*(4), 315–349. https://doi.org/10.1080/02702711.2013.864362

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

Bråten, I., Stadtler, M., & Salmerón, L. (2018). The role of sourcing in discourse comprehension. In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *Routledge handbooks in linguistics. The Routledge handbook of discourse processes* (pp. 141–166). Routledge/Taylor & Francis.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Coiro, J., Coscarelli, C., Maykel, C., & Forzani, E. (2015). Investigating criteria that seventh graders use to evaluate the quality of online information. *Journal of Adolescent & Adult Literacy*, *59*(3), 287–297. https://doi.org/10.1002/jaal.448

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects*. Council of Chief State School Officers and National Governors Association. Common Core State Standards Initiative. http://www.corestandards.org/

Crowston, K. (2010). Internet genres. In M. Bates (Ed.), *Encyclopedia of library and information science* (pp. 2583–2596). Taylor & Francis.

Duke, N. K., & Roberts, K. L. (2010). The genre-specific nature of reading comprehension. In D. Wise, R. Andrews, & J. Hoffman (Eds.), *The Routledge international handbook of English, language and literacy teaching* (pp. 74–86). Routledge.

Finnish National Core Curriculum for Basic Education. (2014). *National core curriculum for basic education 2014*. Finnish National Board of Education. (Publication No. 2016:5).

Fisher, R. (2019). Reconciling disciplinary literacy perspectives with genre-oriented activity theory: Toward a fuller synthesis of traditions. *Reading Research Quarterly*, *54*(2), 237–251. https://doi.org/10.1002/rrq.233

Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, *9*(2), 319–342. https://doi.org/10.1177/1461444807075015

Forzani, E. (2020). A three-tiered framework for proactive critical evaluation during online inquiry. *Journal of Adolescent & Adult Literacy*, *63*(4), 401–414. https://doi.org/10.1002/jaal.1004

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328–351.

Hahnel, C., Eichmann, B., & Goldhammer, F. (2020). Evaluation of online information in university students: Development and scaling of the screening instrument EVON. *Frontiers in Psychology*, *11*, 562128. https://doi.org/10.3389/fpsyg.2020.562128

Hendriks, F., Kienhues, D., & Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The muenster epistemic trustworthiness inventory (METI). *PloS ONE*, *10*(10), e0139309. https://doi.org/10.1371/journal.pone.0139309

Hendriks, F., Kienhues, D., & Bromme, R. (2016). Evoking vigilance: Would you (dis) trust a scientist who discusses ethical implications of research in a science blog? *Public Understanding of Science*, *25*(8), 992–1008. https://doi.org/10.1177/0963662516646048

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Kiili, C., Forzani, E., Brante, E. W., Räikkönen, E., & Marttunen, M. (2021). Sourcing on the internet: Examining the relations among different phases of online inquiry. *Computers and Education Open*, *2*, 100037. https://doi.org/10.1016/j.caeo.2021.100037

Kiili, C., Leu, D. J., Utriainen, J., Coiro, J., Kanniainen, L., Tolvanen, A., Lohvansuu, K., & Leppänen, P. H. T. (2018). Reading to learn from online information: Modeling the factor structure. *Journal of Literacy Research*, *50*, 304–334. https://doi.org/10.1177/1086296X18784640

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Leeder, C. (2016). Student misidentification of online genres. *Library & Information Science Research*, *38*(2), 125–132. https://doi.org/10.1016/j.lisr.2016.04.003

Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2019). New literacies: A dual level theory of the changing nature of literacy, instruction, and assessment. In D. E. Alvermann, N. J. Unrau, M. Sailors, & R. B. Ruddell (Eds.), *Theoretical models and processes of literacy* (7th ed., pp. 319–346). Taylor & Francis.

Leu, D. J., Kinzer, C. K., Coiro, J. L., & Cammack, D. W. (2004). Toward a theory of new literacies emerging from internet and other information and communication technologies. In R. B. Ruddell & N. Unrau (Eds.), *Theoretical models and process of reading* (5th ed., pp. 1570–1613). International Reading Association.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949. https://doi.org/10.3758/s13428-015-0619-7

List, A. (2019). Defining digital literacy development: An examination of pre-service teachers' beliefs. *Computers and Education*, *138*, 146–158. https://doi.org/10.1016/j.compedu.2019.03.009

List, A., & Alexander, P. A. (2018). Corroborating students' self-reports of source evaluation. *Behaviour & Information Technology*, *37*(3), 198–216.

