BMJ Open

Public's understanding of swab test results for SARS-CoV-2: an online behavioural experiment during the April 2020 lockdown

Stefania Pighin, Katya Tentori

To cite: Pighin S, Tentori K. Public's understanding of swab test results for SARS-CoV-2: an online behavioural experiment during the April 2020 lockdown. *BMJ Open* 2021;**11**:e043925. doi:10.1136/bmjopen-2020-043925

► Prepublication history and additional material for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-043925).

Received 19 August 2020
Revised 29 November 2020
Accepted 14 December 2020

Center for Mind/Brain Sciences, University of Trento, Rovereto (TN), Italy

Correspondence to
Professor Katya Tentori;
katya.tentori@unitn.it

ABSTRACT

Objective Although widespread testing for SARS-CoV-2 is in place, little is known about how well the public understands these results. We aimed to provide a comprehensive overview of the general public's grasp of the accuracy and significance of the results of the swab test.

Design Web-based behavioural experiment.

Setting Italy during the April 2020 lockdown.

Participants 566 Italian residents.

Main outcome measures Participants' estimates of the SARS-CoV-2 prevalence; the predictive and diagnostic accuracy of the test; the behavioural impact of (positive vs negative) test results; the perceived usefulness of a short-term repetition of the test following positive or negative results; and rankings of causes for false positives and false negatives.

Results Most participants considered the swab test useful (89.6%) and provided predictive values consistent with their estimates of test diagnostic accuracy and infection prevalence (67.0%). Participants acknowledged the effects of symptomatic status and geographical location on prevalence (all $p<0.001$) but failed to take this information into account when estimating the positive or negative predictive value. Overall, test specificity was underestimated (91.5%, 95% CI 90.2% to 92.8%); test sensitivity was overestimated (89.7%, 95% CI 88.3% to 91.0%). Positive results were evaluated as more informative than negative ones (91.6, 95% CI 90.2 to 93.1 and 41.0, 95% CI 37.9 to 44.0, respectively, $p<0.001$); a short-term repetition of the test was considered more useful after a positive than a negative result (62.7, 95% CI 59.6 to 65.7 and 47.2, 95% CI 44.4 to 50.0, respectively, $p=0.013$). Human error and technical characteristics were assessed as more likely to be the causes of false positives ($p<0.001$); the level of the viral load as the cause of false negatives ($p<0.001$).

Conclusions While some aspects of the swab for SARS-CoV-2 are well grasped, others are not and may have a strong bearing on the general public's health and well-being. The obtained findings provide policymakers with a detailed picture that can guide the design and implementation of interventions for improving efficient communication with the general public as well as adherence to precautionary behaviour.

Strengths and limitations of this study

- This study provides a first comprehensive overview of the general public's understanding of the most commonly used test for detecting SARS-CoV-2.
- This study considers not only participants' estimates of the positive predictive value but also of the negative predictive value and the diagnostic accuracy of tests, as well as their grasp of the test results' behavioural consequences.
- This study employs a rigorous experimental design that allowed for control of many potentially confounding variables, such as participants' geographical location, worry and perceived individual risk.
- The findings provide policymakers with a detailed picture that can guide the design of interventions for improving both efficient communication with the public and adherence to precautionary behaviour.
- Further research is needed to extend this investigation to a wider population, including older adults, individuals with pre-existing conditions and those who have already been tested for SARS-CoV-2.

INTRODUCTION

The global outbreak of the COVID-19 has abruptly placed testing at the centre of everyone's thoughts, actions and feelings. Widespread testing has been strongly recommended by WHO,[1] as the rapid identification of possible new infectious cases is considered decisive in reducing clinical progression, in containing onward transmission[2] and, in the long run, in saving lives and resuming normal life.[3] Although employing different strategies and timelines,[2 4] mass testing has now been implemented in many countries: thousands of individuals are tested every day, and even more are seeking to be tested. However, currently, there is not much knowledge on how this testing is perceived. To what extent does the general population consider it to be accurate and informative? What sense is there of the impact of test results on behaviour and of the usefulness of a short-term repetition of

the test? Are possible test errors ascribed to the most probable causes? In this study, we address these questions by investigating, under various experimental conditions, the public's grasp of the accuracy and significance of results of the reference standard for the detection of the novel coronavirus responsible for COVID-19 (SARS-CoV-2): the real-time reverse transcriptase polymerase chain reaction (rRT-PCR) performed on respiratory specimens.[2 5]

A PubMed search[i] from inception to 24 November 2020 did not identify any research article that investigated how the results of molecular (nor serological) tests for SARS-CoV-2 are interpreted or understood. A similar absence is found in the systematic literature review on SARS-CoV-2 by Rajendran and colleagues.[6] Studies on how laypeople and experts understand the accuracy of screening and diagnostic tests are available with reference to other conditions (hypothetical and real life), such as breast cancer and genetic disorders.[7–12] The main result across these investigations, especially those that considered low-prevalence conditions, is that the great majority of individuals—including a wide range of healthcare providers—systematically overestimate the probability that individuals with a positive test result truly have the disease (namely the positive predictive value of the test, hereafter PPV).[7] This robust finding is an expression of a more general tendency to discount or even ignore base rate information in favour of relevant evidence[13] (a phenomenon known as the base rate fallacy or base rate neglect) and can be, at least partially, modulated by various factors, such as the format of the statistical information conveyed[12 14] or the specific probability question posed.[8] For instance, in Garcia-Retamero *et al*'s study,[15] patients' incorrect diagnostic inferences concerning various positive screening test results decreased from more than two-thirds to less than half when the numerical information concerning the prevalence rate, the sensitivity (SE) and the false positive rate was presented along with a visual display representing the overall number of individuals at risks, the number of individuals who obtained a positive result and the number of individuals who have the disease.

Though extremely interesting from a cognitive perspective and useful for regular medical practice, the results of earlier studies cannot be extrapolated to the public's understanding of the extensive testing now underway. Indeed, the situation generated by the COVID-19 pandemic is new in various respects.

First, as the clinical validation of the newly developed rRT-PCR test for detecting the SARS-CoV-2 is still at an early stage, its accuracy is not yet fully known. Preliminary results suggest a high rate of false negatives and a limited rate of false positives. Reported SE varies widely (depending, for example, on the site, quality and timing of sampling), with most values converging on 70%–85%,[16–19] while specificity (SP) has received less attention[20] and is assumed to be greater than 98%.[20–22] The lack of precise statistics describing test performance and, above all, on the true prevalence of this infection in specific populations[23] makes it currently impossible to calculate the exact PPV. In the absence of this normative value, estimates of the accuracy of this test need to be assessed according to different criteria.

Second, given the considerable extent of asymptomatic carriage[2 24] for the SARS-CoV-2 infection, test results are particularly crucial. Previous research focused almost exclusively on the understanding of PPV. Yet, when millions stand to be exposed to an infection, it is in fact the interpretation of negative results that becomes most challenging. This is because the more prevalence rises, the greater the proportion of false negatives among all negative results and, consequently, the lower the predictive value of a negative result (NPV).[22 25] So we cannot assume that previous findings on the difficulty that is encountered in calculating the PPV generalise to the assessment of the NPV. Nor can they be applied to the predictive value of a double negative test result at a 24-hour interval (NNPV), which has been used as a discharge criterion[26] for patients with COVID-19 from hospitals in various countries.

A third element of novelty concerns participants' high personal involvement. The COVID-19 pandemic and associated containment measures have had a tremendous impact on behaviours and priorities,[27] with the consequence that many individuals who are not 'COVID-19 patients', and who may never be, feel threatened and anxious.[28 29] This, together with extraordinary exposure to health-related information,[30] including data and arguments about testing, as well as the utility of implementation on a large scale, raises the question of how the population in the current emergency situation is comparable to participants of previous experiments, who typically have been presented with medical test scenarios quite remote from their direct experience.

Our study provides insight into the questions and novelties outlined above by offering new empirical data and a new methodological approach for evaluating the accuracy of the PPV and NPV estimates in the absence of definitive evidence on the diagnostic accuracy of a test (see the coherence criterion in box 1). It also complements our comprehension of the understanding of test results by exploring the public's grasp of the various implications of these results. Gaining deeper insight into the general public's understanding of the accuracy and significance of the most widespread test for SARS-CoV-2 offers a unique opportunity to improve scientific knowledge on reasoning about medical tests and could have tangible implications for pandemic health policies now by facilitating more efficient risk communication and by promoting adherence to precautionary behaviour.

[i]The literature search was conducted with no restrictions on language using the search string: (2019-nCoV [All fields] OR SARS-CoV-2 [All fields] OR novel coronavirus [All fields] OR Covid-19 [All fields]) AND (testing result* [All fields] OR test result* [All fields]) AND (interpretation [All fields] OR understanding [All fields] OR assessment [All fields] OR predictive value* [All fields]) AND (probability [All fields] OR reasoning [All fields]).

