Supplementary information

How do the UK public interpret COVID-19 test results? Comparing the impact of official information about results and reliability used in the UK, USA and New Zealand: a randomised controlled trial

Gabriel Recchia, Claudia R. Schneider, Alexandra L. J. Freeman

Main variables measured: Main study

Question	Response options
How easy or difficult did you find the information above to understand? How completely do you feel you understood the information above? How much effort did you feel you had to put into understanding the information above?	7-point Likert scale with the ends labelled "None" and "A lot"
What is your best guess as to the percent chance that John actually had COVID-19 at the time of his test, given his test result?	Slider scale with 10% markings with the ends labelled "0% = John definitely **didn't** have COVID-19" and "100% = John definitely **did** have COVID-19" Recorded as an integer between 0-100
How sure or unsure are you about your answer above, regarding the percent chance that John actually had COVID-19?	Slider scale with the ends labelled "very unsure" and "very sure" Recorded as an integer between 0-100
How high or low do you think the quality of the evidence behind John's test result is?	Slider scale with the ends labelled "very low" and "very high" Recorded as an integer between 0-100
To what extent do you think that the result from John's test was certain or uncertain?	Slider scale with the ends labelled "very uncertain" and "very certain" Recorded as an integer between 0-100
To what extent do you think that John's test result was… Accurate? Reliable? Trustworthy?	7-point Likert scale with the ends labelled "Not at all" and "Very much"
Given John's test result, how much do you agree or disagree with the statement that "John should now isolate himself from other people"?	Slider scale with the ends labelled "completely disagree" and "completely agree" Recorded as an integer between 0-100
What are your reasons for your answer above?	Free text
To what extent do you think the people who are responsible for communicating the results of the COVID-19 swab tests are trustworthy?	7-point Likert scale with the ends labelled "Not at all" and "Very much"

To what extent do you think policies and restrictions, such as rules for being allowed to enter a country, should be based on the results of a swab test?	7-point Likert scale with the ends labelled “Not at all” and “Very much”

Stimulus 1 (United States CDC)

COVID-19 Testing Overview

Two kinds of tests are available for COVID-19: viral tests and antibody tests.

• A viral test tells you if you have a current infection.
• An antibody test might tell you if you had a past infection.

Considerations for who should get tested

• People who have symptoms of COVID-19
• People who have had close contact (within 6 feet of an infected person for at least 15 minutes) with someone with confirmed COVID-19
• People who have been asked or referred to get testing by their healthcare provider, local or state health department

Not everyone needs to be tested. If you do get tested, you should self-quarantine/isolate at home pending test results and follow the advice of your health care provider or a public health professional.

How to get tested for current COVID-19 infection

• You can visit your state or local health department's website to look for the latest local information on testing.
• If you have symptoms of COVID-19 and want to get tested, call your healthcare provider first.

Results

• If you test positive, know what protective steps to take to prevent others from getting sick.
• If you test negative, you probably were not infected at the time your sample was collected. The test result only means that you did not have COVID-19 at the time of testing. Continue to take steps to protect yourself.

Stimulus 2 (United Kingdom NHS)

Your coronavirus test result

This information is about the swab test to check if you have coronavirus (COVID-19).

Getting your test result

You'll get a text or email when your result is ready.

Most people get their test results the day after taking the test. Some results might take longer, but you should get them in 72 hours.

There are 3 types of result you can get:

- negative
- positive
- unclear, void, borderline or inconclusive

Stay at home

If you had a test because you had symptoms, you and anyone you live with must stay at home (self-isolate) until you get your result.

Anyone in your support bubble must also self-isolate until you get your result.

Negative test result

A negative result means the test did not find coronavirus.

You do not need to self-isolate if your test is negative, as long as:

- everyone you live with who has symptoms tests negative
- everyone in your support bubble who has symptoms tests negative
- you were not told to self-isolate for 14 days
- you feel well – if you feel unwell, stay at home until you're feeling better
- If you have diarrhoea or you're being sick, stay at home until 48 hours after they've stopped.

If you're a health or care worker, check with your employer before going back to work.

Positive test result

A positive result means you had coronavirus when the test was done.

If your test is positive, you must self-isolate immediately.

