scientific reports

Check for updates

OPEN

The impact of interventions against science disinformation in high school students

Carlo Martini[1,2]✉, Mara Floris[1], Piero Ronzani[3], Luca Ausili[1], Giulio Pennacchioni[1], Giorgia Adorno[1] & Folco Panizza[4]

This work studies the effectiveness of three educational interventions (Civic Online Reasoning (COR), Cognitive Biases (CB), and Inoculation (INOC)) in helping high school students identify science disinformation. We conducted the experiments in classrooms in Northern Italy, the study employed an ecological design with real-world stimuli and a digital cellphone-based platform, engaging 2,288 students. Participants evaluated Instagram posts containing scientific and pseudoscientific content before and after the interventions. While previous research shows the efficacy of these techniques in controlled environments, this study found no significant improvement in students' ability to discern accurate information. However, COR indirectly enhanced accuracy by promoting lateral reading and click restraint, albeit among a small subset. Conversely, the INOC approach increased generalised skepticism, leading to reduced trust in scientifically valid content. We observe that unlike in non-ecological environments, the classroom environment poses challenges, including distractions and varied engagement, highlighting the difficulty of scaling interventions from controlled to real-world settings. Despite these results, the study underscores the importance of refining educational strategies and adapting them to dynamic, real-world contexts to enhance digital literacy.

Empirical studies on the problem of disinformation among teenagers are scarce, even though directing research towards younger generations is potentially very effective. We have evidence from the literature that once information has taken hold, it is difficult to debunk and to cancel its effects[1,2]. Thus, working with young generations and making them more resilient to first exposure to disinformation should appeal to the age-old adage that prevention is better than the cure. Young people spend a considerable amount of time on the internet and social media. Rideout et al.[3] report an average screen use among teenagers (13-18-y.o.) between seven and eight hours, including between 30 and 60 min dedicated to browsing and between 70 and 90 min dedicated to social media use. The Pew Research Center reports that 95% of all teenagers in the United States own a cellphone and 45% report being online "almost constantly"[4]. Time spent online exposes them to fake news and disinformation, which have been of particular concern on social media[5]. Moreover, even though teenagers seem to be very confident and familiar with the use of social media networks, they do not seem to be very accurate when evaluating information online[6].

Our study advances the literature by exploring how various evidence-based interventions from misinformation research[7,8] could be scaled up and adapted to a highly ecological environment in three key ways: transitioning from online to classroom-based experiments, from computer to mobile platforms, and from controlled to real-world stimuli.

Different solutions have been proposed to prevent the young from falling into disinformation traps. Part of the solutions proposed are grounded on the idea that the skills required to approach high-school textbooks, which are obviously trustworthy and conceived to help students' learning, are not as efficient when used to evaluate items of information they find online, which are not always trustworthy, truth oriented, or conceived to help the reader getting accurate information about facts and events. This necessary shift from two different context-dependent attitudes has been widely studied in recent years. Chinn et al.[9] and Wineburg et al.[10], for instance, have designed a specific set of techniques, labelled Civic Online Reasoning, that are useful to make this shift as efficiently and effortlessly as possible for high school students. The Civic Online Reasoning framework has provided meaningful and important results among students in US high schools. The interventions by Wineburg et al.[10] are designed to teach participants how to use specific strategies and techniques, and proved to be effective in improving people's accuracy at detecting scientifically valid information[11].

[1]Vita Salute San Raffaele University, Milan, Italy. [2]University of Helsinki, Helsinki, Finland. [3]International Security and Development Center, Berlin, Germany. [4]IMT School for Advanced Studies Lucca, Lucca, Italy. ✉email: martini.carlo@hsr.it

Besides the work based on the Civic Online Reasoning program, there have been several attempts at improving children's and young adults' ability to distinguish between accurate and inaccurate information, and at doing "sourcing": Brante and Strømsø[12] review 13 studies about sourcing. Sourcing is defined as the ability to evaluate a source, and its related metadata, of an informational content. All the studies they analyse are experimental, and involve a control group; N varies from 40 to 418, although most studies are on the lower end; the interventions vary from 60 h to 30 min, with most interventions contained in 10 h or less. The outcomes reported from the studies under review are mixed for students in primary school, while in lower secondary schools the intervention groups seem to score better in a variety of measures, like passing sourcing tests[13], and identifying a variety of features, like claim validity and author credentials. The studies reviewed in Brante & Strømsø[12] do not all focus on direct evaluation of accuracy, possibly because it is an epistemological concept that can be difficult to define.

In this paper, we focus on the topic of science information and disinformation, as a specific subset of communication whose reports are based on findings from scientific and technical research. While there is no definitive characterization of science disinformation, we consider here the kind of disinformation that "adopts the mannerism of science in order to advocate antiscience"[14]. Science disinformation typically concerns topics like health, medicine, climate, and tends to undermine informed public policy and trust in science[15]. We aim to study interventions that can help high-school students' accuracy in discriminating between information that is accurate and based on scientific findings as opposed to information that is inaccurate either because it is based only on the appearance of science (i.e. pseudoscience) or entirely made up. Lilienfeld et al.[16] define pseudoscience as claims or practices presented as scientific despite lacking empirical support or adherence to the scientific method.

In our experiment we use ecological cues, that is, science information and disinformation that exists is the World Wide Web. We look for and select cues that are either based on science (valid cues) or based on pseudoscience (invalid cues). Scientific validity is the extent to which a claim or source accurately reflects established scientific evidence and reasoning[17]. We check for scientific validity, and, conversely, invalidity, based on a multi-criteria checklist developed in Martini and Andreoletti[18]. To our knowledge, few studies have specifically examined scientific disinformation and critical thinking in young populations. In addition, we test well-known evidence-based interventions taken from the misinformation literature[7,8]. We do this by adapting the approach used in online experiments by Panizza et al.[11] and Ronzani et al.[19] and we apply it here in a field experiment in high-school classrooms in the northern Italian regions of Piedmont and Lombardy. We create a simulated digital information environment in which science information and disinformation appear side to side, and high-school students are tasked with distinguishing between the two by giving a judgement of scientific accuracy on the information they view.

In this study, we operationalise and test three distinct mechanisms underlying susceptibility to misinformation. *Civic Online Reasoning* (COR)[20–24] promotes deliberate source evaluation through fact-checking strategies. The COR approach emphasizes source evaluation — i.e., assessing author credentials, institutional affiliations, and evidence quality — which has been shown to increase digital literacy and critical evaluation skills[25]. Cognitive Biases (CB) approach[8,26–31] focuses on countering intuitive but faulty reasoning via awareness of mental shortcuts. Targeting heuristics such as confirmation bias helps participants recognize and counteract their own cognitive shortcuts, which has been empirically linked to reduced acceptance of false headlines[32]. *Inoculation Framework* (INOC)[33] aims to preemptively build resistance by exposing students to misleading rhetorical tactics. Inoculation theory posits that brief exposure to weakened forms of misinformation, coupled with refutational pre-emphasis, can confer a 'mental antibody' effect, reducing susceptibility in later encounters[34].

We hypothesised that interventions based on these three approaches might be effective at improving students' accuracy, defined as the ability to correctly identify science information and disinformation. In particular, we formulate the following hypotheses:

H1: Students exposed to any of the three interventions (COR, CB, INOC) perform better (i.e., will be more accurate) than students in the control group at distinguishing scientifically valid or invalid social media content.

H1bis: H1bis (within-group persistence): For each intervention group (COR, CB, INOC), we hypothesise that any accuracy improvements observed after Lecture 1 will persist at follow-up (one or four weeks later). For pragmatic reasons, we fixed the follow-up times at one week and four weeks after the first in-class intervention. We included a hypothesis on follow-up in order to understand whether the potential effect of the three interventions would be limited to a one-shot or have at least medium-lasting effects[35]. Since also students in the control group received a delayed intervention at T1, no between-group comparison at follow-up is possible.