Box 1 Test usefulness and qualitative coherence criteria

The restricted data set included all participants whose estimation of test characteristics met criterion 1 and whose judgements of PVs, prevalence and test characteristics met all of the conditions in criterion 2 (n=358, 63.3%).

Criterion 1: test usefulness (n=507, 89.6%)

To establish whether the test characteristics provided by participants were compatible with those of a 'useful' test, we applied the following rule[37]: SE+SP≥1.5, that is, estimated diagnostic accuracy was at least halfway between 1 (a completely useless test) and 2 (a perfect test).

Criterion 2: coherence between probability estimates (n=379, 67%)

To determine whether the PVs, prevalence and test characteristics provided by participants were coherent with each other, at least from a qualitative point of view, we set the following conditions:

Coherence between *PPV*, *prevalence rate* (*PR*) and *test characteristics* (*SE* and *SP*) (n=531, 93.8%)
(PPV>PR∧SE>(1-SP))∨(PPV<PR∧SE<(1-SP))∨(PPV=PR∧SE=(1-SP))∨(PPV=1∧PR=1∧SE>(1-SP))

Coherence between *NPV*, *prevalence rate* (*PR*) and *test characteristics* (*SE* and *SP*) (n=441, 77.9%)
(NPV>(1-PR)∧SP>(1-SE))∨(NPV<(1-PR)∧SP<(1-SE))∨(NPV=(1-PR)∧SP=(1-SE))∨(NPV=1∧(1-PR)=1∧ SP>(1-SE))

Coherence between *NNPV*, *NPV* and *test characteristics* (*SE* and *SP*) (n=486, 85.9%)
(NNPV>NPV∧SP>(1-SE))∨(NNPV<NPV∧SP<(1-SE))∨(NNPV=NPV∧SP=(1-SE))∨(NNPV=1∧NPV=1∧SP>(1-SE))

NNPV, double negative predictive value; NPV, negative predictive value; PPV, positive predictive value; PVs, predictive values; SE, sensitivity; SP, specificity.

METHODS

Study design, stimuli and procedure

Our online behavioural experiment consisted of four parts (online supplemental material 1 fully reports the stimuli, translated from Italian) and was carried out through Prolific Academic (http://prolific.ac), one of the most popular and reliable crowdsourcing platforms for behavioural research.[31] This gave us access to a general population, although with a lower representation of older adults (in Italy, the proportion of Prolific participants older than 65 years is lower than it is in other countries, such as the UK or the USA). The experiment was computer based and participants carried it out at home. There were no time limits, and the task was typically completed in less than 7 min. Participants received £0.60 compensation.

The first part of the experiment aimed to explore whether participants' estimates of predictive and diagnostic accuracy depended on the prevalence rate. To this end, we employed a 4×3 between-subjects design, in which participants were asked to consider the hypothetical case of a person identified by a combination of two factors: her symptomatic status (SX: unspecified, absent, mild or severe) and her geographical location (Italy, Sassari or Bergamo). Italy represents a generic location (for an Italian participant) while Sassari and Bergamo are two well-known cities of comparable population that largely differ in reported infection and death rates (figure 1). Participants in all groups were asked to estimate the prior probability that the person had the SARS-CoV-2 infection (we consider this evaluation generalisable to the subpopulation to which the hypothetical person belongs and, therefore, will refer to it as judged prevalence) and three predictive values (PVs): the probability that a person from the same subpopulation had the SARS-CoV-2

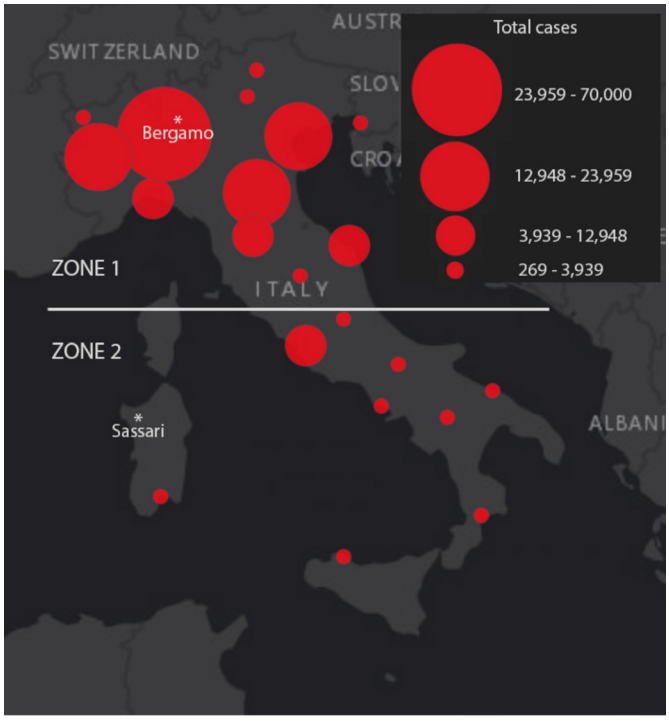

Figure 1 Zones 1 and 2. Geographical distribution of COVID-19 cases in Italy on the first day of data collection (April 6) in the two prevalence areas: zone 1 (latitude ≥42°50), which included the hardest hit regions (more than 500 000 cases, test positivity rate ≥20%), and zone 2 (latitude <42°50), which comprises the least affected regions (fewer than 200 000 cases, test positivity rate ≤10%). Bergamo (in zone 1) and Sassari (in zone 2) are two cities of comparable populations but very different COVID-19 infection and death rates: while the former was the epicentre of the first COVID-19 outbreak in Italy, the latter has passed relatively unscathed, with about 1/20 of the cases of the former.

infection, given a positive test result (judged PPV); the probability that a person from the same subpopulation did not have the SARS-CoV-2 infection, given a negative test result (judged NPV) or given two negative test results at a 24-hour interval (judged NNPV). The phrasing of the questions was adapted from previous studies on Bayesian inferences in the medical domain.[8–10] Participants were then asked to provide their best estimates of the test SE and SP by judging the probability that a person from the same subpopulation who had (did not have) the SARS-CoV-2 infection would receive a positive (negative) test result. Although irrelevant for the last two questions, we kept the reference to the subpopulation in order to see whether participants' estimates were affected by it. To reduce possible misunderstandings, we asked for complementary judgements in pairs and clarified that they had to sum up to 100%; participants' compliance with this requirement also served as an attentional check.

Since emotion-related variables such as worry and perceived risk are acknowledged to drive probability judgements and attitudes towards medical tests,[32] participants were also asked to report their worry for the pandemic, to estimate their likelihood of contracting the infection and to evaluate its severity. The last two measures were then multiplied to obtain a perceived risk score.[33]

The remainder of the experiment was identical for all groups. The second part aimed to investigate whether participants were aware of the asymmetric implications of positive and negative results in terms of impact on behaviour and the usefulness of a short-term test/retest. In this regard, a positive result should be considered rather informative because it implies self-isolation, while a negative result is not expected to substantially affect behaviour, especially during the lockdown. On the other hand, repeating the test might be useful after a negative result (because of the high rate of false negatives) while it appears less justified after a positive result (due to the low rate of false positives, but also because positivity in itself does not impact treatment).

To investigate participants' explanations of test errors, in the third part of the experiment, we asked them to rank, according to their probability, three possible causes of false positives and false negatives (human error, technical characteristics of the test and level of viral load).

Responses to all questions in the first three parts of the experiment were mandatory for completion of the experiment and to receive the Prolific payment.

Finally, in the fourth and final part of the experiment, participants were asked their personal experience with the swab test and COVID-19. Demographic information was obtained from Prolific.ac.

Evaluation criteria

Participants who evaluated the test as useful and who judged prevalence, test characteristics, PPV, NPV and NNPV as qualitatively coherent among each other were identified using the criteria reported in box 1.

Statistical analysis

Participants' characteristics were analysed by means of χ^2 tests for categorical variables and t-tests for continuous variables. For all variables, we calculated descriptive statistics such as means and 95% CIs. Statistical comparisons were evaluated by multivariate analysis of variance (MANOVA) followed by post hoc pairwise comparisons using Tukey's honest significant difference, t-tests and repeated measures analysis of variance (ANOVA) when appropriate. In order to improve the accuracy of the models, the ANOVAs were performed including participants' age, gender and education as covariates. Wilcoxon signed-rank tests were used to compare participants' rankings of causes of false-positive and false-negative test results. Possible effects of participants' characteristics (ie, age, gender and education) on rankings were preliminarily investigated by means of Mann-Whitney tests. All analyses concerning the prevalence, the predictive and the diagnostic accuracy were performed twice: once including all participants (*full data set*) and once including only participants whose judgements met the criteria reported in box 1 (*restricted data set*). Data analysis was performed with SPSS V.23. Only p values below 0.05 were considered significant and reported within the text.

