

## ReviewAid: A Scaffolded Approach to Supporting Readers' Evaluation of Health News

Eun-Young Ko, Yeonsu Kim, Juho Kim  
eunyoungko@kaist.ac.kr, yeonsu.kim@kaist.ac.kr, juhokim@kaist.ac.kr  
Korea Advanced Institute of Science and Technology

**Abstract:** As health news stories affect the audiences' decision and behavior, readers need to understand and evaluate the information. However, assessing the quality of health news stories is challenging for non-expert readers, as it requires both media and scientific literacy. In this paper, we design a web interface, ReviewAid, that guides online readers' evaluation of health news stories with the evaluation criteria designed for non-experts. ReviewAid scaffolds readers' evaluation process by providing 1) explicit guidance that asks readers to distinguish the quality of media and scientific findings and 2) example comments for each criterion. Results from our study with 44 participants show that ReviewAid helps readers better connect the evaluation criteria to the specific news story. ReviewAid also enabled participants to conduct a more grounded and accurate evaluation of the content of scientific research.

### Introduction

The media coverage of scientific research has been an important channel of scientific information for the public, and health is among the most preferred topics for both science journalists and their audience. As health news affects audiences' decision or behavior (Pew Research Center, 2017), health news should deliver accurate information on the scientific findings and discuss its implication for readers. However, health news stories are often criticized for being oversimplified, inaccurate, or biased (Dudo, 2015), and social media worsens the situation by accelerating the spread of sensational and clickbait stories (Waszak et al., 2018). Recent COVID-19 misinformation scandals (e.g., the effectiveness of Hydroxychloroquine in COVID-19 treatment) show how the inappropriate dissemination of scientific research can actually harm public health (Wong, 2020).

Modern science education has been focusing on nurturing the public's ability to understand and evaluate scientific information in their everyday life (NRC, 2012). However, it is known that the public, even those who are scientifically literate, tends to take news stories at face value by deferring the evaluation to the science journalists or relying on their gut feeling (Shah et al., 2017). In fact, assessing health news stories is a complex task for readers, as it necessitates the use of both scientific and media literacy. One needs to gauge the quality of scientific evidence and evaluate how properly the information is delivered and interpreted at the same time.

The most widely used strategy is to use explicit evaluation criteria (see Table 1 for examples) that question the quality of media and scientific research. Evaluation criteria can provide conceptual knowledge on the required inquiry steps and, at the same time, can serve as a basis for quantifying the quality of news stories. Unfortunately, previous research showed that understanding and applying evaluation criteria itself is a challenging task, and people easily make a superficial or incorrect evaluation (Nicolaidou et al., 2011; Zhang et al., 2018). Therefore, in the classroom setting, instructors provide additional guidance on how to apply criteria and correct students' inappropriate assessments. In this paper, we bring this strategy of using evaluation criteria to support readers' assessment of health news stories in an informal, everyday reading context.

However, as there is no such instructor who can guide readers' evaluation process in the informal reading context, it is important to understand the potential misuse of evaluation criteria and provide preventive measures. We conducted a series of observations in which participants were asked to evaluate a health news story following the evaluation criteria by themselves. We found that readers confuse the quality of news stories with the quality of scientific research and fail to apply the evaluation criteria in the context of a specific news story.

Based on the observation, we designed a web interface, ReviewAid, that guides readers' assessment of health news stories. ReviewAid uses criteria designed for non-expert readers and guides readers to evaluate the level of information coverage, the validity of the research, and the quality of the interpretation of the research separately. ReviewAid presents example questions for each evaluation step to help readers understand and apply each criterion to the specific news story. We conducted a between-subjects study with 44 participants to understand the effect of ReviewAid on the quality of readers' assessment and self-efficacy in assessing health news stories. The result shows that participants in the ReviewAid condition better connected the criteria to the specific news story and conducted a more grounded and accurate evaluation of the content of scientific findings. Also, participants in the ReviewAid condition reported a higher level of self-efficacy in explaining how the news story can be improved than participants in the baseline condition.

## Background

We first provide background on what makes public communication of scientific research challenging. Then we review existing evaluation criteria developed by scholars and practitioners and discuss differences between the criteria used in science education and journalism.