In summary, our intervention protocol compares three different boosting interventions with a control group (see experimental design). By "boosting" we mean interventions that target the competence that leads to a behaviour[36]: for example, if a student lacks skills at recognising disinformation (behaviour), we aim at improving those skills in order to change the behaviour. In our intervention we test students' digital critical thinking abilities, defined as their ability to accurately judge as scientifically valid an item of information that is based on accurate and methodologically sound research, and as invalid an item on information that is based on unsound science or pseudoscience, these terms are operationally defined based on Martini and Andreoletti[18]. We test a relatively large number of students (final $N = 2{,}288$) in a field experiment, where the experimenters deliver a series of two lectures per each treatment or control group. We use a sample of Italian students and compare performance retention on the accuracy test at one-week and three-weeks intervals. Finally we also compare students' performance with independent variables like trust in science and conspiracy belief, as both variables have been indicated as possible predictors of disinformation attitudes[37].

Method

Experimental procedure

In January 2023, we sent out emails to school principals and science teachers in the northern Italian provinces of Milan and Turin. We used email addresses that were publicly available through the high school websites. The emails presented our project and asked interested instructors to get in touch with us via email. Based on a preliminary assessment of the number of high school institutions (and therefore students) interested in the study, our target was to recruit ca. 1,000 students through their instructors. We received interest from several dozen schools and we recruited on a first-come-first-serve basis until we reached a sample of about 2,300 students. The sample of interested schools and teachers was self-selected; however, the sample of students was not, as almost all the students, whose instructors decided to opt in our study, gave their consent to participate in the study, and did not choose to opt out, as it was their right to do. We also sought the parents' consent to their children's participation. Most of the schools involved were located in the Northern Italian provinces of Turin and Milan (two of the most populous provinces in Italy); however, some teachers in other provinces heard about our project from their colleagues and asked us to let them participate in the project. We ultimately decided to give them the chance to participate and not to exclude them on the basis of geographical reasons, since these were dictated by convenience rather than methodological considerations.

Experimental design

Experimental design, data collection, and analyses were pre-registered on the open science framework (*osf.io/qkpb5*). Any amendment to the original protocol is presented alongside the original planning.

Experimental treatments

The experiment consisted of a between-subjects design with a follow-up lecture to measure any lasting effects of the interventions weeks apart. Classes included in the study were randomly assigned to three treatment groups and a control group. The three treatment groups were based on three approaches: COR - Civic Online reasoning[24], CB - the Cognitive Biases approach[26], and INOC - the Inoculation framework[33]. Participants assigned to each one of the treatment groups watched a video-recorded lecture based on an established strategy to contrast the spread of disinformation and misinformation.

In the Civic Online Reasoning treatment (COR), inspired by the work of Breakstone et al.[22], we showed a video in which the application of 'lateral reading' and 'click restraint' techniques was illustrated with the help of practical examples. The effectiveness of this type of intervention is well-known in the literature (see for instance Wineburg et al.[10], $d = 0.34$). The proposed COR techniques were based on the reading strategies that professional fact-checkers use. They consist of instructions on how to maximise one's ability to simultaneously compare several pieces of information on the same news item. For example, lateral-reading, as the name suggests, is precisely the activity of opening multiple tabs in the browser to perform the act of sourcing[12], that is checking a source's validity. Showing students examples of these activities should encourage them to compare multiple sources on the same topic. The Civic Online Reasoning approach (COR) is rather standard and very widespread as a technique for "sourcing", namely, for checking a source's credibility and epistemic validity, and it is thought to be crucial in empowering people to navigate the overcrowded and often deceptive online landscape, thus promoting digital literacy.

In the Inoculation treatment (INOC), the video explains how to recognize five different ways through which misinformative contents are commonly disguised, namely: Emotive content, Conspiracy, Impersonification, Discredit, and Trolling. The Inoculation approach (INOC) to disinformation involves preemptively exposing people to weakened forms of misleading arguments or misinformation, along with explanations of why they're false, to build resistance against future manipulation attempts, and it is thought to help develop critical thinking skills and immunity to deceptive tactics, thus reducing susceptibility to misinformation. The efficacy of the inoculation protocol against disinformation has been demonstrated through various studies, showing that it can effectively build resistance against misinformation by exposing individuals to weakened doses of misinformation techniques. Research has highlighted the long-term effectiveness of such inoculations, with findings indicating that these effects can remain stable for at least three months[35]. Inoculation against misinformation has been explored in the context of political debates, health misinformation, and even radical extremist messages, providing evidence of its broad applicability and potential in fostering resilience against various forms of misinformation[38,39].

In the Cognitive Bias treatment (CB) we review the main cognitive biases that influence the way we evaluate online information [see[40,41]]. Along with each theoretical explanation, practical examples were provided. The Cognitive Biases approach (CB) to disinformation focuses on how our inherent mental shortcuts and biases make us vulnerable to false information. By understanding and addressing these biases, such as confirmation bias or the illusion of truth effect, it is thought that we can better recognize and develop resilience against misinformation. Numerous empirical studies suggest that suboptimal decision-making arises not exclusively from erroneous beliefs but derives from cognitive biases[8]. Conversely, interventions targeting decision biases have demonstrated promising outcomes[8]. In light of this, we formulated an intervention designed to increase accuracy by improving the knowledge on cognitive biases. Improving awareness among participants regarding their potential cognitive biases would mitigate their effects, making the participants better at discerning false scientific claims.

The Intervention Groups received two lectures:

LECTURE 1: The experimenter introduced the first lecture and showed a pre-recorded video lesson (COR, CB or INOC). After the video, students took an online accuracy test that measured their level of accuracy in recognising science information and disinformation and we also collected data on the students' demographic data and online habits.

LECTURE 2: In the second lecture, students started by completing a follow-up accuracy test to measure their level of accuracy in recognising science information and disinformation after either one week or four weeks from the first lecture. This was meant to test the effectiveness of the previous intervention after a time. After the test, the students were then offered a second lesson and a debriefing about the examples of information and disinformation they had been exposed to while taking the online accuracy test.

The Control Group also received two lectures:

LECTURE 1: The experimenter introduced the first lecture and then the students took an online accuracy test to measure their level of accuracy in recognising science information and disinformation and we also collected data on the students' demographic data and online habits. After the test, the experimenter showed a pre-recorded video lecture (either COR, CB, or INOC). The lectures were shown to control group students to comply with the schools' request to guarantee all students access to the formative content. As all students were treated at the end of the first lecture, we were able to compare the long-term effects of the treatments only using a pre-post design.

LECTURE 2: The second lecture for the control group follows the exact same procedure as the other intervention groups. Students took a second test with different Instagram posts. In the original pre-registration protocol, students in the control condition would have first watched a second video lecture during the follow-up lecture before responding to the second test. This reverse order was intended to examine the combined effect of two video lectures on the recognition of scientifically valid posts. However, due to unforeseen complications that arose during data collection, the follow-up format was standardised with the other experimental conditions. After the second test, given the potentially beneficial educational value of our intervention, we decided to play a second pre-recorded 15 min video lecture after the test (COR, CB, or INOC see the section Manipulated Variables for details).

The sessions were structured as follows: The experimenter enters the classroom, briefly introduces the research team, tells a short story about a 'historical' fake news story about the Cyclops legend[42], and explains what we mean when we say that a piece of news is reliable and how we define misinformation. In this phase, all experimenters follow a pre-established script and show supporting slides in class (duration approx. 10 min). After the 10-minute introduction the experimenter projects a video containing one of the three interventions (COR, INOC, CB). After finishing the video, the experimenters proceeded to distribute to the class, via a QR code, the link to the test on the Qualtrics© software platform; the approximate completion time was 10 to 15 min. The purpose of the test was to assess students' ability to recognize the scientific validity of a series of Instagram posts. Given the potentially beneficial educational value of our intervention, and to comply with the schools' request to give all students access to the instructional content, we decided to ensure that the control group also had the opportunity to receive the information in a delayed intervention fashion. Classes assigned to the control group first took the test, then watched one of the three video lectures, randomly assigned. After a period of between one and four weeks, depending on the classes' availability, the experimenters returned to the classroom for a second intervention and to test the medium-term effects of the first intervention. Students took a second test with the same structure as the first one, but with different Instagram posts, at the beginning of the second lecture. The test was followed by either watching another one of the video lectures, or by a customised lesson on the same topics based on requests from students and teachers. At the end of the intervention, at the request of the class, the experimenters debriefed the news encountered during the two tests.