- If you had a test because you had symptoms, keep self-isolating for at least 10 days from when your symptoms started.
- If you had a test but have not had symptoms, self-isolate for 10 days from when you had the test.

Anyone you live with, and anyone in your support bubble, must self-isolate for 14 days from when you start self-isolating.

Unclear, void, borderline or inconclusive test result

An unclear, void, borderline or inconclusive result means it's not possible to say if you had coronavirus when the test was done.

Get another coronavirus test as soon as possible if this happens.

If you had a test because you had symptoms, you must keep self-isolating and have another test within 5 days of your symptoms starting.

If you're not able to have another test in time, you must self-isolate for at least 10 days from when your symptoms started. Anyone you live with, and anyone in your support bubble, must self-isolate for 14 days.

If you had a test but have not had any symptoms, you do not need to self-isolate while you wait to get another test. People you live with, and anyone in your support bubble, do not need to self-isolate.

Stimulus 3 (New Zealand Ministry of Health)

COVID-19 test results and their accuracy

Information about how COVID-19 test results are reported, why they are sometimes uncertain, and what they can and cannot tell us

How the results are reported

COVID-19 test results are reported as positive or negative. If the test result is:

- positive - the virus (its genes) was detected in the sample
- negative - the virus was not detected in the sample.

Sometimes the test result might be reported as a 'weak positive', which means a very small amount of virus is in the sample. This could mean the person was tested towards the end of the illness when the level of virus in their body was low.

The accuracy of test results

The viral test for COVID-19 is accurate when taken in ideal conditions. A recent laboratory study found that different COVID-19 testing kits correctly detected COVID-19 in samples more than 95% (and frequently 100%) of the time. When tests were done on samples without the virus, the tests correctly gave a negative result 96% of the time.

But it is important to remember that tests don't work as well in the real world. No viral test is 100% accurate. In real-world use, the viral test for COVID-19 is not 100% 'sensitive' (meaning able to correctly identify people with the disease all of the time). This means that if 100 people who have the disease are tested – some will have a negative result (i.e. a false negative result).

Reasons for a false negative test result could be because:

- the sample was taken at the wrong time (too early or too late)
- the swab did not pick up any pieces of the virus, or
- the sample of mucus (or liquid from the lungs) wasn't big enough.

This means that it is very important to isolate yourself if a health professional asks you to, even if your test result is negative.

The viral test for COVID-19 is much better at correctly identifying people who don't have COVID-19 (this is known as a higher 'specificity'). We expect very few (if any) false positive test results (a false positive being a positive test result for someone who does not have the disease).

What the test results can and cannot tell us

Even when we take the uncertainties of testing into account, the results can tell us a few things.

A positive test tells us that a person either has COVID-19 (whether they have symptoms, or not) or has had COVID-19 recently. We may not be able to distinguish whether the person is currently infectious or not so we will take a precautionary approach.

A positive test cannot tell us:

• if the person is currently infectious
• how ill the person is likely to become.

A negative test can tell us:

• the person was unlikely to be infectious at the time of the test.

A negative test cannot tell us:

• if the person was exposed to the virus or not
• if they are in the early stages of incubating the disease
• if they caught COVID-19 in the past
• if they were infectious in the past
• that they will not get COVID-19 in the future.

Note: if a person has a negative test result and they are at a higher risk of having COVID-19, they may be tested again.

Stimulus 4 ('experimental' based on UK NHS) – Main study only.

Your coronavirus test result

This information is about the swab test to check if you have coronavirus (COVID-19).

Getting your test result

You'll get a text or email when your result is ready.

Most people get their test results the day after taking the test. Some results might take longer, but you should get them in 72 hours.

There are 3 types of result you can get:

- negative - the virus was not detected in the sample.
- positive - the virus was detected in the sample.
- unclear, void, borderline or inconclusive - it was not possible to provide a result for your sample.

Based on data about the test procedure we use, we expect very few (if any) people being given a 'positive' test result if they do not have the disease.

Whilst data shows the test is even better at correctly identifying people who don't have COVID-19, there are other reasons why some people may be given a 'negative' result when they do actually have COVID-19, such as:

- the sample was taken at the wrong time (too early or too late)
- the swab did not pick up any pieces of the virus, or
- the sample of mucus (or liquid from the lungs) wasn't big enough.