Participants and data collection

A total of 591 native Italian speakers residing in Italy were recruited on 6–9 April 2020, during the total lockdown. We excluded from the analyses 22 participants who failed to pass the attention checks (ie, their complementary responses did not sum up to 100) and three participants who assigned extreme (0 and 100) values both to prevalence and test characteristics, making it impossible to compute a meaningful value for some of their expected PVs. The final sample thus included 566 participants (see table 1 for related statistics).

Figure 1 reports the geographical distribution of COVID-19 cases in Italy on the first day of data collection. Since the disease mainly affected the northern regions, we classified participants' locations into two different areas: zone 1 (latitude ≥42°50), which encompasses the hardest hit regions, and zone 2 (latitude <42°50), which includes the regions with the lowest incidence of the disease. All participants provided informed, written consent.

RESULTS

Participants' characteristics

The mean age of participants was 28 years (95% CI 27.5 to 29.0), ranging from 18 to 66 years. Age did not significantly differ for males and females, nor did educational level. Participants in zones 1 and 2 divided 58% to 42%, respectively, and this roughly parallels the percentage split of Italians living in the two zones (about 56% and 44%). Educational level did not differ significantly in the two zones. Only three participants declared they had undergone the swab test, all residing in zone 1. More participants in zone

Table 1 Participants' characteristics

	Zone 1		Zone 2		
	Age 18–25 (n=151)	Age Over 26 (n=176)	Age 18–25 (n=129)	Age Over 26 (n=110)	Overall (n=566)
Demographics					
Gender					
Male (%)	81 (54)	85 (49)	64 (50)	43 (39)	273 (48)
Female (%)	69 (46)	90 (51)	65 (50)	67 (61)	291 (52)
Education					
High school or lower (%)	94 (62)	68 (39)	80 (62)	45 (41)	287 (51)
University degree or higher (%)	57 (38)	107 (61)	48 (38)	65 (59)	277 (49)
Employment					
Full time (%)	13 (9)	66 (38)	12 (9)	35 (32)	126 (22)
Part time (%)	22 (15)	35 (20)	25 (19)	18 (16)	100 (18)
Not in paid work (%)	15 (10)	10 (6)	11 (9)	5 (5)	41 (7)
Unemployed and job seeking (%)	34 (22)	37 (21)	35 (27)	41 (37)	147 (26)
Students or unspecified (%)	67 (44)	28 (15)	46 (36)	11 (10)	152 (27)
Experience with swab test or COVID-19					
Underwent swab test (%)	1 (0.7)	2 (1)	0 (0)	0 (0)	3 (0.5)
Friend/relative/colleague underwent swab test (%)	35 (23)	44 (25)	15 (12)	16 (15)	110 (19)
Friend/relative/colleague had COVID-19 (%)	47 (31)	63 (36)	29 (23)	28 (26)	167 (29)
Worry and risk assessment					
Worry about the pandemic	61.3 (57.6 to 64.9)	68.0 (64.3 to 71.7)	68.0 (64.5 to 71.5)	71.9 (67.8 to 75.9)	66.9 (65.1 to 68.8)
Individual probability (in %)	34.1 (30.1 to 38.2)	36.9 (33.3 to 40.5)	27.3 (23.9 to 30.8)	31.4 (27.7 to 35.1)	32.9 (31.0 to 34.8)
Individual severity	54.2 (50.1 to 58.3)	65.9 (62.1 to 69.6)	62.2 (58.4 to 66.0)	68.9 (64.5 to 73.3)	62.5 (60.5 to 64.6)
Perceived risk	19.9 (17.0 to 22.9)	24.9 (22.1 to 27.7)	18.1 (15.1 to 21.0)	23.2 (19.8 to 26.6)	21.7 (20.2 to 23.2)

Values are n (%) or mean (95% CI). A perceived risk score was computed for each participant by multiplying the probability of being infected with SARS-CoV-2 and the severity of being infected with SARS-CoV-2. Thirteen participants did not declare their age; two their gender; two their educational level; one if she/he had undergone the swab test; and one if she/he knew someone who had undergone the swab test.

1 than in zone 2 reported that a person in their circle (ie, relatives, friends, colleagues) had undergone the swab test or had been diagnosed with COVID-19 (p=0.001 and p=0.012, respectively, χ^2 tests). These results support our partition by confirming a greater spread of the virus in zone 1 than in zone 2. The mean worry for the pandemic was 66.9 (95% CI 65.1 to 68.8). Somewhat surprisingly, participants in zone 1 reported lower worry than those in zone 2 (64.9, 95% CI 62.2 to 67.5 and 69.8, 95% CI 67.2 to 72.4, respectively, p=0.011, independent samples t-test). No matter the zone, younger (18–25 years) participants reported less worry than older (>26 years) ones (64.4, 95% CI 61.8 to 66.9 and 69.5, 95% CI 66.7 to 72.2, respectively, p=0.008, independent samples t-test).

A similar pattern was observed for the severity of contracting the virus, which was lower for participants in zone 1 than in zone 2 (60.5, 95% CI 57.6 to 63.3 and 65.3, 95% CI 62.4 to 68.2, respectively, p=0.022, independent samples t-test) and for younger participants than older ones (57.9, 95% CI 55.0 to 60.7 and 67.0, 95% CI 64.2 to 69.9, respectively, p<0.001, independent samples t-test). Participants in zone 1 estimated the probability of contracting the virus as higher than those in zone 2 (35.6%, 95% CI 33.0% to 38.3% and 29.2%, 95% CI 26.7% to 31.7%, respectively, p=0.001, independent samples t-test). The perceived risk differed between younger and older participants (19.1, 95% CI 17.0 to 21.1 and 24.2, 95% CI 22.1 to 26.4, respectively, p=0.001,

independent samples t-test), but not between zones 1 and 2.

Predictive and diagnostic accuracy: qualitative coherence and test usefulness

The great majority of participants (89.6%) evaluated the test as useful (see criterion 1 in box 1) and provided estimates of predictive accuracy that were coherent with their evaluation of diagnostic accuracy and with their beliefs about prevalence (67.0%) (see box 1). Participants (63.3% of the total) whose judgements met both these criteria were included in the restricted data set. It is worth noting that this does not indicate that the remaining participants hold irrational beliefs; they may simply not have read all questions or response options carefully (in particular, some participants seem to have confused the order of two complementary responses in the NPV question, see the online supplemental material 1).

Predictive and diagnostic accuracy: effects of SX and location

The MANOVA used to investigate the effect of the SX and location on judged prevalence, PPV, NPV, NNPV, SE and SP (with age, gender, educational level, zone, worry and perceived risk as covariates) showed that, in both data sets, participants' prevalence judgements were affected in the expected direction by the manipulation of SX and location. The effect was less systematic for judged PPV, which depended on SX in both data sets but only on location in the restricted data set, as well as for judged NPV, which depended on SX and location in the restricted data set alone (all $p<0.05$, table 2 for the outputs of the MANOVA and figure 2 for Tukey's post hoc tests; see also online supplemental material 2 for mean judgements and 95% CI). These results indicate that participants—at least when they provided qualitatively coherent probability judgements—were sensitive to factors that can affect prevalence, PPV and NPV. Irrespective of the data set, judged SE and SP did not differ significantly across groups (table 2), indicating that participants correctly estimated the test's diagnostic accuracy independently of prevalence. Among the covariates, the most robust effects were those exerted by worry and perceived risk, both on judged prevalence; irrespective of the experimental condition, participants who expressed a greater worry and/or greater perceived risk also provided higher estimates of prevalence.

Predictive accuracy: consistency between judged PVs and expected PVs

Since the exact prevalence of the infection in the considered subpopulations is unknown, objective PVs cannot be computed. The PPV and NPV provided by each participant were therefore compared with the expected PPV and NPV that were obtained by inserting his/her judgements of prevalence and test characteristics into the Bayes theorem. The comparison between judged and expected PVs (table 3, figure 3) reveals that participants overestimated the PPV in the full data set ($p<0.001$, paired sample t-test) but underestimated the NPV in both data sets (all $p<0.001$, paired sample t-test). To control for the base rate fallacy, we performed two further analyses that focused on the judged PPV and NPV of participants who provided low values (≤20) for the prevalence and 1-prevalence, respectively. Regardless of the data set, judged PPV was greater than expected PPV (all $p<0.01$, paired sample t-test). This result supports those of previous research on the base rate fallacy and indicates that participants underweighted their own estimates of prevalence and/or diagnostic accuracy when updating their beliefs based on a positive test result. By contrast, no significant difference was observed between judged and expected NPV when the prevalence was assumed to be high (ie, 1-prevalence ≤20%). Such findings suggest that the base rate fallacy that has been repeatedly observed for the PPV does not extend to the NPV and, if replicated, would require modification of most theoretical models that have proposed to explain this phenomenon.