Video development

Each intervention consisted of a pre-recorded video, shown in class during the lectures. The mean duration of the videos was 18 min and 46 s, with COR being the longest, 20 min and 45 s, INOC the shortest, 17 min and 53 s, and CB the closest to the mean, with 18 min 34 s of duration. All videos are designed to keep the structure unchanged and minimise aesthetic differences. A split screen shows one of the team members explaining the topic, while showing supporting audio-visual material in the other half of the screen. In the case of INOC and CB, ad hoc slideshows were prepared. In the INOC's slides, we collected every form of disinformative content isolated in Van der Linden and Roozenbeek[33] and explained them with made up examples, to describe as clearly as possible the different examples of disinformation, and thus enable the students to recognise future cases of disinformation they may encounter. Similarly, for CB, every selected bias received its own section in the slides, with examples and questions. When questions appeared on the screen, either a little timer came out and the students had time to reply, or the member of our team who was physically present in class during the intervention stopped the video and asked for replies. These questions were conceived to favour the comprehension but also to make the experience more dynamic and engaging. In the case of COR's video, no slideshow was prepared because the activity displayed involved showing the lateral reading technique in real time. On the empty half of the screen, the one without one of the authors (C.M.) speaking, there was the homepage of a common browser (Google Chrome). The video consisted of the author doing various searches about specific news, this time a real one (a report on the current conditions of polar bears), showing to the students how to practise the two main techniques described by the COR approach, namely click-restraint and lateral-reading. The attempt was to give a practical demonstration of how these two actually work, and their efficiency during research about the accuracy of pieces of information found online.

Stimuli selection

The test to evaluate scientific validity was based on the information contained in snapshots of 16 fake Instagram posts. All posts were based on news items that the experimenters picked from the web, and were selected in order to avoid possible bias towards one or the other of the three interventions. Posts were selected and independently evaluated by pairs of researchers from the research team. Half of the posts were based on pseudoscientific claims and contained information that was shown to be unequivocally false or misleading by professional fact-

checking organisations. The remaining half of the posts contained scientifically accurate claims supported by published scientific evidence. Students observed three randomly selected posts in each evaluation, and sampling prevented the presentation of the same posts at follow-up. The English translation of the posts is included in the Supplementary Materials. Because we know that source familiarity is one of the major predictors of accuracy rating[43], we intentionally selected sources that we considered unfamiliar. This was confirmed by familiarity ratings of participants (Mean participants familiar with a source: 11%, standard deviation: 17%).

Measures

Scientific Validity: We ranked each post as scientifically valid or scientifically invalid based on a number of criteria identified in Martini & Andreoletti[18]. The criteria we chose were *authorship*, *pertinence*, *consensus*, and *publisher*. Each criterion raises a number of key questions, which two of the authors evaluated independently for each post and then compared their answers. Posts were selected only if both evaluators agreed. What follows is a brief explanation of the criteria used, adapted from Martini & Andreoletti[18]. Authorship: is the post based on the work of reputable scientists, indicated by their CV? Pertinence: is the expertise displayed by the sources of the post pertinent to the contents of the post itself? Consensus: are the contents of the post generally accepted in the scientific community or is there genuine disagreement on the matter? Publisher: does the evidence supporting the information in the post appear in reputable scientific publishers? Only if all of the above criteria were satisfied did the evaluators rank the post as scientifically valid, whereas posts that were ranked as invalid failed on at least three of the criteria. In sum, we used a set of rather strict conditions to judge posts as scientifically valid or invalid with a high degree of confidence, thus avoiding cases of "contested science".

Accuracy: For each post, students were asked to rate the scientific validity of its content on a scale from 1 ("completely invalid") to 6 ("completely valid"). These ratings were converted into accuracy scores based on the post's actual scientific validity. If a post was scientifically valid, a rating of 6 received the maximum score (1), while a rating of 1 received the minimum score (0). Conversely, if a post was invalid, a rating of 1 earned the maximum score (1), and a rating of 6 earned the minimum score (0). The scoring was conducted at the post level, resulting in three scores per lecture, per student, and a total of six scores per student for the whole experiment.

Exploratory Variables: In addition to the accuracy ratings, we examined several exploratory variables designed to shed light on the decision-making process that guided students' post evaluation. We measured participants' confidence in their rating by asking, "On a scale of 1 to 10, how confident are you in your answer?". To explore the extent of social media engagement, we collected data on "Average Sharing per Week", derived from participants' responses to "On average, how many articles, news, memes, opinions, or other posts did you share last week?". In addition, we examined source familiarity by asking "Did you know [name of source] before the experiment?" to determine participants' familiarity with the sources used. Source familiarity is an aspect proven to be crucial when studying these types of post-evaluations[11]. Source trustworthiness was rated subjectively on a 5-point Likert scale. To examine external search, participants indicated whether they left the test page (yes/no) and provided details of where they obtained external information. Trust in scientists was measured on a 6-point Likert scale adapted from the Edelman Trust Barometer Yearly online survey[44]; "In general, do you consider the information provided by the scientific community to be reliable?", from (1) "not at all" to (6) "A lot"), while the 'conspiracy belief scale' was assessed by a combination of 5-point Likert scales following the approach proposed in Bode & Vraga[45]. The survey also included questions about phone usage time (self assessed, on a slider from 0 to 24 hours) and a question that asked students whether they could check their real screen time and report it on a slider from 0 to 24 hours.

For the participants assigned to one of the three treatment groups, an attention check question was used to monitor attentiveness to the video. The question "Did we show you a video before you answered this questionnaire?" (Yes/No) was followed by video content identification, where participants had to correctly detect elements present in the video.

Participants

Data collection took place during the second half of the 2022/2023 school year. 19 institutions, including 9 higher education institutions and 6 high schools, participated in the study. A total of 100 classes from first to fifth grade were recruited, with around 2,590 students participating in the first lecture and around 2,464 in the follow-up lecture.

Of all the students, some did not start the survey or ended it before evaluating the instagram posts. Some students did not report their class, and this information (along with their assigned treatment) could not otherwise be retrieved. In addition, some students started the survey in the first lecture but completed it only in the follow-up lecture. After removing these participants, we were left with 2,288 responses in the first survey and 2,161 in the follow-up survey. Of these responses, we were able to match the responses of 1,525 students between the two sessions. Thus, 2,288 respondents were included in the analyses for the first lecture (control = 663, COR = 569, INOC = 521, CB = 535) and 1,525 respondents were included in the follow-up analysis (control = 413, COR = 396, INOC = 352, CB = 364).

Analyses

Here we describe the pre-registered analyses for each hypothesis. Any deviation from the pre-registration is reported alongside with an explanation. All tests adopt the standard 5% significance threshold to test against the null hypotheses. Post-hoc tests and multiple analyses are corrected for multiple comparisons using a Benjamini-Hochberg procedure[46]. In the case of missing data for one of the control variables, the remaining data are still included in the analyses whenever the statistical test allows. Following the preregistration, outliers are included in the analyses.