### Challenges in public communication of scientific research

Health news stories are a result of the sequential effort of multiple players (e.g., scientists, PR team, and science journalists) in the production pathway. While there exist many factors that make the public communication of scientific research challenging, the most important and inevitable challenges are those raised by the complex nature of scientific research. Scientific research consists of components such as theoretical background, experiment design, and analysis, and this innate complexity of scientific research makes it hard for journalists to accurately deliver the information (Dunwoody, 1982). Another factor that complicates the public communication of science is the tentativeness of scientific research. The complex and conditional nature of scientific research sometimes limits the scientists' knowledge and control of a subject matter, making some studies have limited validity (Bromme et al., 2014; Peters & Dunwoody, 2016). Health-related research has an especially high level of tentativeness as it often involves human or living subjects, whereas researchers have less control over the condition (Sumner et al., 2014). Therefore, the certainty of the finding is not determined solely by single research but can be only established from a number of studies that align with it (Lee et al., 2012). Communicating the tentativeness of the finding is a very demanding task for journalists as it requires a huge amount of time and effort and sometimes costs readers' engagement and trust in scientific research and the media (Stocking, 1999).

### Evaluation criteria for health news stories

Researchers and practitioners in science education have developed sets of evaluation criteria to teach students how to conduct a scientific inquiry when reading media reports on scientific findings. They aim to teach what to consider in evaluating a scientific claim and what can threaten the validity of the claim delivered through the media (Jarman & McClune, 2007). They guide students to question the *content* of scientific research, such as participants, procedure, measures, or who the researchers are (Oliveras et al., 2013; Tsai et al., 2013). However, previous research showed that students tend to make a superficial or incorrect evaluation of the content of research and have to practice the evaluation multiple times to understand the evaluation criteria and use them appropriately (Nicolaidou et al. 2011; Donnelly et al., 2014).

While criteria developed for students are focused on the *content* of the research, criteria developed for the general public put more emphasis on the *context* of the research. Criteria developed by journalists (e.g., Health News Review.org) or public organizations (e.g., NCCIH (n.d.)) guide readers to ask questions on how news stories deliver scientific research, interpret it, and discuss it in a broader context. They include questions on cost and potential side effects of interventions studied (e.g., new medicine or doing a specific exercise), how strong the study findings are, or how certain and novel the finding is. However, it is difficult to assess the context of research without prior knowledge in the research delivered, and this remains a question as to how well an average health news reader can use these guidelines to evaluate health news stories.

It is important to assess the health news story regarding both *content* and *context* of the research. With evaluation criteria on the content of the research, readers can judge the validity and reliability of the research conducted. On the other hand, the evaluation criteria on the context of the research help readers understand the certainty of the research finding and appropriate application or consequences. In this paper, we design a system that guides readers to evaluate a health news story on both content and context of the research.

### Formative Study: Readers' challenges in evaluating health news stories

To better understand the challenges that readers face while evaluating health news stories, we conducted a series of observations and semi-structured interviews. Specifically, we sought to identify readers' difficulties in understanding and applying given evaluation criteria in the online reading environment, where no additional guidance and support from an expert or instructor is expected.

### Method

We recruited eight participants (6 male, 2 female) from an online community of a technical university in South Korea. Participation was limited to those who are fluent in English and have no prior research experience. Five participants were undergraduate students, two were entering graduate students, and one was a first-year graduate student. We prepared two news stories titled "How Your Morning Coffee Might Slow Down Aging" (Park, 2017) and "Weight-loss pills can help. So why don't more people use them?" (Carroll, 2018), from stories with low

expert ratings (less than 3 out of 5) in HealthNewsReview.org. Participants were asked to choose a news story they are more interested in, evaluate the story by following evaluation criteria provided, and explain rationales for their assessment. After the session, we conducted a short interview on their overall experience and difficulties they faced during the evaluation process. All of the sessions were conducted individually. Each participant was paid KRW 10,000 (approx. \$8) for their hour-long participation.

We asked participants to evaluate one health news story following two sets of criteria, Always Ask (AA) (Jarman & McClune, 2007) and HealthNewsReview.org (HNR). AA is evaluation criteria designed to teach scientific inquiry steps for students and has more focus on the content of research. AA comes with detailed and explicit subcriteria for each high-level criterion. On the other hand, HNR is criteria developed by groups of journalists and scientists and has more focus on the context of research and how the story frames and discusses the scientific research. We used both sets of criteria to observe readers' difficulties in applying criteria on the content and context of the research. Table 1 shows example criteria from AA and HNR. Four participants were guided to follow AA first, and the other four were guided to follow HNR first.