Predictive accuracy: NNPV

To check whether participants acknowledged that a double negative result supports the absence of the infection more than a single negative result does, judged NNPV and NPV were compared. Irrespective of the experimental condition, participants correctly indicated a higher value (all $p<0.001$, paired sample t-test) for judged NNPV (91.9%, 95% CI 90.2% to 93.5% and 97.6%, 95% CI 97.1% to 98.1% in full and restricted data sets, respectively) than judged NPV (81.1%, 95% CI 79.0% to 83.2% and 88.5%, 95% CI 87.0% to 90.1% in full and restricted data sets, respectively). Furthermore, in both data sets, the judged NNPV was lower than 100% (all $p<0.001$, one-sample t-tests), suggesting that participants had a correct grasp of the fact that a double negative test result does not rule out the possibility of an infection.

Diagnostic accuracy: consistency of judged SE and SP with experts' estimates

In line with experts' assessments, irrespective of the experimental condition and data set, participants judged the SP (91.5%, 95% CI 90.2% to 92.8% and 95.9%, 95% CI 95.2% to 96.5% for all and restricted data sets, respectively) as higher than the SE (89.7%, 95% CI 88.3% to 91.0% and 94.4%, 95% CI 93.6% to 95.2% for all and restricted data sets, respectively) according to paired sample t-tests (all $p<0.05$). In both data sets, judged SE was above the upper bound of the current reference range (85%) and judged SP was below the lower bound of the current reference range (98%) (all $p<0.001$, one-sample t-tests). These results complement those obtained for the PPV and NPV and confirm that participants, on one hand, overestimate the test's ability to correctly detect infected individuals and, on the other hand, underestimate the test's ability to correctly identify non-infected individuals.

Table 2 Main results of the MANOVA conducted on participants' judgements in full and restricted data sets

	Prevalence			PPV			NPV			NNPV			SE			SP		
Factors	F	P value	η^2	F	P value	η^2	F	P value	η^2	F	P value	η^2	F	P value	η^2	F	P value	η^2
Full data set																		
Location	37.05	0.000	0.122	0.92	0.400	0.003	2.72	0.067	0.010	0.10	0.909	0.000	1.77	0.172	0.007	2.56	0.078	0.010
SX	102.02	0.000	0.366	4.80	0.003	0.026	2.45	0.063	0.014	0.40	0.754	0.002	1.17	0.322	0.007	1.36	0.253	0.008
SX × Loc	1.84	0.089	0.020	0.55	0.774	0.006	0.76	0.599	0.009	0.33	0.924	0.004	0.29	0.940	0.003	1.42	0.206	0.016
Age	2.27	0.133	0.004	9.13	0.003	0.017	3.67	0.056	0.007	0.17	0.678	0.000	4.51	0.034	0.008	0.31	0.579	0.001
Gender	2.97	0.085	0.006	0.26	0.610	0.000	0.98	0.323	0.002	0.19	0.663	0.000	2.93	0.087	0.005	2.47	0.116	0.005
Education	0.53	0.465	0.001	0.02	0.897	0.000	4.45	0.035	0.008	0.04	0.836	0.000	1.45	0.229	0.003	1.54	0.215	0.003
Zone	0.24	0.627	0.000	0.00	0.997	0.000	1.23	0.269	0.002	0.23	0.630	0.000	1.14	0.286	0.002	2.84	0.092	0.005
General worry	26.37	0.000	0.047	12.58	0.000	0.023	2.52	0.113	0.005	0.00	0.979	0.000	3.51	0.061	0.007	0.05	0.825	0.000
Perceived risk	8.21	0.004	0.015	0.11	0.739	0.000	7.09	0.008	0.013	2.82	0.094	0.005	1.52	0.219	0.003	0.97	0.325	0.002
Restricted data set																		
Location	22.09	0.000	0.118	3.90	0.021	0.023	4.58	0.011	0.027	2.16	0.118	0.013	1.50	0.225	0.009	0.11	0.893	0.001
SX	66.23	0.000	0.377	4.60	0.004	0.040	7.75	0.000	0.99	2.56	0.055	0.023	1.46	0.225	0.013	1.82	0.143	0.016
SX × Loc	1.70	0.122	0.030	1.52	0.172	0.027	0.53	0.785	0.066	1.09	0.366	0.020	1.08	0.376	0.019	0.74	0.622	0.013
Age	0.46	0.500	0.001	2.47	0.117	0.007	2.79	0.096	0.008	0.06	0.801	0.000	0.40	0.527	0.001	0.11	0.747	0.000
Gender	2.37	0.125	0.007	0.41	0.522	0.001	0.02	0.890	0.000	0.66	0.418	0.002	5.42	0.020	0.016	2.77	0.097	0.008
Education	2.89	0.090	0.009	1.62	0.204	0.005	0.14	0.707	0.000	0.18	0.669	0.001	0.73	0.395	0.002	0.09	0.762	0.000
Geographical area	0.31	0.576	0.001	0.44	0.510	0.001	0.26	0.609	0.001	0.83	0.362	0.003	0.36	0.549	0.001	0.70	0.402	0.002
General worry	16.93	0.000	0.049	1.31	0.254	0.004	4.16	0.042	0.012	1.59	0.209	0.005	0.33	0.565	0.001	0.00	0.982	0.000
Perceived risk	6.08	0.014	0.018	0.46	0.497	0.001	0.31	0.580	0.001	0.00	0.983	0.000	0.60	0.440	0.002	0.56	0.456	0.002

MANOVA, multivariate analysis of variance; NNPV, double negative predictive value; NPV, negative predictive value; PPV, positive predictive value; SE, sensitivity; SP, specificity; SX, symptomatic status; SX × Loc, interaction between symptomatic status and geographical location.

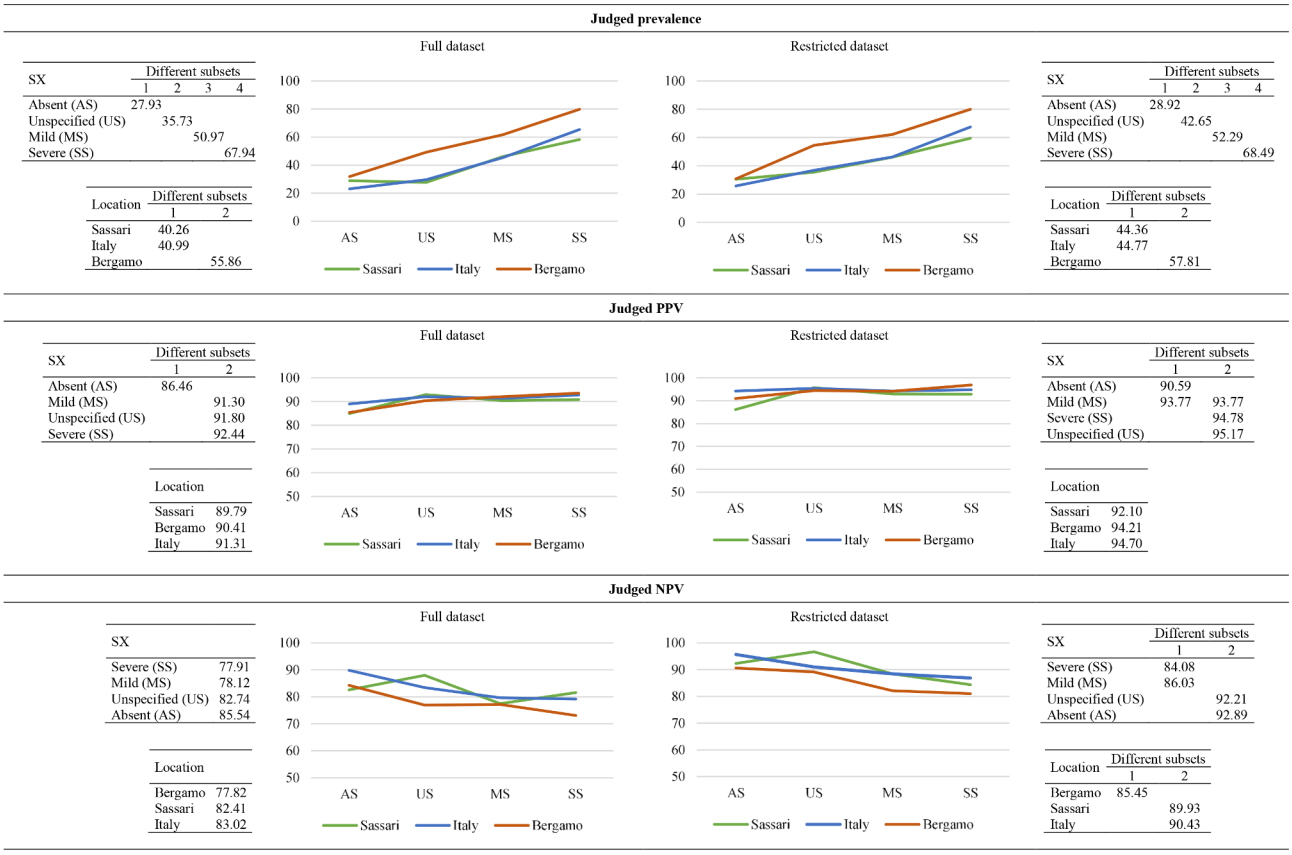

Figure 2 Prevalence, PPV and NPV judgements in the 12 experimental groups, with corresponding results of the Tukey's post hoc tests for the two independent variables (SX and location). Means in one subset significantly differ (at least p<0.05) from those in other subsets. NPV, negative predictive value; PPV, positive predictive value; SX, symptomatic status.

