TREC 2024 Lateral Reading Track Assessing Instructions

Version 1.1, August 20th, 2024
Track Organizers: Dake Zhang, Mark Smucker, Charles Clarke
Track Website: [TREC 2024 Lateral Reading Track Guidelines](#)

I. Assessor Instructions

Assume there is a (general public) reader who is looking through an online news article. Your first task is to suggest questions that the reader should ask to determine its trustworthiness. Your second task is to judge the helpfulness of the questions submitted by track participants for the reader's trustworthiness evaluation. Your third task is to judge the usefulness of retrieved documents to answer the questions pooled from the questions submitted by participants.

II. Background: Lateral Reading

Media literacy and the ability to read critically have long been viewed as important skills for people in the digital age. *Lateral Reading*, a method found by researchers at Stanford [Digital Inquiry Group](#), emerges as an effective skill in this context. Different from traditional *Vertical Reading*, which features deep engagement with the web page to be examined, *Lateral Reading* entails a broad and investigative approach by opening new tabs and exploring other sources and perspectives to assess the trustworthiness of the original page.

In this track, we utilize the idea from Lateral Reading that by placing the news article in a broader context and cross-verifying facts, claims, and the reputation of sources, readers can have a more accurate evaluation of the trustworthiness of online news.

Below are the three key questions of Lateral Reading:

- **Investigate the Source**
    - Question: Who is behind this story?
    - Investigate the media outlet and the author's background.
    - Understand their biases (ideologies, stances, factualness), motivations, and potential conflicts of interest (sponsors).
    - This can help you anticipate the stance they might take.
    - Example: If an article about climate change is published by an oil company, consider its potential bias and motivation to downplay environmental impacts.

- **Validate the Evidence**
    - Question: What is the evidence supporting the claims?
    - Evaluate the quality and context of the evidence mentioned to support claims.
    - Ensure it is NOT fabricated, taken out of context, or misinterpreted.
    - Check the credentials of the evidence. Is the evidence itself trustworthy?
    - Example: An article claims a new drug is highly effective. Check if the supporting evidence comes from peer-reviewed studies or credible medical sources, and verify their findings. You also need to verify that the "credible medical source" has

actually published such findings, i.e., authors sometimes lie about evidence or take it out of context, etc.

- **Cross-Reference**
    - Question: What do other sources say?
    - Is this information consistent with other sources?
    - Cross-check the information with other reputable sources.
    - Consistency across multiple credible sources generally indicates reliability.
    - Check other sources for a more balanced view and look for any missing context or information.
    - Example 1: You read an article from a left-leaning news outlet that discusses a new healthcare policy, highlighting its benefits for low-income families. To get a balanced view, find an article on the same topic from a right-leaning outlet. It might discuss potential drawbacks or financial implications, giving you a fuller picture of the policy.
    - Example 2: Read an article on a new health trend in a lifestyle magazine, then check if the same trend is supported by articles in scientific journals or websites like Mayo Clinic or WebMD.

III. Topics

In this track, each topic is a news article. Each assessor will be assigned 25 topics. The Google Drive Folder link provided to each assessor contains a copy of the assigned 25 Google Documents. The name of each document contains its Document ID. Each document contains the Document ID, its URL, and the plaintext version of the news article. Assessors do not need to judge the trustworthiness of the article.

IV. Task 1: Producing Questions

For each topic (news article), the assigned assessor should scrutinize the article in plaintext as provided in the Google Document and produce 10 questions that the reader should ask to evaluate its trustworthiness, ranked by their importance to the evaluation from the most important to the least important. You may take notes for yourself by leaving comments on the Google Document, but be careful not to accidentally delete or change any text in the original text of the article. If the article contains reader comments at the bottom of the page (there should be few if any of these), please ignore the reader comments and read only the main article to formulate your questions.

Your questions are not limited to those aspects mentioned in Section II, as long as you believe readers should think about those questions when evaluating the trustworthiness of the news article. Your questions should meet the following requirements.