**Table 1**

*Example evaluation criteria used in Always Ask and Health News Review*

Source	Always Ask	Health News Review
Example criteria	How was the research conducted? - What were the subjects of the study? - What was the sample size? - How was the experiment carried out?	Does the story adequately explain the harms of the intervention? Does the story establish the true novelty of the approach?

## Findings

First of all, seven (out of eight) participants said that the evaluation criteria helped them to learn what to consider. Four participants explicitly mentioned that they could realize that some questions are very important only after seeing them. Despite their positive comments on having evaluation criteria, participants had difficulties across various steps, from understanding each criterion to applying the criteria to evaluate the news story.

### Providing a rationale for their evaluation

Participants often failed to provide a rationale for their evaluation. This happened a lot when participants were evaluating the news story with HNR, which does not give explicit subcriteria for each criterion. P8 said, "I think I understood what each (HNR) criterion is asking for, but this does not mean I know what to consider. I can give a score based on my impression, but I cannot explain why I gave that score." Most participants said they are more confident with their evaluation given to AA than HNR as the subcriteria helped them understand concrete questions that they can ask and develop their thoughts. To prevent readers from assessing health news stories by their gut feeling or impression, it is important to guide them with detailed and explicit criteria.

### Facing evaluation criteria that require prior knowledge or external information

Criteria that ask about the value (e.g., novelty or implication) of the delivered research require external information or expert knowledge. To answer questions such as "Does this story establish the true novelty of the approach?" (HNR) or "What is the importance of this study?" (AA), readers need to have prior knowledge in the topic or even knowledge of previous research in the domain. Participants said that they feel helpless when they are asked to evaluate such criteria. P4 said, "It seems that I need to conduct an extensive investigation to really judge the novelty of this research, and I don't think I can correctly answer the question even after the investigation." P2 said that "Some criteria were meaningless to me as there was nothing that I could do."

This does not mean, however, that those criteria should not be given to readers. Having such criteria is still valuable as they suggest and teach readers what to consider with health news stories. Rather, the benefits of such criteria can be maintained without discouraging the readers by adjusting the scope and target of the evaluation. For example, rather than asking readers to evaluate the novelty of the research, readers can still check whether the news story has explained or discussed how novel the research is.

### Distinguishing the quality of news story and the quality of scientific research covered

The quality of research and how it is delivered affect the quality of health news stories. However, participants failed to distinguish the quality of the story from the quality of the research. Six participants mistook the lack of information for lack of validity in the research. For example, there were three participants who concluded the research has limited validity by making a wrong assumption that researchers did not control potential confounding variables. Only two participants said that there is not enough information to evaluate the validity of the research.

It is important for readers to clearly distinguish the source of the problem because the consequences of evaluation can differ a lot. For example, suppose a patient reads a low-quality health news story on potentially effective intervention. If the patient concludes that the finding is insignificant or invalid, the consequence can be just ignoring the information. On the other hand, if the patient concludes that the media did a poor job delivering the research, the patient can find more information about the research or consult with his doctor.

### Applying evaluation criteria to the context of an individual news story

As scientific research varies in its topic and methodology, specific information that is critical for readers can also vary for each health news story. In our observation, participants expressed difficulties in applying criteria to the specific health news story being evaluated. This happens even when participants think they conceptually understand what the evaluation criteria are asking for. Participants said they could point out that some criteria are not met, but they could not explicitly say what should be included or improved in the news story. For example, regarding the criterion “What data were collected” (AA), P5 noted that “I can see that there exists limited information on the data collected but cannot think of what it is specifically.”

## **ReviewAid**

Based on our observations, we designed a web interface called ReviewAid that scaffolds online readers’ evaluation of health news. In ReviewAid, readers review a health news story by following the evaluation criteria designed for non-expert readers. For each subcriterion, ReviewAid provides explicit guidance on medium-research distinction and helps readers connect the evaluation criteria and the target news article with examples.