To test the first set of hypotheses H1/A, H1/B, and H1/C, we measured how accuracy scores differed by experimental treatment after the first lecture. Since accuracy scores were measured on a six-point scale, we treated them as an ordinal rather than a continuous variable. Thus, we ran a cumulative link mixed effects logistic regression (`clmm` function from the R package `ordinal`;[47]) with accuracy scores as the main dependent variable, experimental treatment as the main predictor variable, and random intercepts for Instagram post and nested random intercepts for student, class, and school. Compared to the pre-registration, we decided to include a random intercept for students, which we were not sure we could compute a priori due to the small number of items per subject. The results remained unchanged when conducting the original analysis (Supplementary Materials). We computed the contrasts of the three treatments against the control group using the `trt.vs.ctrl` contrast from the `emmeans` package[48].

In case of a significant effect of treatment, we proceeded to test the persistence of the effect in the follow-up lecture. Accuracy scores in the three treatments at the time of the follow-up lecture were compared to accuracy scores of students in the control group at the time of the first lecture. The model was the same as in testing H1, with the addition of the control variable lecture delay measuring the distance in weeks between the first and the follow-up lecture. Contrary to the pre-registered analysis, no interaction term could be tested as data from baseline always had delay zero.

The analyses were repeated adopting the two following pre-registered robustness checks: the inclusion of a random intercept for the experimenter involved in that class' data collection, and the exclusion of participants failing any of the attention checks included in the study (Supplementary Materials).

Results

Descriptive statistics

Of the participants included in the main analysis, 1,111 were female, 886 were male, and 291 did not report their gender or did not identify as either male or female. Demographic information was missing from the first session due to a coding error ($N = 81$). The remaining missing responses are due to non-completion of the experiment, as these questions were asked at the end of the experiment. Mean, median, and modal age were 16 years (range 13–25 years), 225 students did not specify their age. Since some classes were aggregated and the data was anonymized, it was not possible to determine how many students belonged to each grade.

Participants were shown a series of Instagram posts with scientific and pseudoscientific content and asked to indicate whether the content was scientifically valid, that is, based on accurate science information as opposed to inaccurate information (see Fig. 1), that is information based on the mere appearance of science. Baseline ratings measured in the control group varied widely across Instagram posts, especially between valid and invalid content: students rated invalid posts as invalid 77% of the time (range: 60–98%), whereas scientifically valid posts were rated as valid only 32% of the time (range: 15–58%), displaying a strong scepticism bias[49].

Participants were also asked about their average phone usage and which social media sites they use most frequently. The reported data shows that students spend an average of 5–6 h per day on their phones and that the most frequently used social media are Instagram (83%), Whatsapp (79%), TikTok (68%), BeReal (27%) and Snapchat (13%). Facebook is one of the least used social media, reported by only 3% of students (see Fig. 2).

Treatment balance

We check whether experimental treatments are balanced in terms of the main demographic characteristics and attrition rates. Demographics are balanced across treatments with the exception of gender and age: the proportion of male students is lower in the Cognitive Bias (35%) and COR (33%) treatments than in the Control (45%) and Inoculation (46%) treatments (Chi-squared test, $\chi^2(3) = 23.331$, $p < .001$). Similarly, pairwise comparisons of treatments reveal that COR and Cognitive Bias students have a relatively lower average age than Control and Inoculation treatments (pairwise comparisons contrasts in linear regression, all $p_{\text{corrected}} < 0.043$), although by a small margin (maximum difference between treatments: 5 months). Treatments also differ in the probability of completing the full questionnaire at lecture 1 and attrition rates at lecture 2: completion rates are 97% in Control and Inoculation treatments, versus 93% in COR and Cognitive Bias treatments ($\chi^2(3) = 19.715$, $p < .001$), whereas attrition rates are higher in the control group (38%) compared to the other treatments (30–32%; $\chi^2(3) = 8.543$, $p = .036$). Given these differences, we repeat the pre-registered analyses including covariates such as demographic variables, belief in conspiracy theories and trust in scientists, or completion rates of the first survey, and all results are consistent with the reported results below (see Supplementary Materials).

Treatment effects

Comparing the different treatments to the control group shows that none of the interventions significantly increased news discernment (all $p_{\text{corrected}} > 0.747$; Table 1; Fig. 3). In other words, the average effect of the video lectures did not significantly improve student's ability to recognize scientific false news. The result is robust to using a different metric for accuracy (i.e. categorizing responses as either correct or incorrect, all $p_{\text{corrected}} > 0.704$) or to exclude participants who failed one or more of the attention checks included in the survey (all $p_{\text{corrected}} > 0.618$).

Although the interventions' effect is non-significant overall, it is possible that they may be effective under certain conditions, for specific subsets of our sample, or for certain types of posts. Of particular relevance here is whether the interventions are effective among participants who have been successfully encouraged to adopt good fact-checking techniques, such as lateral reading and click restraint. We therefore examine heterogeneous treatment effects on the adoption of fact checking techniques, as well as other potential indirect effects that can be explained by differences between participants and Instagram posts.

Fig. 1. Illustrates a sample Instagram post used in the experiment, with reactions and statistics of the post obscured.Sourced from *scienzenotizie*, the post caption reads, "The Black Death in the Middle Ages left an indelible imprint on our DNA. A recent study reveals that the Black Death reshaped the evolution of the European population's genome."

Adoption of COR strategies

Since the survey asked participants whether they adopted one of the COR strategies (lateral reading, click restraint), it is possible to monitor whether the COR video lecture increased these behaviours and whether the use of these strategies increased news discernment. Since technique adoption was collected at the participant level, we compute the average accuracy for each participant and test whether the self-reported use of lateral reading and click restraint increase this average score (Figs. 4 and 5). Adopting the two techniques indeed increases accuracy (lateral reading: $\beta = 0.215$, $t(1217) = 3.502$, $p < .001$; click restraint: $\beta = 0.177$, $t(1217) = 2.396$, $p = .017$), confirming evidence from previous studies (Panizza et al., 2022; Wineburg et al., 2022). We then use a mediation analysis to test whether the COR video lecture indirectly increases accuracy by increasing the rate of these strategies (mediation package; Tingley et al., 2014). Analyses show a mediation effect of treatment for this lecture: compared to Control, lateral reading and click restraint are more frequent and increase accuracy (Average Causal Mediation Effects based on 1000 simulations; lateral reading: 5.2% [1.7%,10%], $p < .001$; click restraint: 4% [0.8%,8%], $p = .018$). In other words, a small but significant portion of the effect of the strategies is related to the exposure to the COR video lecture. At the same time, mediation analyses confirm that the total effect is non-significant (all $p > .054$), suggesting that the proportion of students adopting the strategies is still not large enough to ensure a positive effect of the intervention.

Scientific validity of posts

We also explore whether the interventions have a differential effect based on the scientific validity of the Instagram posts shown to participants. We repeat the pre-registered analysis adding as a predictor variable the scientific validity of the posts, and its interaction with the experimental treatment. Contrasts reveal a significant interaction of validity with the Inoculation treatment: when the posts are invalid, accuracy increases significantly compared to Control ($\beta = 0.250$ [0.050,0.449], $z = 3.000$, $p_{corrected} = 0.008$), whereas when posts are valid, accuracy decreases significantly compared to Control ($\beta = -0.239$ [−0.458,−0.019], $z = 2.602$, $p_{corrected} = 0.028$). These results suggest that respondents in the inoculation treatment tend to rate both valid and invalid posts as invalid more frequently than in the control condition. We thus test this hypothesis by looking at a regression with scientific validity (the original rating scale) as the predicted variable. The contrast between Inoculation and