Informativeness of positive and negative test results

To assess whether participants were aware of the differences between the informativeness of the positive and negative results, their perceived usefulness for changing behaviour was compared using a repeated measures ANOVA, and the same was done for the perceived usefulness of repeating the test after a negative or positive result. As expected, participants evaluated positive results as more useful than negative ones for changing current behaviour (91.6, 95% CI 90.2 to 93.1 and 41.0, 95% CI 37.9 to 44.0, respectively, p<0.001). The analysis revealed a significant interaction between participants' judgements and age (p=0.038). More specifically, the difference between the perceived usefulness of positive and negative results was greater in younger (91.8, 95% CI 89.8 to 93.8 and 38.6, 95% CI 34.4 to 42.7, respectively) than in older participants (91.4, 95% CI 89.5 to 93.4 and 43.6, 95% CI 38.9 to 47.8, respectively). Yet, both evaluations (especially the one concerning a possible negative result) appear surprisingly high, given that participants were under lockdown and the tested person was assumed to be asymptomatic. Even less easily understandable is that participants found a short-term repetition of the test more useful after a positive rather than a negative

Table 3 Means of judged and expected PPV and NPV (and 95% CI) in full and restricted data sets, with corresponding P values of paired sample t-tests

	Full data set			Restricted data set		
PVs	**Judged**	**Expected**	**P value**	**Judged**	**Expected**	**P value**
PPV						
All	90.5 (89.3 to 91.6)	86.5 (84.7 to 88.3)	0.000	93.7 (92.7 to 94.6)	93.8 (92.8 to 94.9)	0.760
Prevalence ≤20%	87.0 (83.9 to 90.1)	69.8 (64.6 to 75.0)	0.000	92.1 (89.1 to 95.2)	85.2 (81.4 to 89.0)	0.002
NPV						
All	81.1 (79.0 to 83.2)	89.4 (88.0 to 90.9)	0.000	88.5 (87.0 to 90.1)	92.1 (90.7 to 93.5)	0.000
1-prevalence ≤20%	71.8 (65.1 to 78.5)	79.0 (73.8 to 84.1)	0.054	81.5 (75.5 to 87.4)	82.4 (76.6 to 88.1)	0.778

NPV, negative predicted value; PPV, positive predictive value; PVs, predictive values.

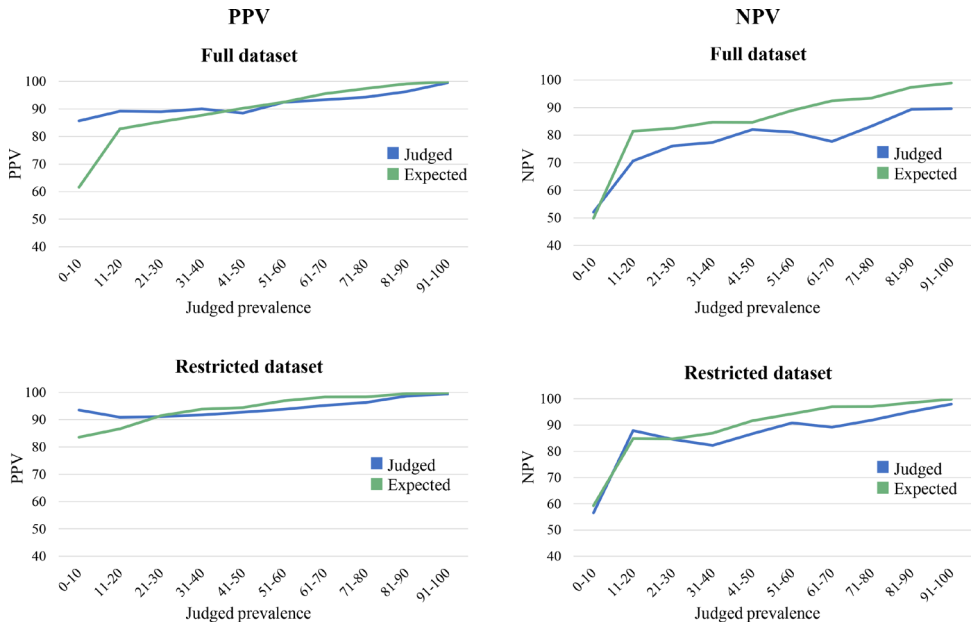

Figure 3 Judged and expected PPV (left panel) and NPV (right panel) as a function of judged prevalence, in full and restricted data sets. NPV, negative predictive value; PPV, positive predictive value.

result (62.7, 95% CI 59.6 to 65.7 and 47.2, 95% CI 44.4 to 50.0, respectively, p=0.013). These results are worth investigating in greater depth and, if confirmed in future studies, would indicate that participants do not consider some crucial behavioural implications of test results.

Causes of test errors

The Mann-Whitney tests did not reveal significant effects of age, gender and education on participants' rankings of possible causes of test errors with one exception: the technical characteristics of the test were considered to be more probable as a cause of false positives by participants with a higher education level than those with a lower education level (2.12, 95% CI 2.03 to 2.21 and 1.95, 95% CI 1.87 to 2.03, respectively, p=0.007). Wilcoxon signed-rank tests showed that participants properly distinguish between the causes of false-positive and false-negative test results (figure 4): the level of viral load was considered more likely to be the cause of false negatives, while human error and technical characteristics of the test were assessed as more likely to generate false positives (all p<0.001). These evaluations appear in line with experts' assessments.[20 34]

DISCUSSION

This study provides a first comprehensive overview of the general public's understanding of the most commonly

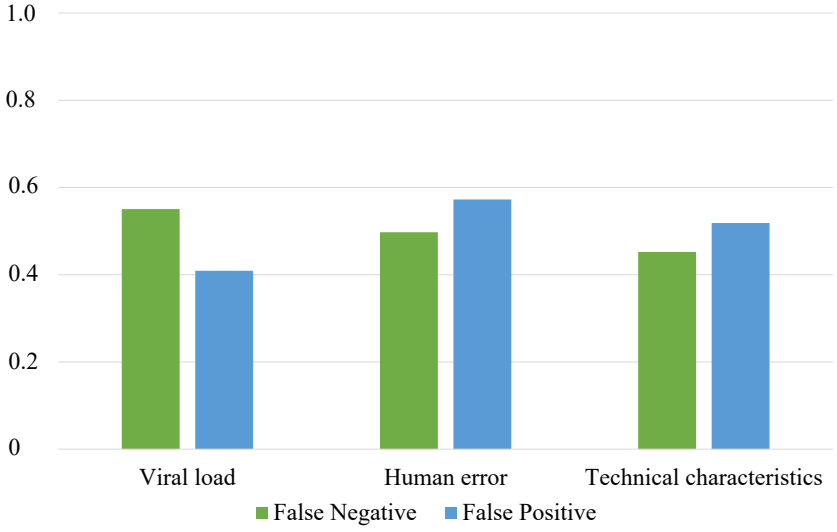

Figure 4 Ranking of causes for test errors. To display participants' rankings on a scale between 0 and 1, we assigned each cause a score from 1 (least probable) to 3 (most probable), and then normalised total scores using the MinMax normalisation method.

Box 2 Results in a nutshell

Because the (minimal) differences observed in the two data sets can be reasonably attributed to noise, we report here only the results of the restricted data set.

Participants' most relevant attitudes towards SARS-CoV-2 and COVID-19

- Older participants and those residing in the least affected area expressed greater worry about the COVID-19 pandemic.
- Older participants and those residing in the least affected area assessed a possible SARS-CoV-2 infection as more severe.
- Participants residing in the most affected area indicated a higher perceived probability of contracting the SARS-CoV-2 infection, while there was no difference between younger and older participants.
- Participants' prevalence judgements were predicted by their worry about the COVID-19 pandemic and their perceived risk of a possible SARS-CoV-2 infection.

Aspects of the test that are well understood by participants

- Fairly good qualitative coherence between judgements of prevalence, test characteristics and predictive values (PVs).
- Dependency of judged prevalence, positive predictive value (PPV) and negative predictive value (NPV) on geographic location and severity of symptoms.
- No base rate fallacy for judged NPV (when prevalence >80).
- Double negative results acknowledged both as supporting the absence of infection more than a single negative result and as not ruling out the possibility of infection.
- Estimates of test characteristics (sensitivity (SE) and specificity (SP)) compatible with those of a useful test (ie, SE+SP≥1.5) and independent of symptomatic status and geographic location.
- Higher estimates for false-negative than false-positive rates.
- Positive results evaluated as more informative than negative ones with respect to an asymptomatic person's current behaviour.
- Human error and technical characteristics of the test judged more likely causes for false positives; level of viral load for false negatives.