In ReviewAid, readers evaluate a health news story by following seven evaluation criteria that we developed by restructuring existing criteria, namely AA, HNR, and the criteria based on the taxonomy developed in Korpan et al. (1997) (e.g., Zimmerman et al., 2001). Criteria were chosen and designed so that they support scientific inquiry on both the *content* and *context* of the research. Criteria 1-5 are introduced to scaffold the scientific inquiry process explicitly. Criterion 6 is on its connection with other research or theory, and criterion 7 asks about its application and implication to the real world. To help readers better understand each criterion, ReviewAid provides two to four subcriteria (see Figure 1-(C), for example) for each criterion. When a reader selects a criterion to review, corresponding subcriteria are prompted. Readers evaluate the story for each subcriterion and then assess the high-level criterion based on it.

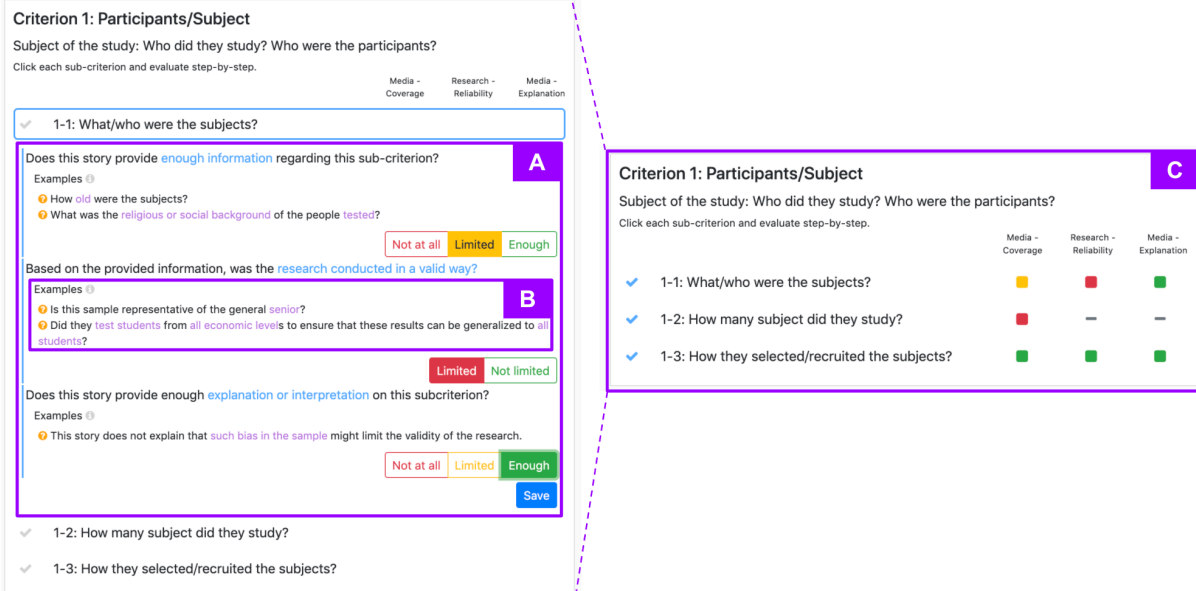
1. Participants/Subject: Who did the research study? Who were the participants?
2. Research Design: How and why is this research designed, and how is it done?
3. Measure: How were the factor (putative cause) and effect defined and measured?
4. Data & Statistics: How were the raw data and statistical results?
5. Social Context: Is there any social factor that may have influenced the research? Who did the research, funded it, or who is advertising it?
6. Theory & Related Research: Does the finding align with other research? Is there a theory that can explain the observed effect?
7. Application & Implication: What is the implication of this study finding, and how should I relate this finding to the real world?

When a reader selects a subcriterion to evaluate, ReviewAid provides procedural support to prevent them from confusing the quality of media and research. ReviewAid guides readers to evaluate the news article on 1) how well the story provides information (coverage), 2) how reliable the research is (reliability), and 3) how well the story explains or interprets the reliability of the research (explanation). The system prompts the evaluations of these aspects in order of coverage, reliability, and explanation. To prevent potential confusion between the media coverage and research validity, the step for research reliability is skipped if the reader answers there is no information. For the criteria on the contextual information (6 and 7) that require prior knowledge or external information, readers are not asked to assess them by themselves but to assess how the news story discusses them (explanation). Figure 1-(A) shows the evaluation steps ordered by the aspects.