Question Requirements:
- Should be self-contained and explain the full context, i.e., one can understand this question without reference to the article.
  Examples:

- ❌What are the details of the study mentioned in the article?
- ✅What are the details of Dr. Jane Doe's recent study on diet as published in the Health Journal, such as data and methods?
- Should be at most 120 characters long.
- Should be reasonably expected to be answered by a single web page.
  Examples:
  - ❌Are Jason Samenow, Ian Livingston, and Jeff Halverson experts in meteorology or related fields?
  - ✅Is Jason Samenow an expert in meteorology or related fields?
- Compound questions should be avoided. In general, each question should focus on a single topic.
  Examples:
  - ❌Who is Dr. Jane Doe and is she a recognized expert in health?
  - ✅Is Dr. Jane Doe a recognized expert in health?

For each article, the assigned assessor should submit their questions through this Google Form: https://forms.gle/ApNh46NPwNRMuoGy8.

Tip: We expect each topic to take you 30-45 minutes to read the article and write your prioritized list of 10 questions, including any time needed to search the Web for possible answers to your questions to help you determine the relative importance of the questions. You will probably want to write and edit your 10 questions in a text editor of your choice, and then copy and paste the questions into the Google Form only after you are satisfied with the questions and their ordering.

V. Task 1 Example

On February 21, 2023, the New York Times published an opinion article by Bret Stephens entitled "The Mask Mandates Did Nothing. Will Any Lessons Be Learned?". This Google Document contains the plaintext version of this article (also attached in the appendices of this document). Stephens makes an argument that mask mandates during the COVID pandemic did not work. Given the importance of this issue, the reader would be advised to examine the trustworthiness of the information.

As suggested by Lateral Reading, we want to ask about sources, evidence, and what others say about the issue. This example file (also attached in the appendices) shows the 10 questions we came up with to evaluate the trustworthiness of this article, based on its plaintext version in the Google Document. In working to answer these questions, the reader would likely learn that Stephens is a conservative, that Tom Jefferson had previously published articles using other studies as evidence against masks, which received criticism from other scientists, that Maryanne Demasi is a journalist who has faced criticism for reports that go against scientific consensus, e.g. Wi-Fi is dangerous, and that the Cochrane study was misinterpreted as it was inconclusive about the question of if interventions to encourage mask wearing worked or not.

VI. Task 2: Judging Questions from Participants

The track participants have developed methods to produce 10 questions for each article similar to your first task to support readers in evaluating the trustworthiness of online news. For this second task, you need to assess the helpfulness of their questions. You will be presented with a ranked list of 10 questions from a run each time. Your assessments should be on two aspects: "**quality**" and "**redundancy**".

Your "**quality**" assessments should be from the following grades.

- **Flawed [-1]**:
  - The question can not be used as an independent query without reference to the article, i.e., not self-contained (contextualized). Example: Does the article show any political or financial bias?
  - **OR** The question is not related to the article.
  - **OR** The question has other issues that are not listed above and therefore can not be comprehended.
- **Not Helpful [0]**:
  - The question does not help in evaluating the trustworthiness of the article, even if it is somewhat related.
  - The question may be compound or may not be expected to be answered by a single web page.
- **Okay [1]**:
  - The question provides helpful information for understanding topics in the article but is not crucial for the reader to judge trustworthiness. In other words, it might be helpful for some readers, but not important for most readers during their trustworthiness evaluation.
  - The question may be compound or may not be expected to be answered by a single web page.
- **Good [2]**:
  - While not critical, an answer to the question will enhance the reader's confidence in their judgment of the article's trustworthiness.
  - The question may be compound or may not be expected to be answered by a single web page.
- **Very Good [3]**:
  - The question addresses the core aspects of the article's trustworthiness. An answer to the question will be critical for the reader to form a reliable judgment about the article's trustworthiness. Its answer can potentially flip the reader's trustworthiness perception of the article.
  - The question may be compound or may not be expected to be answered by a single web page.
- **Excellent [4]**:
  - The question addresses the core aspects of the article's trustworthiness. An answer to the question will be critical for the reader to form a reliable judgment about the article's trustworthiness. Its answer can potentially flip the reader's trustworthiness perception of the article.