Participants' main errors

- Base rate fallacy for judged PPV (when prevalence ≤20).
- Judged NPV lower than expected based on judged prevalence and characteristics of the test.
- General underestimation of false-negative rate.
- General overestimation of false-positive rate.
- General overestimation of the impact of both positive and negative results on an asymptomatic person's behaviour.
- Confusion about the utility of a short-term repetition of the test after positive or negative results.

used test for detecting SARS-CoV-2. Overall, some aspects of the test appear to be fairly well grasped while others are not (for a detailed summary of the main results, see box 2). With regard to the latter, consistent with earlier research that considered different conditions and medical tests, our data show that, although laypeople are sensitive to several factors that can influence prevalence, they are not always able to integrate this information with evidence provided by test results. Our data also indicate that the estimate of the NPV can be flawed in a different way from that generally observed for the PPV. Moreover, the examination of participants' beliefs about the diagnostic accuracy of the test allowed us to document an overestimation of the false-positive rate together with an underestimation of the false-negative rate. Finally, the high behavioural impact attributed to test results in the absence of symptoms appears to be unjustified, especially in the case of negative outcomes, as is the utility assigned to a short-term repetition of the test after a positive result. Among the aspects of the current SARS-CoV-2 testing that participants best understood are: the dependence of prevalence but not of SE and SP on SX and geographical location; the evaluation of the false-negative rate as higher than the false-positive rate; and proper probabilistic ordering of causes of false-positive and false-negative test results.

As noted in the Introduction section, the findings of earlier studies cannot be extrapolated to the testing now underway. In particular, previous research has focused almost exclusively on whether participants properly calculate the PPV when explicitly provided with information about the prevalence of a condition and the diagnostic performance of a test used to detect it. Estimates of NPV or of the diagnostic accuracy of tests have not been studied, nor have the various behavioural consequences of the comprehension of test results. For the first time in literature, participants' perception of the accuracy of a medical test was explored and combined with the behavioural impact of its positive and negative results, the perceived usefulness of a short-term repetition of the test following positive or negative results, and ranking of causes for false positives and false negatives. Thanks to the wider range of measures considered, this study extends scientific knowledge of how the general public interprets test results and challenges most theoretical models that have been proposed to explain the difficulties in computing the PPV and, more generally, base rate neglect. Furthermore, our study expanded existing methodology by introducing a qualitative coherence criterion that allows documentation of the base rate fallacy, even in the absence of normative values for test characteristics and prevalence.

The main limitations of this study are the narrow age range of participants (more than 95% younger than 50 years). Another limit is that it included mainly participants who had not undergone the swab test at the time of data collection. Finally, although there is no apparent reason to expect substantial cross-national differences in the accuracy of adults' performance in these kinds of tasks,[8] the generalisation to other countries cannot be taken for granted. Future research could extend our research questions to users of different healthcare systems and, especially, to specific subsets of the population, including older adults and patients with pre-existing conditions, and even to primary care physicians or specialists,[ii] who—

[ii]A preliminary analysis on our data indicated that, overall, the judgements of participants who have a medical/health-related university degree (eg, medicine, neuroscience, biology, neurobiology, psychology; n=82) did not differ from those of all other participants in any of our dependent measures (ie, prevalence

together with public health agencies—have key roles in helping laypeople understand the reasons behind recommendations and obligations. Finally, it may also be of interest to consider other tests for SARS-CoV-2 infection (eg, rapid antigen tests or serological tests for the detection of antiviral antibodies).

Undoubtedly, mass testing plays a major role in the collection of epidemiological information and in the management of pandemics. However, it also has unavoidable effects at an individual level, as test results might well influence personal inferences and decisions. The aspects of the test that escape common understanding may indeed have a strong bearing on the public's health and well-being. For example, the systematic underestimation of the false-negative rate could well lead to neglect of precautions and, in the event of subsequent development of symptoms, to mistrust of medical services and institutions. Similarly, the disproportionate behavioural impact attributed to test results in the absence of symptoms and the confusion about the utility of a short-term repetition of the test after a positive result could give rise to overtesting, with all the serious consequences this entails.

In conclusion, certainly the dissemination of correct medical information[35] and the implementation of health literacy interventions[36] are essential for dealing with this (or any) pandemic emergency. Yet, for these policies to be truly effective, they must be grounded in empirical evidence that indicates where exactly the difficulties lie, and hopefully provides precise guidance on how to overcome them. As more than 50 years of cognitive studies on human rationality have shown, the problem is more complicated than providing laypeople with accurate information but, beyond this, encompasses comprehending how they use this information in their reasoning.

Contributors Both authors contributed equally to the design of the study. SP collected data and performed the statistical analysis. Both authors drafted the manuscript for important intellectual content, approved the final version submitted and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. KT acts as guarantor and corresponding author.

Funding The study was supported by the MIUR project "Dipartimenti di eccellenza".

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study was approved by the Human Research Ethics Committee of the University of Trento (protocol number 2019/2020-026). All participants provided informed, written consent and data obtained were fully anonymous.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data sharing: the authors support data sharing and queries in this regard can be addressed to the corresponding author.

ORCID iD

Katya Tentori http://orcid.org/0000-0002-5968-9936

rate, PVs, informativeness of a positive and negative test result and utility of short-term repetition after a positive or negative test result, all p>0.05).

REFERENCES

1 WHO. *Director-general's opening remarks at the media briefing on COVID-19*. Geneva: World Health Organization, 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---16-march-2020
2 Cheng MP, Papenburg J, Desjardins M. Diagnostic testing for severe acute respiratory syndrome–related coronavirus-2: a narrative review. *Ann Intern Med* 2020;172:726–34.
3 Peto J. Covid-19 mass testing facilities could end the epidemic rapidly. *BMJ* 2020;368:m1163.
4 Cohen J, Kupferschmidt K. Countries test tactics in 'war' against COVID-19. *Science* 2020;367:1287–8.
5 Sharfstein JM, Becker SJ, Mello MM. Diagnostic testing for the novel coronavirus. *JAMA* 2020;323:1437–8.
6 Rajendran DK, Rajagopal V, Alagumanian S, *et al*. Systematic literature review on novel corona virus SARS-CoV-2: a threat to human era. *Virusdisease* 2020;31:161–73.
7 Eddy DM. Probabilistic reasoning in clinical medicine: Problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment under uncertainty: heuristics and biases*. Cambridge, England: Cambridge University Press, 1982: 249–67.
8 Pighin S, Tentori K, Savadori L, *et al*. Fostering the understanding of positive test results. *Ann Behav Med* 2018;52:909–19.
9 Pighin S, Gonzalez M, Savadori L, *et al*. Improving public interpretation of probabilistic test results: distributive evaluations. *Med Decis Making* 2015;35:12–15.
10 Pighin S, Gonzalez M, Savadori L, *et al*. Natural frequencies do not foster public understanding of medical test results. *Med Decis Making* 2016;36:686–91.
11 Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med* 1998;73:538–40.
12 Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. *BMJ* 2006;333:284–6.
13 Kahneman D, Tversky A. On the psychology of prediction. *Psychol Rev* 1973;80:237–51.
14 Navarrete G, Mandel DR, eds. *Improving bayesian reasoning: what works and why?* Lausanne: Frontiers Media, 2016.
15 Garcia-Retamero R, Cokely ET, Hoffrage U. Visual AIDS improve diagnostic inferences and metacognitive judgment calibration. *Front Psychol* 2015;6:932.
16 Fang Y, Zhang H, Xie J, *et al*. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology* 2020;296:E115–7.
17 Kucirka LM, Lauer SA, Laeyendecker O, *et al*. Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. *Ann Intern Med* 2020;173:262–7.
18 Long C, Xu H, Shen Q, *et al*. Diagnosis of the coronavirus disease (COVID-19): rRT-PCR or CT? *Eur J Radiol* 2020;126:108961.
19 Wang W, Xu Y, Gao R, *et al*. Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA* 2020;323:1843–4.
20 Cohen AN, Kessel B. False positives in reverse transcription PCR testing for SARS-CoV-2. *MedRxiv* 2020.
21 Ren X, Liu Y, Chen H, *et al*. Application and optimization of RT-PCR in diagnosis of SARS-CoV-2 infection. *SSRN* 2020.
22 Kamae I. A coronavirus pandemic alert: massive testing for COVID-19 in a large population entails extensive errors, 2020. Available: https://cigs.canon/en/article/20200402_6324.html
23 Subramanian SV, James KS. Use of the demographic and health survey framework as a population surveillance strategy for COVID-19. *Lancet Glob Health* 2020;8:e895.