For evaluating coverage, reliability, and explanation, ReviewAid shows example questions or comments related to each aspect. While each criterion comes in a general form that can be applied to any news story on health-related research, examples serve as concrete instantiations of each criterion that readers can refer to understand the criterion and apply it to their context. We constructed the example pool based on students’ questions raised on scientific news briefs collected in Korpan et al. (1994) and questions and comments left in expert reviews in HealthNewsReview.org. Phrases that are specific to the news stories that each question or comment is raised on were indicated separately (in purple color), as shown in Figure 1-(B).

**Figure 1**

Overview of the scaffolded evaluation process in ReviewAid. (A) For each subcriterion, the user evaluates the story for each of media coverage, research validity, and media interpretation (B) Example comments raised in other news stories are shown. Terms that are specific to the source news story are colored in purple. (C) Evaluations on the subcriteria are summarized.



**Criterion 1: Participants/Subject**  
Subject of the study: Who did they study? Who were the participants?  
Click each sub-criterion and evaluate step-by-step.

Media - Coverage    Research - Reliability    Media - Explanation

✓ 1-1: What/who were the subjects?

Does this story provide enough information regarding this sub-criterion?  
Examples ⓘ  
How old were the subjects?  
What was the religious or social background of the people tested?  
Not at all    Limited    Enough

Based on the provided information, was the research conducted in a valid way?  
Examples ⓘ  
Is this sample representative of the general senior?  
Did they test students from all economic levels to ensure that these results can be generalized to all students?  
Limited    Not limited

Does this story provide enough explanation or interpretation on this subcriterion?  
Examples ⓘ  
This story does not explain that such bias in the sample might limit the validity of the research.  
Not at all    Limited    Enough

Save

✓ 1-2: How many subject did they study?  
✓ 1-3: How they selected/recruited the subjects?

**Criterion 1: Participants/Subject**  
Subject of the study: Who did they study? Who were the participants?  
Click each sub-criterion and evaluate step-by-step.

Media - Coverage    Research - Reliability    Media - Explanation

✓ 1-1: What/who were the subjects?    ●    ●    ●  
✓ 1-2: How many subject did they study?    ●    -    -  
✓ 1-3: How they selected/recruited the subjects?    ●    ●    ●

## Evaluation

ReviewAid system scaffolds readers' assessment with 1) explicit and detailed evaluation criteria, 2) explicit procedural support to distinguish media and the research, and 3) examples that instantiate each criterion to a specific news story. We conducted a study to assess the effect of ReviewAid on the readers' evaluation of health news stories. As the benefits of having explicit and detailed criteria are well studied in previous research (e.g., Reiser et al., 2001), we focus on the effect of explicit guidance on medium-research distinction and examples. Specifically, we aim to understand readers' experience in evaluating health news stories with ReviewAid and the quality of assessment that they make.

## Method

We conducted a between-subjects study with two conditions: Baseline and ReviewAid. In both conditions, participants were asked to evaluate a health news story with seven criteria with 2-4 subcriteria for each. In the ReviewAid condition, participants were asked to follow the designed evaluation process for each subcriterion before giving a score to the story. In the Baseline condition, participants were shown the subcriteria, but the scaffolded evaluation process in ReviewAid was not presented.

### Participants and procedure

As ReviewAid is designed to support health news readers in their informal reading context, we tried to recruit participants from the general population. Previous research (Berinsky et al., 2012) showed that population samples recruited in Amazon Mechanical Turk (MTurk) tend to be more representative of the U.S. population than samples from in-person recruitment. We recruited 44 participants (20 male, 24 female) from MTurk and randomly assigned them to one of the two conditions. We had 21 and 23 participants for the Baseline and ReviewAid conditions, respectively. For their 40-60 minutes-long participation, participants were paid \$8. The average age was 36.5 (SD: 8.69). A total of 35 participants had tertiary education, and three among them had postgraduate education.

In the pre-survey, we asked questions on factors that may affect their experience and outcome of the main task, such as education level, prior experience in research-related activities, perceived media literacy, and how they prefer cognitive work (NFC scales in REI-10). In the main task, each participant evaluated a health news story titled "Tofu might harm memory in elderly." (Kahn, 2008) (adapted from Leung et al. (2015)) with seven criteria. For each criterion, participants were asked to rate the story in a 1-5 scale and explain their rationale for the score. In the post-survey, participants were asked to describe how the evaluation criteria and scaffolded evaluation process affected their evaluation. Also, we asked participants to rate how confident they are about 1)

judging the quality of a health news story, 2) pointing out inadequacies of a health news story, and 3) suggesting ways to improve a health news story in a 7-points Likert scale.