- ○ **AND** You could expect the question to be answered by a single web page, even if the question is compound. For example, "Where does Bret Baier work and what is his job?" could be expected to be answered in a single web page.

As you are judging a list of 10 questions from one run, "**redundancy**" refers to whether an answer to a question earlier in the list (within one run) largely suffices as an answer to this question.

- - **Yes [1]** or **No [0]**: For example, "What is the credibility of Sheera Frenkel in reporting on Middle Eastern conflicts?" and "Is Sheera Frenkel a reliable journalist with expertise in Middle Eastern conflicts?" are redundant.

VII. Task 3: Judging Documents Retrieved by Participants

The track participants have developed methods to retrieve documents from a specified web collection to answer some given questions to support readers' trustworthiness evaluation of online news. For this third task, you need to assess the **usefulness** of the retrieved documents to answer the question itself. This is similar to the evaluation of traditional ad-hoc retrieval tasks.

- - **Very Useful [2]**: The document is very useful for answering the question because it directly addresses the question with an explicit and complete answer or an answer can be derived easily from this document alone.
- - **Useful [1]**: The document is useful for answering the question because it contains useful information. Other documents might be needed to derive a reliable answer to the question, i.e., the document itself only contains a partial answer.
- - **Not Useful [0]**: The document is irrelevant to the question or useless for deriving an answer to the question. The document is also considered useless if it is not written in English, contains inappropriate contents (e.g., adult materials), or is unreadable.

VIII. Appendices

A. Example News Article from Bret Stephens

Document ID: clueweb22-fake-id
URL: https://www.nytimes.com/2023/02/21/opinion/do-mask-mandates-work.html

----

OPINION
BRET STEPHENS

The Mask Mandates Did Nothing. Will Any Lessons Be Learned?
Feb. 21, 2023

3.8K
Bret Stephens

By Bret Stephens

Opinion Columnist

The most rigorous and comprehensive analysis of scientific studies conducted on the efficacy of masks for reducing the spread of respiratory illnesses — including Covid-19 — was published late last month. Its conclusions, said Tom Jefferson, the Oxford epidemiologist who is its lead author, were unambiguous.

"There is just no evidence that they" — masks — "make any difference," he told the journalist Maryanne Demasi. "Full stop."

But, wait, hold on. What about N-95 masks, as opposed to lower-quality surgical or cloth masks?

"Makes no difference — none of it," said Jefferson.

What about the studies that initially persuaded policymakers to impose mask mandates?

"They were convinced by nonrandomized studies, flawed observational studies."

What about the utility of masks in conjunction with other preventive measures, such as hand hygiene, physical distancing or air filtration?

"There's no evidence that many of these things make any difference."

These observations don't come from just anywhere. Jefferson and 11 colleagues conducted the study for Cochrane, a British nonprofit that is widely considered the gold standard for its reviews of health care data. The conclusions were based on 78 randomized controlled trials, six of them during the Covid pandemic, with a total of 610,872 participants in multiple countries. And they track what has been widely observed in the United States: States with mask mandates fared no better against Covid than those without.

No study — or study of studies — is ever perfect. Science is never absolutely settled. What's more, the analysis does not prove that proper masks, properly worn, had no benefit at an individual level. People may have good personal reasons to wear masks, and they may have the discipline to wear them consistently. Their choices are their own.

But when it comes to the population-level benefits of masking, the verdict is in: Mask mandates were a bust. Those skeptics who were furiously mocked as cranks and occasionally censored as "misinformers" for opposing mandates were right. The mainstream experts and pundits who supported mandates were wrong. In a better world, it would behoove the latter group to acknowledge their error, along with its considerable physical, psychological, pedagogical and political costs.

Don't count on it. In congressional testimony this month, Rochelle Walensky, director of the Centers for Disease Control and Prevention, called into question the Cochrane analysis's reliance on a small number of Covid-specific randomized controlled trials and insisted that her agency's guidance on masking in schools wouldn't change. If she ever wonders why respect for the C.D.C. keeps falling, she could look to herself, and resign, and leave it to someone else to reorganize her agency.