This means that it is very important to follow the instructions below, even if your test result is negative.

Stay at home

If you had a test because you had symptoms, you and anyone you live with must stay at home (self-isolate) until you get your result.

Anyone in your support bubble must also self-isolate until you get your result.

Negative test result

A negative result means the test did not find coronavirus.

You do not need to self-isolate if your test is negative, unless:

- you have any symptoms – if you feel unwell, stay at home until you're feeling better
- someone you live with, or who is in your support bubble, has symptoms and has not yet been tested

• someone you live with, or who is in your support bubble, has tested positive
• you have been told to self-isolate for 14 days
• you have diarrhoea or you're being sick. If this is the case, stay at home until 48 hours after they've stopped.

If you're a health or care worker, check with your employer before going back to work.

Positive test result

A positive result means you almost certainly had coronavirus when the test was done.

If your test is positive, you must self-isolate immediately.

• If you had a test because you had symptoms, keep self-isolating for at least 10 days from when your symptoms started.
• If you had a test but have not had symptoms, self-isolate for 10 days from when you had the test.

Anyone you live with, and anyone in your support bubble, must self-isolate for 14 days from when you start self-isolating.

Unclear, void, borderline or inconclusive test result

An unclear, void, borderline or inconclusive result means it's not possible to say if you had coronavirus when the test was done.

Get another coronavirus test as soon as possible if this happens.

If you had a test because you had symptoms, you must keep self-isolating and have another test within 5 days of your symptoms starting.

If you're not able to have another test in time, you must self-isolate for at least 10 days from when your symptoms started. Anyone you live with, and anyone in your support bubble, must self-isolate for 14 days.

If you had a test but have not had any symptoms, you do not need to self-isolate while you wait to get another test. People you live with, and anyone in your support bubble, do not need to self-isolate.

Supplementary analyses

Supplementary Analysis 1

After removing participants who answered between 48-52% and also gave a rating of their confidence in their answer at 50% or below – which possibly indicates that they were indicating 'I don't know' rather than that they thought the likelihood of the recipient having COVID-19 was around 50% – the average estimate of the likelihood that John had COVID-19 was 74.3% for a positive test result and 37.1% for a negative test result. A factorial ANOVA on these participants did not reveal a main effect of wording nor an interaction with test result. Further analysis using two one-way ANOVAs for positive and negative test result respectively confirmed this pattern, i.e. no significant differences between the wording conditions emerged.

Supplementary Analysis 2

Relationships with numeracy

We found significant but very weak associations between numeracy and higher quality of evidence perception (b=0.07, SE=0.02, p<.001, R^2=0.01) and higher trustworthiness ratings (b=0.09, SE=0.02, p<.001, R^2=0.01). When numeracy was added to the main effects models of test result and test wording, the previously observed main effect of test result remained significant for both perceived trustworthiness (F(1, 1713) = 37.9, p < .001, η_p^2 = 0.022) as well as quality of evidence perception (F(1, 1713) = 33.4, p < .001, η_p^2 = 0.019). Additionally, the previously observed small main effect of wording for perceived trustworthiness likewise remained significant after the inclusion of numeracy into the model (F(4, 1713) = 2.4, p = .049, η_p^2 = 0.006). Numeracy also emerged as a significant predictor in those models for both quality of evidence (F(1, 1713) = 17.5, p < .001, η_p^2 = 0.010) and trust perception (F(1, 1713) = 21.8, p < .001, η_p^2 = 0.013).

Relationships with education

Educational level ranged from 'No formal education above age 16' to 'Doctorate' on a 6-point Likert scale. Exploring the effects of education, we find significant but very weak associations between educational level and higher quality of evidence perception (b=0.07, SE=0.03 p = .013, R^2=0.003) and higher trustworthiness ratings (b=0.06, SE=0.03, p = .037, R^2=0.002). When education was added to the main effects models of test result and test wording, the previously observed main effect of test result remained significant for both perceived trustworthiness (F(1, 1735) = 40.9, p < .001, η_p^2 = 0.023) as well as quality of evidence perception (F(1, 1735) = 38.0, p < .001, η_p^2 = 0.021). Additionally, the previously observed small main effect of wording for perceived trustworthiness likewise remained significant after the inclusion of educational level into the model (F(4, 1735) = 2.4, p = .044, η_p^2 = 0.006). Education level also emerged as a significant, albeit very small, predictor in those models for both quality of evidence (F(1, 1735) = 7.3, p = .007, η_p^2 = 0.004) and trust perception (F(1, 1735) = 5.4, p = .020, η_p^2 = 0.003).