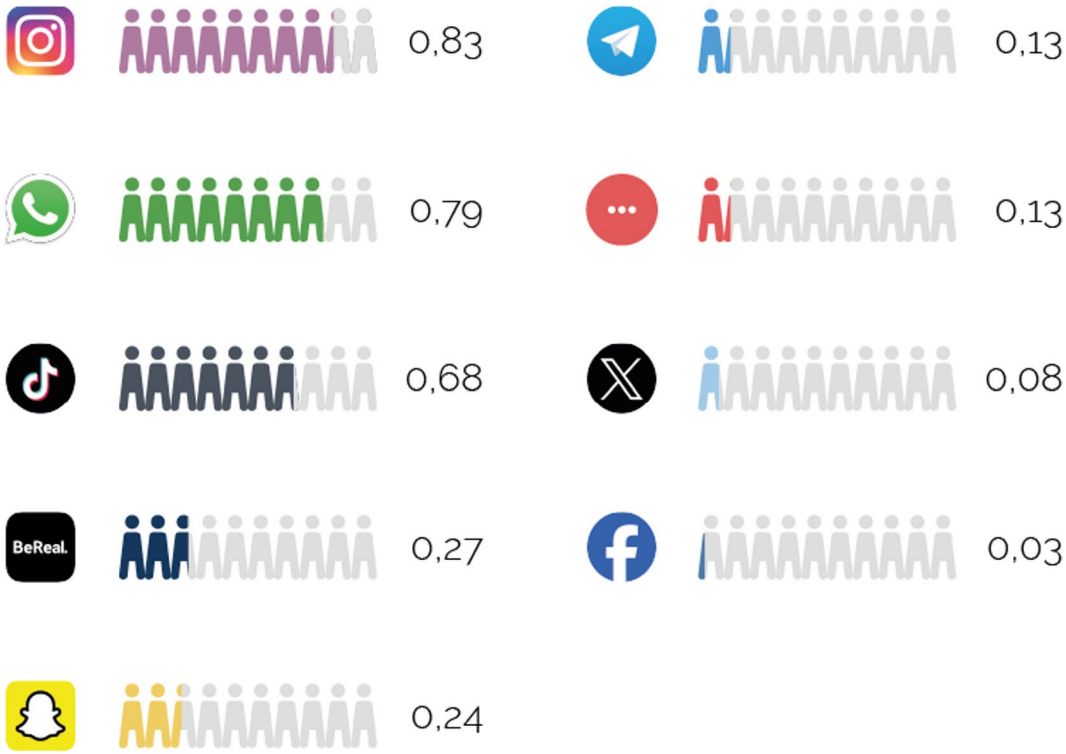

Fig. 2. Depicts the proportions of social network usage among the students who participated in the experiment. From the top left: Instagram, Whatsapp, TikTok, BeReal, Snapchat, Telegram, the three-dotted red circle indicates all other social media, X, and Facebook.

Intervention	Estimated effect [95% C.I.]	Z statistic	*p* value
CB	−0.0196 [−0.165, 0.126]	−0.322	0.747
COR	0.0279 [−0.116, 0.172]	0.465	0.747
INOC	0.0283 [−0.119, 0.176]	0.460	0.747

Table 1. Average treatment effect of the three video lectures on the recognition of scientific misinformation (p values corrected for multiple comparisons).

Control is indeed significant. ($\beta = -0.279$ [−0.485,−0.074], $z = 3.257$, $p_{corrected} = 0.003$): the video lecture about Inoculation appears to further increase the skepticism compared to an already skeptic baseline.

Group size

Since the test was conducted during school time, responses might be influenced by the size of the classes (or groups of classes) tested. We thus repeated the pre-registered analysis this time including as a predictor variable group size and its interaction with experimental treatment. Group size appears to interact with the COR treatment: as class size shrinks, average accuracy increases ($\beta = -0.015$, $z = -2.971$, $p_{corrected} = 0.003$). When clustering participants based on group size, those who were in groups smaller than 25 students had an average accuracy score of 4.07 [3.93, 4.22] on a 1 to 6 scale, whereas those who were in groups of 25 or more had an average accuracy score of 3.80 [3.71, 3.90], a 5.4% difference. Despite this significant interaction, we find no association of group size with attention checks (linear regression, all $p > .775$) nor with the adoption of COR strategies (all $p > .067$), suggesting that if class size is indeed influencing the effectiveness of the COR lecture, it is not through increased attention nor through increased adoption of COR strategies.

Follow-up lecture

Since we found no significant effect in lecture 1, it was not possible to establish any lasting effect of the interventions in the follow-up lecture. Despite this, we test whether the results from the exploratory analyses are still significant weeks apart. The tests suggest that the effects are no more significant (all $p > .734$).

Discussion

The problem of disinformation has been at the forefront of much public discussion in the past decade, with media organisations and policy makers decrying the problem and seeking solutions at the institutional and

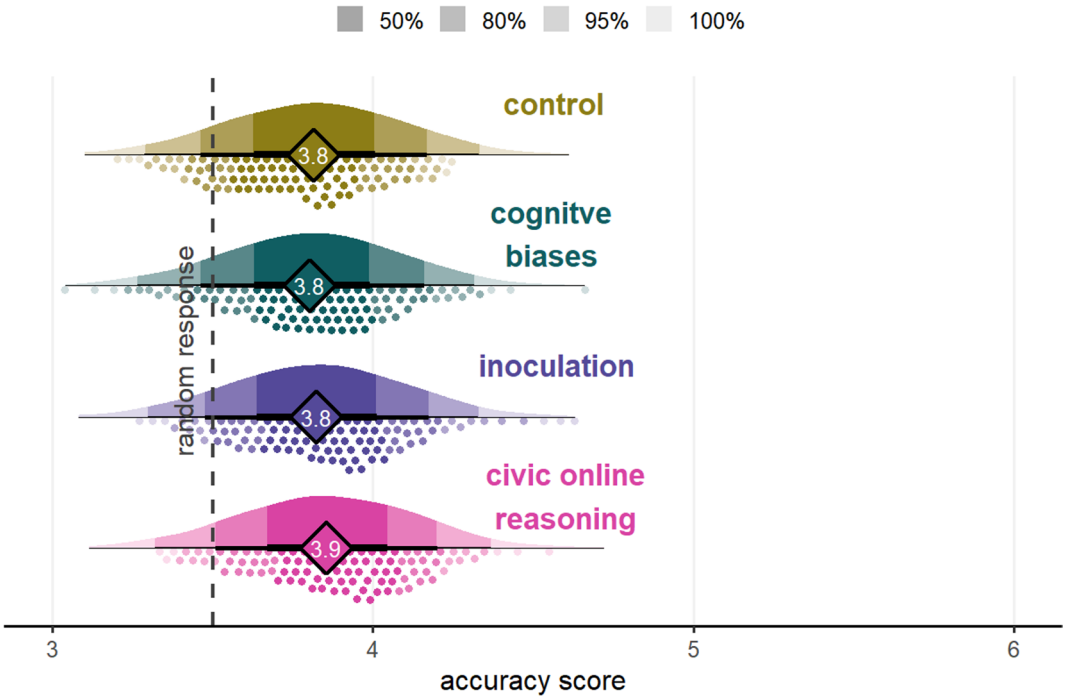

Fig. 3. Raincloud plot estimating the average accuracy score across conditions (min. 1, max. 6, random response: 3.5). The density curve on top represents the distribution of bootstrapped estimates of the average accuracy in each condition, whereas the bottom density represents the same distribution but discretized in 100 points. Shades represent different confidence intervals.