### Measuring the quality of assessment

To measure the quality of participants' assessment, we conducted a discourse analysis of the evaluation comment written by each participant. We measured the number of rationales provided in each comment, the number of rationales specific to the news story, and the number of incorrect rationales. The analysis was conducted with 88 comments on criterion #1 (participants/subject) and criterion #6 (theory and related research). These two criteria were chosen so that we can analyze comments on both content and context of research.

The analysis was done in three steps. One of the authors and one external coder (Ph.D. student in biology) worked together in every step. In the first step, we split each comment into multiple arguments so that each argument contains a single intention or meaning, and a total of 88 comments were divided into 417 arguments. In the second step, we marked whether each argument contains a rationale. To ensure consistency between coders, the two coders divided the first 10% of comments together and then divided the next 10% independently, compared the result, and discussed to reach an agreement. After the consistency building session, each coder independently coded the remaining arguments, and the inter-coder agreement was 0.79 (Cohen's Kappa). The two coders finalized the result by discussion. Out of 417 arguments, 318 (76.3%) were of rationale-type. In the last step, we labeled whether each rational-type argument is specific to the news story (binary) and correct (binary). The inter-coder agreements (Cohen's Kappa) were 0.86 and 0.83 for specificity and accuracy, respectively.

### Results

Overall, participants showed a moderate to high level of confidence in their media literacy (Mdn=5). There were no between-group differences in media literacy, as well as in the perception of health news and the need for cognition. The average time spent on the reviewing task was 19 minutes (longest: 45 mins, shortest: 7 mins). The average completion times for the Baseline and ReviewAid conditions were 17.3 and 21.0 minutes, respectively.

**Table 2**

*The numbers of participants who provided rationale, specific rationale, and incorrect rationale and the number of arguments provided respectively, with the median.*

		Criterion #1		Criterion #6	
		Baseline	ReviewAid	Baseline	ReviewAid
Participants	(1) Total	21	23	21	23
	(2) Rationale	17 (81.0%)	22 (95.7%)	21 (100.0%)	23 (100.0%)
	(3) Specific rationale	9 (42.9%)	19 (82.6%)	10 (47.6%)	18 (78.3%)
	(4) Incorrect rationale	10 (47.6%)	5 (21.7%)	3 (14.3%)	4 (17.4%)
Arguments	(5) Total	97, Mdn=4	149, Mdn=7	81, Mdn=3	90, Mdn=3
	(6) Rationale	60, Mdn=3	116, Mdn=5	66, Mdn=3	76, Mdn=3
	(7) Specific rationale	32, Mdn=0	89, Mdn=4	20, Mdn=0	28, Mdn=1
	(8) Incorrect rationale	14, Mdn=0	7, Mdn=0	5, Mdn=0	5, Mdn=0

Table 2 shows 1) the number of participants who provided rationale, specific rationale, and incorrect rationale and 2) the number of arguments with a rational, specific rationale, and incorrect rationale in each condition and criterion. Participants in the ReviewAid condition provided significantly more rationales (Mdn=5) than the Baseline participants (Mdn=3) for criterion #1 (Mann-Whitney (MW) Test,  $U=139.5$ ,  $p < .005$ ). However, there was no significant difference in the number of rationales for criterion #6 (Mdn=3 in both conditions), which is about the context of the research (MW Test,  $U=226.5$ ,  $p > .05$ ).

The proportion of participants who provided a rationale that is specific to the news story (see Table 2, row (3)) was higher in the ReviewAid condition, for both criteria ( $\chi^2$  Test,  $\chi^2(1, N=44)=7.50$ ,  $p < 0.01$  and  $\chi^2(1, N=44)=4.45$ ,  $p < .05$ , for criteria #1 and #6 respectively). The median number of specific rationale provided in each review was 0 and 4 for criterion #1, 0 and 1 for criterion #6 for Baseline and ReviewAid, respectively. The difference was statistically significant for criterion #1 but not for criterion #6 (MW Test,  $U=122$ ,  $p < .005$  and  $U=187.5$ ,  $p=0.184$ , respectively).