Relationships with ethnicity

Participants indicated their ethnicity as belonging to a range of specific groups that were given in the survey. Specifically, there were several sub-categories of 'white', 'mixed', 'Asian', 'Black', and 'Arab' to choose from, in addition to the option of choosing 'any other ethnic group'. For the purposes of exploring the effects of ethnicity in a meaningful manner, we aggregated the sub-categories into a binary measure of White/Non-White (as many of the sub-categories had too few observations in them for a reliable analysis). Looking at whether there was a difference in reported quality of evidence or trustworthiness between white and non-white participants did not reveal any significant effects (Quality of evidence: t = -0.4, df = 190.89, p = 0.696; Trustworthiness: t = -0.5, df = 194.62, p = 0.644). Non-parametric analyses results, conducted as a robustness check, were in line with the parametric findings. When adding ethnicity to the main effects models of test result and test wording, the previously observed main effect of test result remained significant for both perceived trustworthiness ($F(1, 1736) = 40.0$, $p < .001$, $\eta_p^2 = 0.023$) as well as quality of evidence perception ($F(1, 1736) = 36.6$, $p < .001$, $\eta_p^2 = 0.021$). Additionally, the previously observed small main effect of wording for perceived trustworthiness likewise remained significant after the inclusion of ethnicity into the model ($F(4, 1736) = 2.4$, $p = .045$, $\eta_p^2 = 0.006$). Ethnicity did not emerge as a significant predictor in those models, neither for quality of evidence nor trust perception.

Reading time across conditions

The text which participants saw in the different experimental conditions varied in length. As expected, average reading time between our experimental conditions varied as well.[1] On average, participants took longest to read the New Zealand text, followed by the experimental wording text, the UK text, and the US text was quickest to read. See table below.

Condition	Number of words in the text	Average reading time in minutes (95% CI)	Normalized reading time (in seconds per word) (95% CI)

[1] Note: The control group did not get any text to read, hence, the presented analysis of reading time is only based on the four experimental groups which received a piece of text.

Experimental wording	613	1.97 (1.80,2.15)	0.19 (0.18,0.21)
NZ wording	548	2.22 (1.98,2.47)	0.24 (0.22,0.27)
UK wording	443	1.92 (1.72,2.12)	0.26 (0.23,0.29)
US wording	225	1.66 (1.47,1.85)	0.44 (0.39,0.49)

Pilot study

Hypotheses

The first hypothesis preregistered in the pilot study (H1) was that real-world messaging which clearly acknowledges the presence of uncertainty (Ministry of Health New Zealand advice) would facilitate a better understanding of the true level of uncertainty associated with positive and negative PCR test results than messaging that does not (UK National Health Service advice).

The second preregistered hypothesis (H2) was that reading messaging which defines the meaning of positive and negative test results but fails to acknowledge the presence of uncertainty (UK National Health Service advice) would lead participants to underestimate the true probability of having COVID-19 after a negative test result. The full preregistration can be found at https://osf.io/8n62f.

Methods

Procedure (preregistered)

Participants were recruited from two sources, Prolific and Respondi, and presented with a questionnaire in Qualtrics. They were randomised to their condition via the Qualtrics 'randomise' function. After consenting to take part, they were presented with a scenario in which a fictional individual ('John') has been feeling ill. Participants were told that based on symptoms alone, a knowledgeable doctor believes that this fictional individual has a 50-50 chance of having COVID-19. They were then told the result of this individual's COVID-19 viral test result (either positive or negative). Depending on condition, they were then also presented with information about interpreting a COVID-19 test result published by the U.S. Centers for Disease Control, the U.K. National Health Service, the New Zealand Ministry of Health, or no information at all (see above for full text of each). That is, each participant was randomly assigned to one of eight conditions in a between-subjects 2 (test result: positive, negative) x 4 (message: CDC, NHS, Ministry of Health NZ, or none) factorial design.