List, A., Alexander, P. A., & Stephens, L. A. (2017). Trust but verify: Examining the association between students' sourcing behaviors and ratings of text trustworthiness. *Discourse Processes*, *54*(2), 83–104. https://doi.org/10.1080/0163853X.2016.1174654

Mason, L., Junyent, A. A., & Tornatora, M. C. (2014). Epistemic evaluation and comprehension of web-source information on controversial science-related topics: Effects of a short-term instructional intervention.

*Computers & Education*, *76*, 143–157. https://doi.org/10.1016/j.compedu.2014.03.016

McCrudden, M. T. (2018). Text relevance and multiple-source use. In J. L. G. Braasch, I. Bråten, & M. T. McCrudden (Eds.), *Handbook of multiple source use* (pp. 168–183). Routledge.

McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory & Research in Social Education*, *46*(2), 165–193. https://doi.org/10.1080/00933104.2017.1416320

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

Nussbaum, E. M. (2020). Critical integrative argumentation: Toward complexity in students' thinking. *Educational Psychologist*, *56*(1), 1–17. https://doi.org/10.1080/00461520.2020.1845173

Official Statistics of Finland. (2020). *Educational structure of population. Population with educational qualification by level of education, field of education and gender*. Statistics Finland.

Okan, O., Bollweg, T. M., Berens, E. M., Hurrelmann, K., Bauer, U., & Schaeffer, D. (2020). Coronavirus-related health literacy: A cross-sectional study in adults during the COVID-19 infodemic in Germany. *International Journal of Environmental Research and Public Health*, *17*(15), 5503. https://doi.org/10.3390/ijerph17155503

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Pieschl, S., & Sivyer, D. (2021). Secondary students' epistemic thinking and year as predictors of critical source evaluation of internet blogs. *Computers & Education*, *160*, 104038. https://doi.org/10.1016/j.compedu.2020.104038

Potocki, A., de Pereyra, G., Ros, C., Macedo-Rouet, M., Stadtler, M., Salmerón, L., & Rouet, J. F. (2020). The development of source evaluation skills during adolescence: Exploring different levels of source processing and their relationships. *Journal for the Study of Education and Development*, *43*(1), 19–59. https://doi.org/10.1080/02103702.2019.1690848

Purcell-Gates, V., Duke, N., & Martineau, J. (2007). Learning to read and write genre-specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly*, *42*, 8–46. https://doi.org/10.1598/RRQ.42.1.1

Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist*, *52*(3), 148–166. https://doi.org/10.1080/00461520.2017.1322968

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51–67. https://doi.org/10.1207/s15327906mbr2301_3

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

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

Sundin, O., & Francke, H. (2009). In search of credibility: Pupils' information practices in learning environments. *Information Research*, *14*(4). http://informationr.net/ir/14-4/paper418.html

Tangcharoensathien, V., Calleja, N., Nguyen, T., Purnat, T., D'Agostino, M., Garcia-Saiso, S., Landry, M., Rashidian, A., Hamilton, C., AbdAllah, A., Ghiga, I., Hill, A., Hougendobler, D., van Andel, J., Nunn, M., Brooks, I., Sacco, P. L., De Domenico, M., Mai, P., … Briand, S. (2020). Framework for managing the COVID-19 infodemic: Methods and results of an online, crowdsourced WHO technical consultation. *Journal of Medical Internet Research*, *22*(6), e19659. https://www.jmir.org/2020/6/e21820/

Thomm, E., & Bromme, R. (2016). How source information shapes lay interpretations of science conflicts: Interplay between sourcing, conflict explanation, source evaluation, and claim evaluation. *Reading and Writing*, *29*(8), 1629–1652. https://doi.org/10.1007/s11145-016-9638-8

Tierney, R. J., & Pearson, P. D. (2021). *A history of literacy education: Waves of research and practice*. Teachers College Press.

Zarefsky, D. (2019). *The practice of argumentation: Effective reasoning in communication*. Cambridge University Press.

**How to cite this article:** Kiili, C., Räikkönen, E., Bråten, I., Strømsø, H. I., & Hagerman, M. S. (2023). Examining the structure of credibility evaluation when sixth graders read online texts. *Journal of Computer Assisted Learning*, 1–16. https://doi.org/10.1111/jcal.12779

## APPENDIX A

### A.1 | Task assignment translated from Finnish

In this task, you will read four different online texts about the effects of sugar. Your job is to rank the texts based on how credible you think the texts are.

Fact checker Max will help you. Next, read what he has to say to you.