Regarding the accuracy of evaluation, for criterion #1, ten and five participants provided incorrect rationales in the Baseline and ReviewAid conditions, respectively. Among those who provided at least one rationale, 58.7% (10 out of 17) in Baseline and 22.7% (5 out of 22) in ReviewAid made incorrect arguments in their comments and the difference between the conditions is statistically significant ( $\chi^2$  Test,  $\chi^2(1, N=44)=5.27$ ,  $p < 0.05$ ). For criterion #6, 3 (out of 21) and 4 (out of 23) participants provided incorrect rationales in Baseline

and ReviewAid conditions ( $\chi^2$  Test,  $\chi^2(1, N=44)=0.12, p=0.720$ ). The median number of incorrect rationale in each comment was 0 for both conditions and both criteria.

Overall, participants showed a high level of self-efficacy with median score 5 for questions on how confident they are with 1) judging the quality of a health news story, 2) pointing out inadequacies of a health news story, and median score 6 for the question on 3) suggesting ways to improve a health news story. There was no difference between the conditions for the first (Mdn=5 in both conditions) and second (Mdn=6 in both conditions). For the third question on suggesting ways to improve, participants in the ReviewAid condition (Mdn=6) gave significantly higher scores than participants in the Baseline (Mdn=5) (MW Test,  $U=172, p<0.05$ ).

In both conditions, participants said that having detailed evaluation criteria helped them conduct a grounded evaluation. P19 in the Baseline said, “The subcriteria were very helpful in that it made it simple to know which direction to go in my evaluation.” Participants said that the scaffolded evaluation process in ReviewAid helped them conduct detailed and accurate assessments. P21 noted, “I think it (distinguishing media and research) added to my ability to see piece by piece what was affecting my opinion and better articulate my thoughts in my reviews. It helped me notice what was missing.” Also, P14 said, “Examples helped me fully understand what was being asked and how it applied to this specific article.”

## Discussion and future work

In our evaluation, ReviewAid had different effects on the type of evaluation criteria. For criterion #1, which is on the content of the research, participants in the ReviewAid condition provided more rationales and more specific rationales than those in the Baseline condition. Also, compared to the Baseline, the proportion of participants who provided wrong rationale was significantly smaller in the ReviewAid condition. For criterion #6, however, there was no difference between conditions in the number of rationales and the ratio of participants with incorrect rationales. As criterion #6 is on the context of research, ReviewAid does not present the guidance on medium-research distinction and the only difference between the Baseline and ReviewAid condition was the presence of examples. Our results illustrate that the examples given for criterion #6 helped participants better connect the criteria to the specific news story but not in providing more ground or accurately assessing the news story.

In our study, participants in the Baseline condition showed a moderate to high level of self-efficacy in assessing health news stories and said that having detailed evaluation criteria helped their assessment. However, we saw that such confidence does not guarantee the accuracy of their assessment. As readers are confident in their evaluation, it is hard for them to realize that their evaluation is incorrect. This illustrates how simply adapting tools from formal learning can result in undesired outcomes in the informal learning context. It is important to understand the differences between the two learning conditions and specific challenges in the informal setting.

Although our study results show the benefits of having the scaffolded evaluation process, we note that the observed effect needs to be further verified in future studies. The effect of having the ReviewAid system should be further evaluated in studies with large numbers of participants and on various health news stories. Another limitation of this study is the lack of consideration of readers’ prior knowledge and beliefs on the subject matter. Prior research showed that prior knowledge and beliefs on the subject matter could significantly affect one’s reasoning process, especially with socio-scientific issues such as climate change and genetic engineering (Christenson et al., 2014). Future studies can investigate online readers’ challenges in relation to their prior knowledge or belief and design a system that supports their evaluation of health news stories.

We also emphasize the need for designing a system that guides online readers’ evaluation of health news stories that is more practical to be used by online readers. In our study, participants spent 20 minutes evaluating a single news story on average. This questions the practicality and impact of the ReviewAid system. In our future work, we will investigate a more practical use of the ReviewAid system by supporting collaborative and shared evaluation of health news stories among readers.

Lastly, while this work focuses on the health news stories, the idea of supporting non-expert readers to distinguish media and the source information and conduct contextualized evaluation can be transferred to other domains where expert knowledge is disseminated and communicated to the public, such as economy or policy. We envision a future where non-experts can actively engage in a structured evaluation process with various types of information.