Participants who failed an attention check question were excluded.

After reading the information, participants were asked the following questions:

Question	Response options
How easy or difficult did you find the information above to understand? How completely do you feel you understood the information above? How much effort did you feel you had to put into understanding the information above?	7-point Likert scale with the ends labelled "None" and "A lot"
What is your best guess as to the percent chance that John actually had COVID-19 at the time of his test, given his test result?	Free-text (limited to numerical answers only)

(Please type a number without the '%' sign)	
How high or low do you think the quality of the evidence for the accuracy of 'PCR' or 'swab' testing in general is?	Slider scale with the ends labelled "very low" and "very high" Recorded as an integer between 0-100
How high or low do you think the quality of the evidence behind John's test result is?	Slider scale with the ends labelled "very low" and "very high" Recorded as an integer between 0-100
To what extent do you think that the result from John's test was certain or uncertain?	Slider scale with the ends labelled "very uncertain" and "very certain" Recorded as an integer between 0-100
To what extent do you think that John's test result was… Accurate? Reliable? Trustworthy?	7-point Likert scale with the ends labelled "Not at all" and "Very much"
To what extent do you think the people who are responsible for communicating the results of the COVID-19 'PCR' or 'swab' tests are trustworthy?	7-point Likert scale with the ends labelled "Not at all" and "Very much"
To what extent do you think policies and restrictions, such as rules for being allowed to enter a country, should be based on the results of a 19 'PCR' or 'swab' test?	7-point Likert scale with the ends labelled "Not at all" and "Very much"

Prior research (van der Bles et al., 2020) suggested that explicitly communicating uncertainty increases the degree to which participants perceive a number as uncertain, with effect sizes from d = .37 to .72, although we expected we may see smaller effect sizes given differences in our experimental task and setup. Our power analysis was based on the tests for which we have clear hypotheses. For Analysis 1, to achieve 95% power with an alpha of .05, d = .3 for those tests, power analysis in G*Power indicated that we would require 1176 participants. To account for attrition due to failing of the attention check and to buffer for potentially smaller observed effect sizes of the effects for which we do not have clear directional hypotheses or prior research findings, we sampled a total of 1850 participants.

Analyses (preregistered)

Analysis 1a

Analysis 1a, the preregistered analysis intended to test H1, was a 2 (test result) x 4 (message) factorial ANOVA, with participants' understanding of the predictive value of the test result as the dependent variable, with a planned comparison comparing participants' understanding of the predictive value of the test result for participants viewing the NHS message vs. the Ministry of Health NZ message.

The key dependent variable for this analysis, "*participants' understanding of the predictive value of the test result*," was operationalized as the absolute difference between each participant's estimate of the probability of COVID ("What is your best guess as to the percent chance that John actually had COVID-19 at the time of his test?") and an expert estimate, i.e., the estimate of the posterior probability of having COVID-19 (corresponding to a prior of 50%) as determined by Watson, Whiting & Brush (2020), Table 1: 93% for a positive test result, and 24% for a negative test result.

Analysis 1b

The preregistered analysis corresponding to H2 was a one-sample t-test comparing participants' estimates to the expert estimates, for participants who saw the negative test result and the NHS message. While this hypothesis was restricted to participants in the negative test result condition who viewed the NHS message, we noted in the preregistration that we would conduct similar tests for participants who saw the negative test result as well as either the CDC message, Ministry of Health NZ message, or no message; we had no clear hypothesis about whether participants would underestimate the probability of COVID in these conditions.

Analysis 2

We preregistered 2 (test result) x 4 (message) factorial ANOVAs with participants' understanding, quality of evidence perception, perceived trustworthiness, and decision making as the dependent variables. We identified a number of measures that we intended to use to operationalize these constructs. We noted in our preregistration that if correlations between measures corresponding to the same construct were at least 0.7 and Cronbach's alpha was sufficient, we would combine multiple items into a single index for that construct. The measures are shown in the table above.

Results

Due to slight accidental oversampling 1869 participants were sampled, of whom 1657 passed the attention check and were included in analyses. 895 were recruited via Prolific and 762 from Respondi.