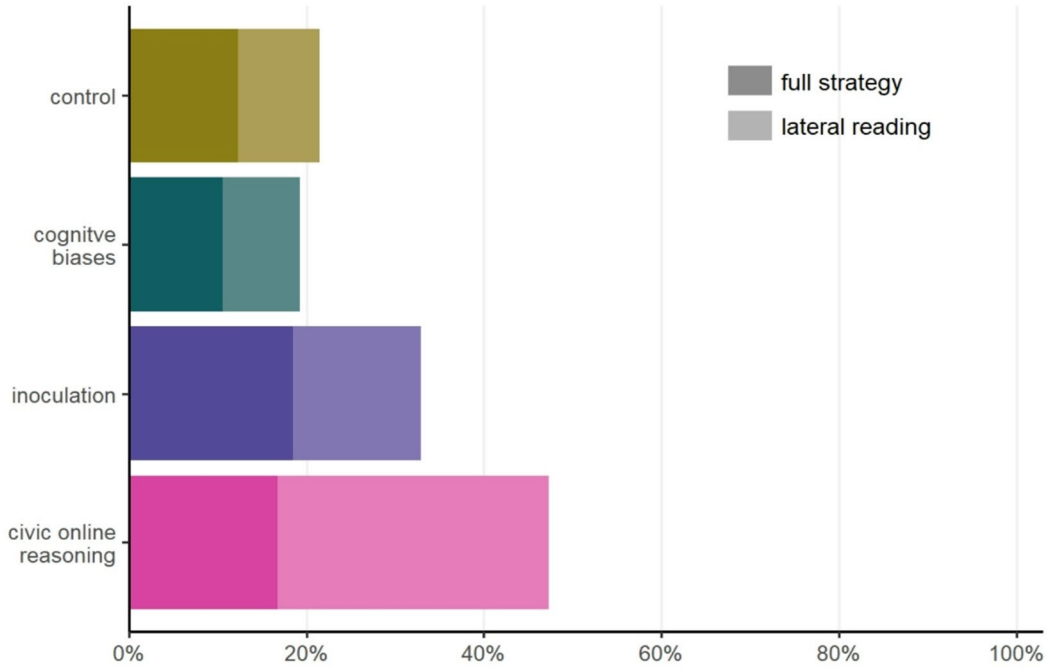

Fig. 4. Frequency of adoption of COR strategies, by experimental group. Lateral reading: participants who reported using lateral reading; full strategy: participants who reported using both lateral reading and click restraint.

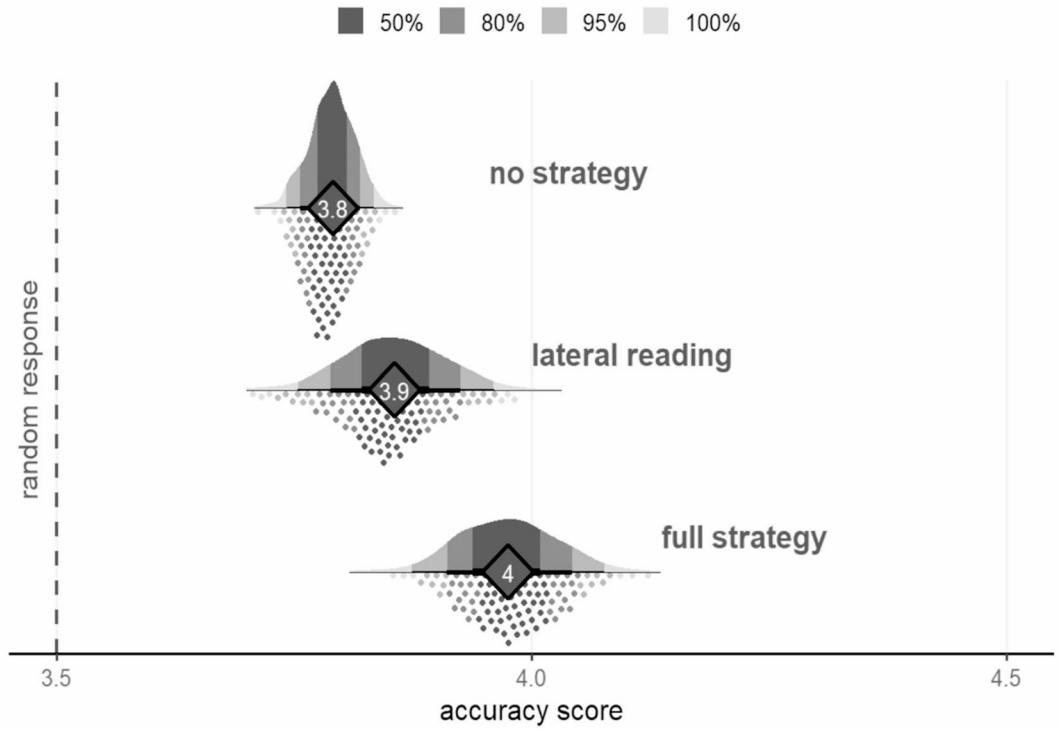

Fig. 5. Raincloud plot estimating the average accuracy per participants based on their adoption of COR techniques (min. 1, max. 6, random response: 3.5). No strategy: participants who reported not using any COR strategy; lateral reading: participants who reported using lateral reading; full strategy: participants who reported using both lateral reading and click restraint. The density curve on top represents the distribution of bootstrapped estimates of the average accuracy, whereas the bottom density represents the same distribution but discretized in 100 points. Shades represent different confidence intervals.

technological level. The solutions typically include: (a) use and support of fact-checking, including automated fact-checking[50] and human-powered organisations or groups of journalists tasked with assessing the accuracy of news stories and claims[51]; (b) holding social media platforms accountable for malicious and false content[52], often through regulatory measures[9,53]; (c) promoting moderation practices; (d) proposing counter narratives and debunking, possibly through the institution of independent communication centres, working through international collaborations[54] .

The proposed solution we focus on in our paper relies on media literacy education, the idea that we can promote critical thinking skills, and we can do so from a young age. Our study used a novel approach running interventions directly in the classrooms, and reaching a relatively high number of students compared to the studies reviewed that used direct interventions. We also use a multi-theoretical approach by comparing different interventions that are often proposed in the literature: the Civic Online Reasoning techniques, the Cognitive Biases program, and the Inoculation Approach. Our study used an ecological experimental design: We ran interventions in the classroom, during school hours, using instagram, one of the most used social media among teenagers, as the template platform where we simulated a digital environment. In addition, we used experimental stimuli (examples of true and false news) that we found online, rather than creating them ad hoc for the experiment. We think that this approach is highly ecological and ought to be used in future repetitions of studies on boosting people's critical thinking and sourcing skills.

The results from our study reveal the difficulties with an ecological approach targeting critical thinking and reading competences, but also potential avenues for improvement. We intentionally decided to focus on science disinformation, which may arguably be considered a rather difficult type of disinformation to recognise and debunk[18]. We started with the hypothesis that students receiving the interventions (COR, CB, and INOC) ought to think and read online news more critically, but none of the interventions significantly increased news discernment, and the result is robust using different metrics for accuracy. In other words, our video lectures did not significantly improve student's ability to recognize scientific false news. Moreover, the follow up lectures did not yield any significant effect.

In general we observe that the interventions we used have previously worked in other contexts[11,33], usually online experiments, which is typically the favourite medium on which most of the literature relies. Nonetheless, our results show that those interventions cannot be easily translated to the classroom settings. This could be interpreted in various ways. Moving from a laptop platform, which was used, for example in[25], to a mobile platform, like we used in our classroom experiments, could lead to additional challenges and could affect the ability of students to do lateral reading, one of the main strategies that are proven to improve accuracy assessing content validity.

We should also note that 10% of the respondents did not correctly recall one of the examples made in the video lecture that was previously presented to them, thus failing the first attention check. And 15.8% of the respondents failed at least one of the additional two attention checks we included in the test, thus denoting scarce attention, and possibly a distracting classroom environment. Future work should try to mitigate potential sources of distraction in the classroom, for instance by working with smaller groups of students, in order to focus on the potential effects of the interventions.

We also know from informal interactions with the researchers who did the classroom interventions that there was extreme variability in the level of attention that students had in class, and that classroom management was at times challenging; moreover, across all classes attention failure (i.e. students failing at least one of the three attention checks) averaged 23.3%. The previous considerations suggest that in order to go from lab (e.g. online laboratory experiment) to the field (e.g. the classrooms where we ran the interventions) we need to adapt materials and stimuli and make them more effective, and possibly more interactive. This is an important lesson about the scalability of online interventions: Since it is likely that any large-scale digital critical thinking intervention would be carried out during curricular hours, the interventions need to be tailored to their natural environment, namely the classroom.