## References

- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com’s Mechanical Turk. *Political analysis*, 20(3), 351-368.
- Bromme, R., & Goldman, S. R. (2014). The public’s bounded understanding of science. *Educational Psychologist*, 49(2), 59-69.
- Carroll (2018, September 2). Weight-loss pills can help. So why don’t more people use them?. *NBC News*.

- Christenson, N., Chang Rundgren, SN., & Zeidler, D. L. (2014). The relationship of discipline background to upper secondary students' argumentation on socioscientific issues. *Research in Science Education*, 44(4), 581-601.
- Donnelly, D. F., Linn, M. C., & Ludvigsen, S. (2014). Impacts and characteristics of computer-based science inquiry learning environments for precollege students. *Review of Educational Research*, 84(4), 572-608.
- Dudo, A. (2015). Scientists, the media, and the public communication of science. *Sociology Compass*, 9(9), 761-775.
- Dunwoody, S. (1982). A question of accuracy. *IEEE Transactions on Professional Communication*, (4), 196-199.
- Jarman, R., & McClune, B. (2007). *Developing Scientific Literacy: Using News Media In The Classroom: Using News Media in the Classroom*. McGraw-Hill Education (UK).
- Kahn (2008, July 5). Tofu might harm memory in elderly. *The Telegraph*.
- Korpan, C. A., Bisanz, G. L., Bisanz, J., & Henderson, J. M. (1997). Assessing literacy in science: Evaluation of scientific news briefs. *Science Education*, 81(5), 515-532.
- Lee, P. N., Forey, B. A., & Coombs, K. J. (2012). Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC cancer*, 12(1), 1-90.
- Leung, J. S. C., Wong, A. S. L., & Yung, B. H. W. (2015). Understandings of nature of science and multiple perspective evaluation of science news by non-science majors. *Science & Education*, 24(7), 887-912.
- National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
- National Center for Complementary and Integrative Health (n.d.). Checklist for Understanding Health News Stories. <https://www.nccih.nih.gov/health/know-science/facts-health-news-stories/page-10>
- Nicolaidou, I., Kyza, E. A., Terzian, F., Hadjichambis, A., & Kafouris, D. (2011). A framework for scaffolding students' assessment of the credibility of evidence. *Journal of Research in Science Teaching*, 48(7), 711-744.
- Oliveras, B., Márquez, C., & Sanmartí, N. (2013). The use of newspaper articles as a tool to develop critical thinking in science classes. *International Journal of Science Education*, 35(6), 885-905.
- Park (2017, January 16). How Your Morning Coffee Might Slow Down Aging. *Time*.
- Peters, H. P., & Dunwoody, S. (2016). *Scientific uncertainty in media content: Introduction to this special issue*. Pew Research Center (2017, September). Science news and information today. Pew Research Center.
- Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B. K., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. *Cognition and instruction: Twenty-five years of progress*, 263-305.
- Shah, P., Michal, A., Ibrahim, A., Rhodes, R., & Rodriguez, F. (2017). What makes everyday scientific reasoning so challenging?. In *Psychology of learning and motivation* (Vol. 66, pp. 251-299). Academic Press.
- Stocking, S. H. (1999). How journalists deal with scientific uncertainty. *Communicating uncertainty: Media coverage of new and controversial science*, 23-41.
- Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., ... & Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, 349.
- Tsai, P. Y., Chen, S., Chang, H. P., & Chang, W. H. (2013). Effects of Prompting Critical Reading of Science News on Seventh Graders' Cognitive Achievement. *International Journal of Environmental and Science Education*, 8(1), 85-107.
- Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media—the pilot quantitative study. *Health policy and technology*, 7(2), 115-118.
- Wong (2020, April 7). Hydroxychloroquine: how an unproven drug became Trump's coronavirus 'miracle cure'. *The Guardian*.
- Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., Adams, N.B, Vincent, E., Lee, J., Robbins, M., Bice, E., Hawke, S., Karger, D. & Mina, A. X. (2018, April). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of The Web Conference 2018* (pp. 603-612).
- Zimmerman, C., Bisanz, G. L., Bisanz, J., Klein, J. S., & Klein, P. (2001). Science at the supermarket: A comparison of what appears in the popular press, experts' advice to readers, and what students want to know. *Public Understanding of Science*, 10(1), 37-58.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1C1C1007587).