Analysis 1a

The ANOVA found a main effect of test result, $F(1, 1649) = 137.6$, $p < .001$, $\eta_p^2 = .077$, with participant estimates being farther from the expert estimate for positive test results (mean = 33.7%, SD = 20.8%) than for negative test results (mean = 24.1%, SD = 12.3%). This may have been influenced by the fact that half (50.0%) of participants provided an answer of 50% as their estimate—a response that other research has found participants commonly use to indicate "I don't know" rather than to indicate their true best guess (De Bruin et al., 2000)—and 50% was closer to the expert estimate for negative test results. There was also a main effect of message, $F(3, 1649) = 12.1$, $p < .001$, $\eta_p^2 = .022$; differences between participant estimates and expert estimates are given in Table S1. The planned comparison comparing participants' understanding of the predictive value of the test result for participants viewing the UK National Health Service message vs. the Ministry of Health NZ message found a greater discrepancy between participant and expert estimates for those viewing the UK message (mean = 31.7%, SD = 16.4%) vs. those viewing the NZ message (mean = 27.2%, SD = 19.3%), $t(1649) = 3.93$, $p < .001$.

In addition to this difference between those viewing the NZ and UK messages, Tukey's post-hocs showed that differences between UK/control, US/control, and US/NZ were all significant, $p_{adj} < .05$. There was also a significant interaction, $F(3, 1649) = 19.8$, $p < .001$, $\eta_p^2 = .035$, Table S1. For example, for those who were told John's test result was positive, the difference between expert and participant estimates was significantly greater for participants who had seen the US wording than for the control group ($p_{adj} < .001$), but there was no such difference for those who were told John's test result was negative ($p_{adj} = 1.0$).

Message	*Absolute differences between participant and expert estimates, Mean (SD)*		
UK	31.7% (16.4%)	*Negative test results*	23.7% (10.6%)
		Positive test results	39.7% (17.3%)
US	30.8% (18.2%)	*Negative test results*	23.2% (9.1%)
		Positive test results	38.3% (21.6%)
NZ	27.2% (19.3%)	*Negative test results*	26.8% (16.0%)
		Positive test results	27.5% (22.4%)
Control	25.9% (16.2%)	*Negative test results*	22.4% (11.7%)

Positive test results	29.0% (19.0%)

Table S1. Absolute differences between participant and expert estimates by wording group and test result.

Given that the distribution was highly non-normal, due to a high proportion of participants estimating John's chance of COVID to be 50%, we also conducted a non-parametric aligned ranks transformation ANOVA, which also found both main effects and the interaction to be significant.

Analysis 1b

For an individual having a 50% prior probability of COVID, who then tested negative, the expert estimate of that individual's probability of having COVID despite the negative test is 24%. Participants who were told that John had tested *negative* had significantly *higher* mean estimates than this irrespective of message condition (UK message: 39.4%, $p < .001$; US message: 36.7%, $p < .001$; NZ message: 41.5%, $p < .001$; control: 38.3%, $p < .001$). Participants who were told that John had tested *positive* had *lower* mean estimates than the expert estimate of 93% (UK message: 54.1%, $p < .001$; US message: 56.3%, $p < .001$; NZ message: 67.3%, $p < .001$; control: 65.9%, $p < .001$).

Analysis 2

Answers to the questions 'How easy or difficult did you find the information that we just showed you to understand?' and 'How completely do you feel you understood the information?' were highly correlated (r = .81) and combined into an index. The effort investment item, "How much effort did you feel you had to put into understanding the information above?", was correlated at less than .40 for both ease and completeness of understanding, and analysed as a single item. The ANOVA on the ease and completeness of understanding index measure revealed a main effect of message, F(2, 1236) = 18.6, p < .001, η_p^2 = .029. There was no main effect of test result and no interaction. Post-hoc Tukey tests suggested that participants felt the NZ message (mean = 5.03, SD = 1.38) and the UK message (mean = 5.19, SD = 1.49) were more difficult to understand than the US message (mean = 5.61, SD = 1.31), both p < .001. The ANOVA on the effort invested for understanding measure likewise revealed a main effect of message, F(2, 1238) = 14.88, p < .001, η_p^2 = .023. There was no main effect of test result and no interaction. Post-hoc Tukey tests suggested that participants felt they had to invest more effort into understanding the NZ message (mean = 4.91, SD = 1.58) and the UK message (mean = 4.69, SD = 1.86) compared to the US message (mean = 4.25, SD = 1.85), both p < .001. The control condition was not included in these analyses as participants were not asked these questions in the control condition.