24 Day M. Covid-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ* 2020;369:m1375.
25 Goldstein ND, Burstyn I. On the importance of early testing even when imperfect in a pandemic such as COVID-19. OSF Preprints 2020.
26 European Centre for Disease Prevention and Control. Guidance for discharge and ending isolation in the context of widespread community transmission of COVID-19 – first update, 2020. Available: https://www.ecdc.europa.eu/en/publications-data/covid-19-guidance-discharge-and-ending-isolation
27 Fiorillo A, Gorwood P. The consequences of the COVID-19 pandemic on mental health and implications for clinical practice. *Eur Psychiatry* 2020;63:1–2.
28 Holmes EA, O'Connor RC, Perry VH, *et al*. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *The Lancet Psychiatry* 2020;7:547–60.
29 Mazza C, Ricci E, Biondi S, *et al*. A nationwide survey of psychological distress among Italian people during the COVID-19 pandemic: immediate psychological responses and associated factors. *Int J Environ Res Public Health* 2020;17:e3165.
30 Ashrafi-Rizi H, Kazempour Z. Information typology in coronavirus (COVID-19) crisis; a commentary. *Arch Acad Emerg Med* 2020;8:e19.
31 Palan S, Schitter C, Prolific SC. Prolific.ac—A subject pool for online experiments. *J Behav Exp Finance* 2018;17:22–7.
32 Slovic PE. *The perception of risk*. London: Earthscan publications, 2000.
33 Lowrance WW. The nature of risk. In: Schwing RC, Albers WA, eds. *How safe is safe enough*. New York: Plenum Press, 1980: 5–14.
34 Lippi G, Simundic A-M, Plebani M. Potential preanalytical and analytical vulnerabilities in the laboratory diagnosis of coronavirus disease 2019 (COVID-19). *Clin Chem Lab Med* 2020;58:1070–6.
35 Earnshaw VA, Katz IT. Educate, amplify, and focus to address COVID-19 misinformation. *JAMA Health Forum* 2020;1:e200460.
36 Paakkari L, Okan O. COVID-19: health literacy is an underestimated problem. *Lancet Public Health* 2020;5:e249–50.
37 Power M, Fell G, Wright M. Principles for high-quality, high-value testing. *Evid Based Med* 2013;18:5–10.

Supplementary material 1 - Stimuli and questions for the online behavioral experiment

The section headings (grey background) and the variable labels (in italics) were not displayed to participants. All materials are translated from Italian.
SX = Symptomatic status; PPV = Positive predictive value; NPV = Negative predictive value; NNPV = Double negative result predictive value.

INTRODUCTION
(Common to all participants.)

As you certainly know, in recent weeks, Italy has been facing a major health emergency related to the **COVID-19** pandemic.

The most frequently used test for the detection of **SARS-CoV-2**, the infection that causes COVID-19, is based on the use of a **naso/oropharyngeal swab**: a sterile swab is inserted deep inside the tested person's nose and/or throat to collect a mucus sample that is then analyzed in specialized laboratories. If the analysis detects the presence of the virus, the test result is "**positive**", while if the analysis does not detect the presence of the virus, the test result is "**negative**". The test is quite accurate, but, as with all medical tests, can produce a number of errors.

In what follows, you will be asked some questions about your perception of the accuracy and implications of this test. Please don't search for the correct answers on the web but respond based on what you already know about this test.

PART 1

Twelve experimental scenarios

(= 4 symptomatic statuses (SX) x 3 geographical locations ("Italy" corresponds to the generic location, "Bergamo" to the high-risk location, and "Sassari" to the low-risk location. Each participant was presented with only one scenario.)

Unspecified SX	Imagine a person who lives in [**Italy/Bergamo/Sassari**].
Absent SX	Imagine a person who lives in [**Italy/Bergamo/Sassari**] and who, at the moment, **does NOT show any symptoms** compatible with the SARS-CoV-2 infection that causes COVID-19 (the person does not have a temperature higher than 37.5°C, no cough or respiratory difficulties).
Mild SX	Imagine a person who lives in [**Italy/Bergamo/Sassari**] and who, at the moment, **is showing some symptoms** compatible with the SARS-CoV-2 infection that causes COVID-19 (the person has a temperature higher than 37.5°C and/or mild cough but does not have breathing difficulties).
Severe SX	Imagine a person who lives in [**Italy/Bergamo/Sassari**] and who, at the moment, **is showing several symptoms** compatible with the SARS-CoV-2 infection that causes COVID-19 (the person has a temperature higher than 37.5°C, severe cough, and breathing difficulties).

Test questions

(The experimental scenario was repeated before each of the questions below in order to make it clear that these questions were independent of each other and there was no accumulation of evidence. Questions were common to all participants and their order was the same for all participants.)

Prevalence

In your opinion, what is the probability (in %) that this person

has the SARS-CoV-2 infection? _____ (Insert a value from 0 = "impossible" to 100 = "certain")

does NOT have the SARS-CoV-2 infection? _____ (Insert a value from 0 = "impossible" to 100 = "certain")

(The sum of your responses has to equal 100)

PPV

This person is **randomly screened** for the SARS-CoV-2 infection and undergoes the **swab test**.
The **swab test** result is **positive**.
In your opinion, given that the swab test result is **positive**, what is the probability (in %) that she

has the SARS-CoV-2 infection? _____ (Insert a value from 0 = "impossible" to 100 = "certain")

does NOT have the SARS-CoV-2 infection? _____ (Insert a value from 0 = "impossible" to 100 = "certain")

(The sum of your responses has to equal 100)

NPV

This person is **randomly screened** for the SARS-CoV-2 infection and undergoes the **swab test**.
The **swab test** result is **negative**.
In your opinion, given that the swab test result is **negative**, what is the probability (in %) that she

has the SARS-CoV-2 infection? _____ (Insert a value from 0 = "impossible" to 100 = "certain")

does NOT have the SARS-CoV-2 infection? _____ (Insert a value from 0 = "impossible" to 100 = "certain")

(The sum of your responses has to equal 100)

NNPV

This person is **randomly screened** for the SARS-CoV-2 infection and undergoes the **swab test twice** (at a 24-hour interval).
In both cases, the **swab test** result is **negative**.
In your opinion, given that both the swab test results are **negative**, what is the probability (in %) that she

has the SARS-CoV-2 infection? _____
(Insert a value from 0 = "impossible" to 100 = "certain")

does NOT have the SARS-CoV-2 infection? _____
(Insert a value from 0 = "impossible" to 100 = "certain")

(The sum of your responses has to equal 100)

Sensitivity

This person **has** the **SARS-CoV-2** infection.
This person is **randomly screened** for the SARS-CoV-2 infection and undergoes the **swab test**.
In your opinion, given that she **has** the SARS-CoV-2 infection,
what is the probability (in %) that her **swab test** is

positive? _____
(Insert a value from 0 = "impossible" to 100 = "certain")

negative? _____
(Insert a value from 0 = "impossible" to 100 = "certain")

(The sum of your responses has to equal 100)

Specificity

This person **does NOT have** the **SARS-CoV-2** infection.
This person is **randomly screened** for the SARS-CoV-2 infection and undergoes the **swab test**.
In your opinion, given that she **does NOT have** the SARS-CoV-2 infection,
what is the probability (in %) that her **swab test** is

positive? _____
(Insert a value from 0 = "impossible" to 100 = "certain")

negative? _____
(Insert a value from 0 = "impossible" to 100 = "certain")

(The sum of your responses has to equal 100)

General worry

In general, how **worried** are you about the COVID-19 pandemic? _____
(Insert a value from 0 = "not worried at all" to 100 = "extremely worried")

Individual probability

In your opinion, how **likely** is it that, within the next 2 months, you will contract the SARS-CoV-2 infection that causes COVID-19? _____
(Insert a value from 0 = "impossible" to 100 = "certain")

Individual severity

Imagine contracting, within the next 2 months, the SARS-CoV-2 infection that causes COVID-19. How **severe** would you consider it? _____
(Insert a value from 0 = "not severe at all" to 100 = "extremely severe")

PART 2

Test questions

(Common to all participants. The order of the questions was randomized across participants.)

Informativeness of a positive result

Imagine that a person who **does NOT show any symptoms compatible with COVID-19** undergoes a swab test during a screening for the SARS-CoV-2 infection that causes COVID-19.

The **swab test** result is **positive**.

To what extent, in your opinion, does this result provide this person with information **useful in changing her behaviour** (compared to that currently in place)? _____
(Insert a value from 0 = "not informative at all" to 100 = "extremely informative")

To what extent, in your opinion, would it be **useful to repeat** the **swab test** on this person again in 1-2 days? _____
(Insert a value from 0 = "not useful at all" to 100 = "extremely useful")

Informativeness of a negative result

Same as above except for the sentence "The **swab test** result is **negative**"

PART 3

Test questions

(Common to all participants. The order of the questions and of the response options were randomized across participants.)