Hi!

I am fact-checker Max. My job is to evaluate the credibility of various texts. I am often asked to rank the texts according to their credibility. It may sound easy, but it is not. There are many things to keep in mind when making judgements!

In your task, you can try out the set of guiding questions that I have developed. Read each web page and answer carefully to the related questions. When you are done, use your answers when ranking the texts based on their credibility.

I will also read the texts and rank them. In the end, let us see if we were of the same opinion about the order!

Good luck!

## APPENDIX B

**TABLE B1** Wilcoxon rank tests between the more and less credible texts

| Credibility aspect | Texts | *Z* | *p* | Effect size (*r*) |
|---|---|---|---|---|
| Expertise | Popular science text versus Blog text | 9.88 | <0.001 | 0.61 |
| | Popular science text versus Commercial text | 10.20 | <0.001 | 0.63 |
| | Newspaper article versus Blog text | 8.86 | <0.001 | 0.54 |
| | Newspaper article versus Commercial text | 9.43 | <0.001 | 0.58 |
| Benevolence | Popular science text versus Blog text | 13.19 | <0.001 | 0.81 |
| | Popular science text versus Commercial text | 11.55 | <0.001 | 0.71 |
| | Newspaper article versus Blog text | 12.74 | <0.001 | 0.78 |
| | Newspaper article versus Commercial text | 10.67 | <0.001 | 0.66 |
| Quality of evidence | Popular science text versus Blog text | 11.21 | <0.001 | 0.69 |
| | Popular science text versus Commercial text | 10.91 | <0.001 | 0.67 |
| | Newspaper article versus Blog text | 11.02 | <0.001 | 0.68 |
| | Newspaper article versus Commercial text | 10.61 | <0.001 | 0.65 |

## APPENDIX C

**TABLE C1** Spearman correlations among students' credibility evaluation scores and prior beliefs ($N = 265$)

| | 2. | 3. | 4. | 5. | 6. | 7. | 8 | 9. | 10. | 11. | 12. | 13. | 14. | 15. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Blog expertise | 0.285*** | 0.620*** | 0.225*** | 0.284*** | 0.242*** | −0.144* | −0.034 | 0.052 | −0.027 | 0.010 | −0.021 | −0.384*** | 0.026 | 0.234*** |
| 2. Blog benevolence | | 0.382*** | 0.160** | 0.208*** | 0.180** | −0.119 | −0.188** | −0.038 | −0.106 | −0.042 | −0.103 | −0.194** | 0.051 | 0.133* |
| 3. Blog evidence | | | 0.245*** | 0.336*** | 0.359*** | −0.181** | −0.166** | −0.042 | −0.079 | 0.037 | −0.001 | −0.340*** | 0.014 | 0.293* |
| 4. Company expertise | | | | 0.598*** | 0.588*** | −0.054 | −0.060 | −0.095 | −0.061 | 0.023 | −0.104 | −0.156* | −0.042 | 0.226*** |
| 5. Company benevolence | | | | | 0.630*** | −0.076 | −0.116 | −0.074 | 0.030 | 0.064 | 0.040 | −0.177** | −0.100 | 0.338*** |
| 6. Company evidence | | | | | | −0.148* | −0.099 | −0.101 | −0.029 | −0.018 | −0.058 | −0.161** | −0.054 | 0.301*** |
| 7. News expertise | | | | | | | 0.447*** | 0.369*** | 0.266*** | 0.164** | 0.228*** | 0.107 | 0.016 | 0.054 |
| 8. News benevolence | | | | | | | | 0.338*** | 0.213*** | 0.269*** | 0.174** | 0.150* | 0.041 | 0.057 |
| 9. News evidence | | | | | | | | | 0.256*** | 0.267*** | 0.308*** | 131* | 0.087 | 0.132* |
| 10. Science expertise | | | | | | | | | | 0.507*** | 0.484*** | 0.057 | 0.060 | 0.177** |
| 11. Science benevolence | | | | | | | | | | | 0.537*** | 0.46 | 0.051 | 0.267*** |
| 12. Science evidence | | | | | | | | | | | | 0.051 | 0.037 | 0.180** |
| 13. Prior belief—Sugar causes hyperactivity | | | | | | | | | | | | | −0.032 | −0.116 |
| 14. Prior belief—Sugar improves memory | | | | | | | | | | | | | | 0.002 |
| 15. Ranking score | | | | | | | | | | | | | | |

*$p < 0.05$; **$p < 0.01$; ***$p < 0.00$.