That, too, probably won't happen: We no longer live in a culture in which resignation is seen as the honorable course for public officials who fail in their jobs.

But the costs go deeper. When people say they "trust the science," what they presumably mean is that science is rational, empirical, rigorous, receptive to new information, sensitive to competing concerns and risks. Also: humble, transparent, open to criticism, honest about what it doesn't know, willing to admit error.

The C.D.C.'s increasingly mindless adherence to its masking guidance is none of those things. It isn't merely undermining the trust it requires to operate as an effective public institution. It is turning itself into an unwitting accomplice to the genuine enemies of reason and science — conspiracy theorists and quack-cure peddlers — by so badly representing the values and practices that science is supposed to exemplify.

It also betrays the technocratic mind-set that has the unpleasant habit of assuming that nothing is ever wrong with the bureaucracy's well-laid plans — provided nobody gets in its way, nobody has a dissenting point of view, everyone does exactly what it asks, and for as long as officialdom demands. This is the mentality that once believed that China provided a highly successful model for pandemic response.

Yet there was never a chance that mask mandates in the United States would get anywhere close to 100 percent compliance or that people would or could wear masks in a way that would meaningfully reduce transmission. Part of the reason is specific to American habits and culture, part of it to constitutional limits on government power, part of it to human nature, part of it to competing social and economic necessities, part of it to the evolution of the virus itself.

But whatever the reason, mask mandates were a fool's errand from the start. They may have created a false sense of safety — and thus permission to resume semi-normal life. They did almost nothing to advance safety itself. The Cochrane report ought to be the final nail in this particular coffin.

There's a final lesson. The last justification for masks is that, even if they proved to be ineffective, they seemed like a relatively low-cost, intuitively effective way of doing something against the virus in the early days of the pandemic. But "do something" is not science, and it shouldn't have been public policy. And the people who had the courage to say as much

deserved to be listened to, not treated with contempt. They may not ever get the apology they deserve, but vindication ought to be enough.

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow The New York Times Opinion section on Facebook, Twitter (@NYTopinion) and Instagram.

Bret Stephens has been an Opinion columnist with The Times since April 2017. He won a Pulitzer Prize for commentary at The Wall Street Journal in 2013 and was previously editor in chief of The Jerusalem Post.

B. Example Questions for the Above Article

clueweb22-fake-id   groupPrefixRuntag   1   Are reviews by Cochrane, a British non-profit, a reliable source of health care data?

clueweb22-fake-id   groupPrefixRuntag   2   Did Cochrane, a British non-profit, publish a study in 2023 indicating that mask mandates are not effective for reducing the spread of respiratory illnesses — including Covid-19?

clueweb22-fake-id   groupPrefixRuntag   3   Could Tom Jefferson, the Oxford epidemiologist, be considered an expert on mask mandates and the spread of respiratory illnesses?

clueweb22-fake-id   groupPrefixRuntag   4   Could Rochelle Walensky, director of the Centers for Disease Control and Prevention, be considered an expert on mask mandates and the spread of respiratory illnesses?

clueweb22-fake-id   groupPrefixRuntag   5   What evidence is there that wearing masks can protect against respiratory illnesses — including Covid-19?

clueweb22-fake-id   groupPrefixRuntag   6   Are N-95 masks better than lower-quality surgical or cloth masks at protecting against respiratory illnesses — including Covid-19?

clueweb22-fake-id   groupPrefixRuntag   7   What is the guidance from the Center for Disease Control and Prevention on mask mandates in schools?

clueweb22-fake-id   groupPrefixRuntag   8   What are the political leanings of Bret Stephens, the New York Times opinion columnist?

clueweb22-fake-id   groupPrefixRuntag   9   What are the political leanings of the journalist Maryanne Demasi?

clueweb22-fake-id   groupPrefixRuntag   10   Does China provide a highly successful model for pandemic response?