Despite those difficulties, in our intervention we confirm the indirect positive effect of the COR techniques, in support of previous literature[10,11]. We also observe that the Inoculation approach increases generalised scepticism in students, leading to an undesired side effect from the interventions: instead of increasing critical thinking we end up increasing the tendency to consider a news' item false[49].

Limitations and future work

Our study tried to create an ecological choice environment: (1) from artificial to natural stimuli, thus using real true and false news we collected online; (2) from labs and online labs to classrooms; (3) from computers platforms to mobile ones, which is the place where most young people consume news. These have advantages, like adherence to reality and high external validity, but also impose some difficult tradeoffs.

The school environment implies risks of spillovers, that is, students copying each other. It also makes it hard to monitor individual behaviour and provides little incentives for performance. It is also possible that the use of video lectures, chosen in order to minimise experimenter's effect in the classroom, made the interventions less engaging. In other words, it is possible that a video lecture is less effective than more interactive lecture and practice sessions, where students could practise, for example, the COR techniques, and thus "learn by doing". All of the above could explain the high level of noise we observed, that is, the fact that a high proportion of students didn't pass the attention checks.

Nonetheless, despite the noise generated by the ecological environment we used, our analysis shows that even after excluding the subjects who failed the attention checks there were no significant direct effects of the interventions. This excludes the possibility that noise and ecological conditions alone were the only drivers of the failure to confirm the experimental hypotheses.

Despite our effort to select posts from a wide variety of topics, it is then possible that the selection of stimuli was not fully representative of the natural information environment the students typically find themselves in. It would be advisable for future studies to use the direct feed that subjects encounter in their own digital environment, selected thus from students' news diet according to their truth value and thematics to ensure the content was science-based.

Further, in order to ensure standardization across a large, multi-site sample of the content delivered to the classrooms, our interventions consisted of brief, passive videos, and lack of interactivity is likely to have reduced the impact of interventions. While passive interventions in online experiments have shown positive results in past work[11], it is possible that in classroom settings interactivity is needed in order to maximise engagement. Future work should strive to maximise interactivity and future iterations of the work, possibly through gamified interventions[55], should carefully evaluate the tradeoff between standardization of content delivery and interaction.

A potential risk associated with our interventions, which may explain the absence of a direct effect on accuracy, is also an increase in students' scepticism, particularly resulting in lower ratings for true content. This is evident looking at the average validity ratings of true content posts that have been all evaluated as more invalid than valid. In all treatments, between 66 and 72% of all respondents rated scientifically valid posts as invalid, indicating a strong scepticism bias[56]. Another element that may have contributed to this generalised scepticism is the fact that, in order to control for potential influences of the source, we selected sources that participants were mostly unfamiliar with. A recent systematic meta analysis[43] confirms that source is used as a cue for veracity and excluding it might have played in favour of high rates of scepticism.

It is also possible that the choice to focus on scientific information and disinformation may have influenced the mostly null findings. This is because scientific literacy among students is generally low and may have limited the effectiveness of the interventions. While our intervention strategies could succeed in domains such as political or social misinformation — where background knowledge requirements are lower — they may not compensate for limited understanding of scientific principles. This hypothesis could be explored in future studies by measuring baseline scientific reasoning — e.g., through standardized tests or STEM grades — to assess how prior knowledge interacts with intervention success.

Lastly, our study was conducted in a limited geographical area, the Northern Italian provinces of Milan and Turin, and broader-reaching experiments would be welcome. Clearly, the practical limitations and costs of directly reaching a high number of high school students implies that there are structural limitations to these kinds of interventions.

Conclusion

Our study aimed to assess the effectiveness of three distinct interventions — Civic Online Reasoning (COR), Cognitive Biases (CB), and Inoculation (INOC) — in increasing high school students' ability to identify science disinformation. Despite the ecological validity of our experimental design, which included real-world stimuli and a mobile platform, to simulate students' natural digital environments, none of the interventions demonstrated a significant overall improvement in students' ability to discern scientifically valid information from invalid information.

Several factors may have contributed to these findings. The transition from controlled, online experimental settings to the ecological, dynamic, and potentially distracting classroom environment posed challenges. Additionally, the video lecture format, chosen to minimise experimenter bias, may have reduced the engagement necessary for effective learning, compared to more interactive but less controllable methods.

Nevertheless, our study revealed important insights. The COR intervention indirectly improved accuracy through increased adoption of lateral reading and click restraint strategies, albeit in a small subset of students. The finding supports previous literature on the efficacy of these techniques. Conversely, the INOC intervention inadvertently increased general scepticism among students, leading to lower accuracy ratings for both valid and invalid posts. This suggests that while inoculation against misinformation can build resistance, it may also foster an overly critical attitude that undermines trust in legitimate science communication.

Our study shows that interventions that were shown to be effective in relatively less ecological experiments may face scalability challenges in highly ecological environments. It also underscores the complexity of contrasting science disinformation among young people and highlights the need for more engaging, interactive, and contextually adapted educational interventions. Future research should explore alternative methods and more diverse settings to better understand how to effectively enhance digital critical thinking and digital literacy skills in classrooms.

While our interventions did not yield the anticipated improvements in students' ability to identify scientific misinformation, they provided valuable lessons for designing more effective educational strategies, as well as suggesting novel and ecological experimental designs to conduct experimental studies on disinformation. By refining these approaches and addressing the challenges identified in this study, we can hope to equip young people with the critical thinking skills necessary to navigate an increasingly complex digital world.

Data availability

Data and analysis scripts are available at the online OSF repository: https://osf.io/qkpb5.

Received: 5 December 2024; Accepted: 18 August 2025
Published online: 01 October 2025

References

1. Johnson, H. M. & Seifert, C. M. Sources of the continued influence effect: when misinformation in memory affects later inferences. *J. Exp. Psychol.: Learn. Memory Cogn.* **20** (6), 1420–1436 (1994).
2. Chan, M. P. S., Jones, C. R., Jamieson, H., Albarracín, D., Debunking & K. and A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychol. Sci.* **28** (11), 1531–1546 (2017).
3. Rideout, V. & Robb, M. B. The common sense census: media use by tweens and teens, *Common Sense Media* (2019). https://www.commonsensemedia.org/sites/default/files/research/report/2019-census-8-to-18-full-report-updated.pdf.
4. Vogels, E. A., Gelles-Watnick, R., Massarat, N. & Teens social media and technology 2022. *Policy Commons* (2022, accessed 27 May 2025). https://policycommons.net/artifacts/2644169/teens-social-media-and-technology-2022/3667002/.
5. Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A. & Petersen, M. B. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *Am. Polit. Sci. Rev.* **115** (3), 999–1105 (2021).
6. Dissen, A., Qadiri, Q. & Middleton, C. J. I read it online: understanding how undergraduate students assess the accuracy of online sources of health information. *Am. J. Lifestyle Med.* **16** (5), 641–654 (2022).
7. Kozyreva, A. et al. Toolbox of individual-level interventions against online misinformation. *Nat. Hum. Behav.* **8**, 1044–1052 (2024).
8. French, A. M., Storey, V. C. & Wallace, L. The impact of cognitive biases on the believability of fake news. *Eur. J. Inform. Syst.* **34** (1), 72–93 (2023).
9. Chinn, C. A., Barzilai, S. & Duncan, R. G. Education for a post-truth world: new directions for research and practice. *Educ. Res.* **51** (1), 51–60 (2020).
10. Wineburg, S., Breakstone, J., McGrew, S., Smith, M. D. & Ortega, T. Lateral reading on the open internet: a district-wide field study in high school government classes. *J. Educ. Psychol.* **114** (5), 893–909 (2022).
11. Panizza, F. et al. Lateral reading and monetary incentives to spot disinformation about science. *Sci. Rep.* **12** (1), 5678 (2022).
12. Brante, E. W. & Strømsø, H. I. Sourcing in text comprehension: a review of interventions targeting sourcing skills. *Educ. Psychol. Rev.* **30** (3), 773–799 (2018).
13. Britt, M. A. & Aglinskas, C. Improving students' ability to identify and use source information. *Cogn. Instr.* **20** (4), 485–522 (2002).
14. Dornan, C. Science disinformation in a time of pandemic. *Public Policy Forum* (2020, accessed 27 May 2025). https://ppforum.ca/wp-content/uploads/2020/06/ScienceDisinformation-PPF-June2020-EN.pdf.
15. Kahan, D. Fixing the communications failure. *Nature* **463**, 296–297 (2010).
16. Lilienfeld, S. O., Ammirati, R. & David, M. Distinguishing science from pseudoscience in school psychology: science and scientific thinking as safeguards against human error. *J. Sch. Psychol.* **50** (1), 7–36 (2012).
17. Oreskes, N. *Why Trust Science??* (Princeton University Press, 2021).
18. Martini, C. & Andreoletti, M. Genuine versus bogus scientific controversies: the case of Statins. *Hist. Philos. Life Sci.* **43**, 1–23 (2021).
19. Ronzani, P., Panizza, F., Morisseau, T., Mattavelli, S. & Martini, C. How different incentives reduce scientific misinformation online. *Harvard Kennedy School Misinf. Rev.* **5**, 1 (2024).
20. Baer, A. & Kipnis, D. G. Teaching lateral reading with an online tutorial: preliminary study findings. *Libraries Scholarship, vol. 27* (2020, accessed 27 May 2025). https://rdw.rowan.edu/lib_scholarship/27.
21. Brodsky, J. E. et al. Improving college students' fact-checking strategies through lateral reading instruction in a general education civics course. *Cogn. Res.: Principles Impl.* **6**, 1–18 (2021).