With respect to perceptions of the quality of the evidence for the accuracy of 'PCR' or 'swab' testing generally *('How high or low do you think the quality of the evidence for the accuracy of 'PCR' or 'swab' testing in general is?')*, there was a main effect of test result, F(1, 1648) = 19.0, p < .001, η_p^2 = .011, with higher quality perceived by those who saw a positive test result (mean = 73.0, SD = 19.3) vs. those shown a negative test result (mean = 68.7, SD = 20.9). There was also an interaction between test result and message, F(3, 1648) = 2.6, p = .049, η_p^2 = .005; post-hoc Tukey tests suggested those shown the positive result perceived higher quality than those shown the negative result in the control condition (p < .001) and in the NZ condition (p = .029), but the differences were not significant in the UK and US conditions. There was no main effect of message.

With respect to the specific test result they were provided *('How high or low do you think the quality of the evidence behind John's test result is?')*, there was a main effect of test result, $F(1, 1646) = 33.5$, $p < .001$, $\eta_p^2 = .020$, with higher quality perceived by those who saw a positive test result (mean = 73.3, SD = 18.7) vs. those shown a negative test result (mean = 67.8, SD = 20.5). There was also an interaction between test result and message, $F(3, 1646) = 5.2$, $p = .002$, $\eta_p^2 = .009$. Post-hoc Tukey tests comparing the dependent variable for the different levels of test result, conditioned on message, suggested that positive tests were perceived as higher-quality than negative tests in the control condition ($p < .001$) and in the NZ condition ($p = .003$), but the differences were not significant in the UK and US conditions. There was no main effect of message.

Answers to the question '*To what extent do you think that the result from John's test was certain or uncertain?*' and to the questions '*To what extent do you think that John's test result was trustworthy?'/'accurate?'/'reliable?*' were all highly correlated (all $r > 0.7$, Cronbach's alpha = 0.95); these were therefore rescaled to 1-7 scales and averaged into a single "perceived trustworthiness" index. Answers to the question '*To what extent do you think the people who are responsible for communicating the results of the COVID-19 'PCR' or 'swab' tests are trustworthy?*' were much less correlated and this question was therefore not included in the index. The perceived trustworthiness ANOVA revealed a main effect of test result, $F(1, 1643) = 47.3$, $p < .001$, $\eta_p^2 = .028$, with higher trustworthiness for those who saw a positive test result (mean = 5.37, SD = 1.14) vs. those shown a negative test result (mean = 4.96, SD = 1.27). There was also an interaction between test result and message, $F(3, 1643) = 4.0$, $p = .007$, $\eta_p^2 = .007$. Post-hoc Tukey tests comparing the dependent variable for the different levels of test result, conditioned on message, suggested that those shown the positive result reported higher trust than those shown the negative result in the control ($p < .001$), NZ ($p < .001$), and US ($p = .005$) conditions, but the difference was not significant in the UK condition. There was no main effect of message.

Finally, concerning responses to the question *'To what extent do you think policies and restrictions, such as rules for being allowed to enter a country, should be based on the results of a 'PCR' or 'swab' test?'*, there was a main effect of test result, $F(1, 1652) = 6.5$, $p = .011$, $\eta_p^2 = .004$, with greater agreement for those told that John's test result was positive (mean = 5.42, SD = 1.45) than for those told it was negative (mean = 5.23, SD = 1.48). There was also a main effect of message, $F(3, 1652) = 2.9$, $p = .036$, $\eta_p^2 = .005$. There was no interaction, and differences between message conditions (NZ mean 5.21, SD 1.53; control mean 5.23, SD 1.46; UK mean 5.42, SD 1.44; US mean 5.44, SD 1.43) were not significant in a post-hoc Tukey test that included all six pairwise comparisons.