22. Breakstone, J. et al. Lateral reading: college students learn to critically evaluate internet sources in an online course. *Harv. Kennedy School Misinf. Rev.* **2** (1), 1–17 (2021).
23. Breakstone, J., Smith, M., Ziv, N. & Wineburg, S. Civic Preparation for the digital age: how college students evaluate online sources about social and political issues. *J. High. Educ.* **93** (7), 963–988 (2022).
24. Wineburg, S., Mcgrew, S. L. & Reading Reading less and learning more when evaluating digital information. *Teachers Coll. Record* **121**, 1452 (2019).
25. McGrew, S. & Breakstone, J. Civic online reasoning across the curriculum. *AERA Open.* **9**, 1–14 (2023).
26. Abrami, P. C. et al. Strategies for teaching students to think critically: a meta-analysis. *Rev. Educ. Res.* **85** (2), 275–314 (2015).
27. Beaulac, G. & Kenyon, T. Critical thinking education and debiasing. *Informal Log.* **34** (4), 341 (2014).
28. Marin, H. Pedagogy for developing critical thinking in adolescents: explicit instruction produces greatest gains. *Think. Skills Creativity*. **6** (1), 1–13 (2011).
29. Mehta, S. R. & Al-Mahrooqi, R. Can thinking be taught? Linking critical thinking and writing in an EFL context. *RELC J.* **46** (1), 23–36 (2014).
30. Niu, L., Behar-Horenstein, L. S. & Garvan, C. W. Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educ. Res. Rev.* **9**, 114–128 (2013).
31. Van Peppen, L. M. et al. Effects of self-explaining on learning and transfer of critical thinking skills. *Front. Educ.* **3**, 100 (2018).
32. Piksa, M. et al. The impact of confirmation bias awareness on mitigating susceptibility to misinformation. *Front. Public. Health*. **15**, 12 (2024).
33. Van der Linden, S. & Roozenbeek, J. Psychological inoculation against fake news. In *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation* (eds. Greifeneder, R. et al.) (Routledge, 2020).
34. Banas, J. A. & Rains, S. A. A meta-analysis of research in inoculation theory. *Commun. Monogr.* **77**, 281–311 (2010).
35. Maertens, R., Roozenbeek, J., Basol, M. & van der Linden Long-term effectiveness of inoculation against misinformation. *J. Exp. Psychol.: Appl.* **27**, 1 (2021).
36. Espinosa, V. I., Wang, W. H. & de Soto, H. Principles of nudging and boosting. *Sustain.: Sci. Pract. Policy*. **14** (4), 2145 (2022).
37. Agley, J. & Xiao, Y. Misinformation about COVID-19 and trust in science. *BMC Public. Health*. **21**, 1–12 (2021).
38. Traberg, C. S., Roozenbeek, J. & van der Linden Psychological inoculation against misinformation. *Annals Am. Acad. Political Social Sci.* **700**, 136–151 (2022).
39. Boman, C. D. Protecting against disinformation. *J. Public. Relations Res.* **35**, 162–181 (2023).
40. Lau, A. Y. & Coiera, E. W. Cognitive biases in information search. *J. Am. Med. Inform. Assoc.* **14**, 599–608 (2007).
41. Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021).
42. Paglieri, F. Empedocle, i ciclopi e Gli elefanti. *Sistemi Intell.* **29**, 655–680 (2017).
43. Sultan, M. et al. Susceptibility to online misinformation: a systematic meta-analysis of demographic and psychological factors. *Proc. Natl. Acad. Sci.* **121**, 47 (2024).
44. Edelman. *Edelman trust barometer* (2022, accessed 27 May 2025). https://www.edelman.com/trust/2022-trust-barometer.
45. Bode, L. & Vraga, E. K. See something, say something. *Health Commun.* **33**, 1131–1140 (2018).
46. Benjamini, Y. & Hochberg, Y. Multiple hypotheses testing with weights. *Scand. J. Stat.* **24**, 407–418 (1997).
47. Christensen, R. H. Regression models for ordinal data. *CRAN Repository.* (2022, accessed 28 May 2025). https://cran.r-project.org/web/packages/ordinal/ordinal.pdf.
48. Lenth, R. Package 'emmeans': Estimated Marginal Means, aka Least-Squares Means *CRAN Repository* (2023, accessed 28 May 2025). https://cran.r-project.org/web/packages/emmeans/emmeans.pdf.
49. Van der Meer, T. G., Hameleers, M. & Ohme, J. Fighting misinformation and spillover effects. *Journal. Stud.* **24**, 803–823 (2023).
50. Guo, Z., Schlichtkrull, M. & Vlachos, A. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguist.* **10**, 178–206 (2022).
51. Mena, P. Principles and boundaries of fact-checking. *Journal. Pract.* **13**, 657–672 (2019).
52. Saurwein, F. & Spencer-Smith, C. Combating disinformation on social media. *Digit. Journal.* **8**, 820–841 (2020).
53. Tan, C. Regulating false news on Google. *Comput. Law Secur. Rev.* **46**, 105738 (2022).
54. ALLEA. Fact or fake? Tackling science disinformation. ALLEA Discussion Paper 5, Berlin (2021, accessed 27 May 2025). https://allea.org/wp-content/uploads/2021/04/Fact-or-Fake-Discussion-Paper.pdf.
55. Axelsson, C. A. W., Nygren, T., Roozenbeek, J. & van der Linden S. Bad news in the civics classroom. *J. Res. Technol. Educ.* **2024**, 1–27 (2024).
56. Acerbi, A. & Altay, S. Mercier. H. Fighting misinformation or fighting for information? *Harvard Kennedy School Misinf. Rev.* **3**, 1 (2022).

Author contributions

C.M., M.F. conceived the work and the design of the experiment. C.M. and L.A. created the interventions. G.P., L.A., G.A. M.F. run the experimental interventions and collected the data. P.R. and F.P. helped designing the experiment and analysed the data. C.M., M.F., P.R. and F.P. wrote the article and interpreted the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-16565-6.

Correspondence and requests for materials should be addressed to C.M.