Non-preregistered analyses

Estimates of John's true probability of having COVID-19

While Analysis 1a considered the absolute difference between participant and expert estimates, participants' actual estimates of John's probability of having COVID-19 are also of interest. While we had investigated these to some degree in Analysis 1b, we felt it worthwhile to conduct an exploratory 2x4 ANOVA to look for potential differences by test result and by message. Nonparametric aligned ranks transformation ANOVA was chosen due to the non-normality of the distribution. There was a main effect of test result, $F(1, 1649) = 392.6$, $p < .001$, $\eta_p^2 = .925$, with participants estimating John to have a higher probability of having COVID-19 if his test result was positive (mean = 60.9%, SD = 23.1%) than if it was negative (mean = 39.0%, SD = 22.5%). There was also a main effect of wording, $F(3, 1649) = 9.9$, $\eta_p^2 = .535$, and a significant interaction, $F(3, 1649) = 8.6$, $p < .001$. Nonparametric post-hoc tests (Mann-Whitney U tests) suggested that averaged across test result, differences between all pairs of wording groups were significant (all $p < .001$, Table S2), with the exception of the UK/US and control/NZ comparisons.

Zooming in on participants who were told that John's test result was positive, nonparametric post-hocs suggested differences between NZ/US, NZ/UK, control/US, and control/UK (all $p < .001$), with estimates for the NZ and control groups higher than those for the US and the UK (Table S2). Differences among answers for those told that John's test result was negative were not nearly so stark, with unadjusted nonparametric post-hocs only finding a possible difference between the NZ and US estimates (Table S2), $p = .04$.

Message	*Participant estimates of the probability that John has COVID-19, Mean (SD)*		
UK	46.8% (21.4%)	*Negative test result*	39.5% (20.9%)
		Positive test result	54.2% (19.2%)

US	46.6% (24.9%)	*Negative test result*	36.7% (21.5%)
		Positive test result	56.3% (24.1%)
NZ	53.8% (28.3%)	*Negative test result*	41.5% (25.9%)
		Positive test result	67.3% (24.5%)
Control	52.7% (25.4%)	*Negative test result*	38.3% (20.9%)
		Positive test result	65.9% (21.7%)

Table S2. Participant estimates of the probability that John has COVID-19, by wording group and test result.

Further exploratory analysis revealed that the distributions behind the means of each wording group were different. Overall, participants differed in their propensity to give a categorical '100%' or '0%' answer by message condition; a chi-squared test comparing participants giving categorical vs. noncategorical answers by wording condition was significant, $X^2(3, N = 1657) = 16.8$, with the NZ wording group having the lowest propensity to answer categorically overall, although results were not consistent when split out into participants in the positive vs. negative test result conditions.

Who was most likely to answer 50%?

As mentioned, half (50.0%) of participants provided an answer of 50% as their estimate of John's probability of having COVID-19. These participants may simply have been indicating that they did not know the answer (De Bruin et al., 2000), or they may have felt that the test result was completely uninformative (or simply been confused by it) and therefore reproduced the pre-test percentage. In any of these cases, answering precisely 50% seems to be indicative of a lack of understanding that John's probability of having COVID-19 should be adjusted upwards on receipt of a positive test result and downwards on a negative result. We therefore felt it worthwhile to look for differences across message conditions with respect to participants' propensity to answer "50%," and indeed a X^2 test showed that the number of participants answering 50% vs. any other value differed by condition, (X^2(3, N = 1657) = 48.1, p < .001, Table S3). Follow-up X^2 tests comparing the UK message condition to each other group in turn found that those in the UK message group were more likely to answer 50% than those in the US (p = .018), NZ (p < .001), or Control (p < .001) message groups.

Message	*Number of participants answering 50% (percentage)*		
UK	260/420 (61.9%)	*Negative test results*	119/210 (56.7%)
		Positive test results	141/210 (67.1%)
US	220/411 (53.5%)	*Negative test results*	105/204 (51.5%)
		Positive test results	115/207 (55.6%)
NZ	164/412 (39.8%)	*Negative test results*	90/215 (41.9%)
		Positive test results	74/197 (37.6%)
Control	184/414 (44.4%)	*Negative test results*	88/198 (44.4%)
		Positive test results	96/216 (44.4%)

Table S3. Number of participants answering 50% when asked to express the probability that John has COVID-19, by wording group and test result.