Explanation of false positives

Imagine that a person who **does NOT have** the **SARS-CoV-2** infection undergoes the swab test. A health care professional collects the sample and brings it to the laboratory, where the sample is analyzed. The result of the swab test is **positive**.

The **causes** of this **testing error** (which is technically called a "**false positive**") can vary, and below you will find some listed. Rank these possible causes from the one that you consider the most probable ("1") to the one you that consider the least probable ("3"):

___ **human error** during the sample collection, transport or analysis (such as insufficient material, improperly stored material, and/or cross-contaminated material);

___ **technical characteristics of the test**, i.e. the specific methods used to analyze the sample and the criteria adopted to interpret the results of these analyses;

___ **level of the viral load**, i.e. the amount of virus in the tested person's nose and/or pharynx at the time of the sample collection.

Explanation of false negatives

Same as above except for the sentences "Imagine that a person who **has** the **SARS-CoV-2** infection", "The **swab test** result is **negative**", and **"**The **causes** of this **testing error** (which is called a "**false negative**") …"

PART 4

Demographic information
(Obtained from Prolific.ac)

Questions concerning participant's experience with the swab test and COVID-19
(Common to all participants. The order of the questions was the same for all participants.)

Underwent swab test

Have you undergone the swab test for SARS-COV-2? Y/N

Friend/relative/colleague underwent swab test

Has someone in your circle (i.e., relatives, friends, colleagues) undergone the swab test for SARS-COV-19? Y/N

Friend/relative/colleague had COVID-19

Has anyone in your circle (i.e., relatives, friends, colleagues) been diagnosed with COVID-19? Y/N

Supplementary material 2 - Mean judgments of prevalence, PVs, and test characteristics (and 95% CI) in full and restricted datasets for the 12 experimental groups.
PPV = Positive predictive value; NPV = Negative predictive value; NNPV = Double negative result predictive value; SE = Sensitivity; SP = Specificity

Groups		Participant judgments											
		Prevalence (%)		PPV (%)		NPV (%)		NNPV (%)		SE (%)		SP (%)	
		Full dataset	Restricted dataset	Full dataset	Restricted dataset	Full dataset	Restricted dataset	Full dataset	Restricted dataset	Full dataset	Restricted dataset	Full dataset	Restricted dataset
Unknown SX													
	Sassari	27.7 (21.4–34.1)	35.5 (27.5–43.5)	92.9 (89.7–96.2)	95.7 (93.4–97.9)	88.0 (81.9–94.2)	96.7 (95.2–98.1)	94.0 (88.5–99.6)	99.6 (99.3–100)	89.8 (84.0–95.5)	96.6 (94.5–98.8)	95.8 (93.6–98.0)	97.8 (96.4–99.1)
	Italy	29.7 (23.6–35.7)	36.9 (29.2–44.6)	92.1 (88.1–96.1)	95.4 (92.7–98.2)	83.4 (75.7–91.1)	91.0 (87.0–95.0)	93.1 (87.7–98.4)	97.6 (95.5–99.7)	86.3 (79.7–92.9)	95.7 (92.9–98.5)	89.6 (83.8–95.4)	95.5 (91.9–99.0)
	Bergamo	49.3 (42.2–56.3)	54.5 (45.7–63.2)	90.4 (86.8–94.0)	94.5 (92.3–96.6)	77.0 (68.9–85.0)	89.2 (84.4–94.0)	92.4 (87.5–97.4)	98.1 (97.0–99.1)	91.5 (88.0–95.0)	95.6 (93.4–97.8)	87.5 (81.5–93.6)	95.8 (93.7–97.8)
	Total	35.7 (31.7–39.7)	42.7 (37.7–47.6)	91.8 (89.8–93.8)	95.2 (93.8–96.5)	82.7 (78.5–86.9)	92.2 (90.0–94.4)	93.2 (90.2–96.2)	98.4 (97.7–99.2)	89.2 (86.2–92.3)	96.0 (94.6–97.3)	91.0 (88.1–93.8)	96.3 (95.0–97.7)
Absent SX													
	Sassari	28.8 (23.2–34.5)	30.5 (23.1–37.8)	85.0 (80.5–89.4)	86.1 (79.9–92.3)	82.6 (75.4–89.8)	92.3 (89.3–95.3)	89.1 (81.6–96.6)	98.6 (97.9–99.4)	89.0 (84.4–93.6)	92.2 (88.6–95.7)	93.7 (91.1–96.4)	96.0 (94.7–97.4)
	Italy	23.1 (17.8–28.4)	25.7 (19.6–31.8)	89.0 (84.3–93.7)	94.2 (91.3–97.1)	89.8 (84.7–94.9)	95.7 (94.0–97.3)	93.1 (87.5–98.7)	99.2 (98.6–99.8)	85.4 (78.0–92.7)	95.8 (94.0–97.7)	94.2 (91.7–96.7)	97.6 (96.2–98.9)
	Bergamo	31.9 (26.3–37.4)	30.8 (23.5–38.0)	85.4 (80.4–90.5)	91.0 (87.6–94.4)	84.3 (78.2–90.3)	90.6 (86.2–94.9)	93.1 (88.3–97.8)	97.4 (94.4–100)	90.1 (85.9–94.4)	95.0 (92.8–97.3)	87.5 (80.9–94.2)	96.4 (94.7–98.2)
	Total	27.9 (24.8–31.1)	28.9 (25.1–32.8)	86.5 (83.8–89.1)	90.6 (88.1–93.1)	85.5 (82.0–89.1)	92.9 (91.1–94.7)	91.7 (88.3–95.2)	98.4 (97.4–99.4)	88.2 (85.0–91.3)	94.4 (92.9–95.9)	91.8 (89.3–94.3)	96.7 (95.9–97.5)
Mild SX													
	Sassari	46.1 (40.0–52.2)	46.1 (38.9–53.2)	90.4 (85.7–95.2)	93.0 (89.0–96.9)	77.5 (68.8–86.2)	88.5 (81.7–95.2)	91.2 (84.1–98.4)	97.5 (95.7–99.3)	88.4 (83.0–93.9)	93.9 (90.2–97.6)	92.7 (89.5–95.9)	95.5 (93.3–97.7)
	Italy	45.3 (37.9–52.7)	46.3 (36.1–56.5)	91.4 (87.8–94.9)	94.3 (91.2–97.3)	79.7 (72.0–87.4)	88.4 (81.9–94.9)	91.0 (85.1–96.9)	97.5 (95.5–99.4)	88.6 (82.6–94.7)	95.8 (93.0–98.6)	92.9 (88.5–97.3)	97.2 (95.2–99.2)
	Bergamo	61.7 (57.0–66.5)	62.2 (56.8–67.6)	92.1 (89.0–95.2)	94.1 (91.6–96.7)	77.2 (69.4–84.9)	82.1 (76.4–87.9)	91.8 (86.3–97.2)	95.0 (92.6–97.4)	91.3 (87.5–95.1)	93.6 (91.2–96.1)	94.7 (92.1–97.4)	95.8 (93.6–98.0)
	Total	51.0 (47.3–54.7)	52.3 (47.9–56.7)	91.3 (89.1–93.5)	93.8 (92.0–95.6)	78.1 (73.6–82.6)	86.0 (82.5–89.6)	91.3 (87.8–94.3)	96.6 (95.3–97.8)	89.5 (86.5–92.4)	94.3 (92.6–96.0)	93.4 (91.5–95.4)	96.1 (94.9–97.3)
Severe SX													
	Sassari	58.3 (52.1–64.4)	59.4 (53.0–65.9)	90.9 (85.9–95.9)	92.9 (89.0–96.8)	81.5 (75.3–87.8)	84.4 (78.2–90.5)	92.9 (87.9–97.9)	97.3 (95.6–99.0)	90.3 (86.9–93.7)	91.5 (88.0–95.1)	90.8 (85.4–96.1)	94.9 (92.3–97.5)
	Italy	65.4 (59.0–71.9)	67.5 (59.5–75.5)	92.8 (89.8–95.8)	94.8 (92.4–97.3)	79.2 (71.2–87.2)	86.8 (80.8–92.9)	90.9 (85.0–96.7)	96.6 (94.2–99.0)	92.1 (89.1–95.1)	92.9 (89.5–96.3)	90.5 (85.3–95.7)	94.0 (90.9–97.1)
	Bergamo	79.8 (75.7–83.9)	79.9 (74.9–85.0)	93.6 (90.3–96.9)	96.9 (95.2–98.5)	73.1 (64.4–81.8)	81.0 (73.0–89.0)	89.8 (82.7–96.8)	97.8 (96.6–99.0)	92.9 (89.8–96.0)	95.6 (93.5–97.7)	88.3 (81.8–94.8)	95.0 (91.8–98.2)
	Total	67.9 (64.4–71.4)	68.5 (64.5–72.5)	92.4 (90.2–94.6)	94.8 (93.1–96.5)	77.9 (73.5–82.3)	84.1 (80.3–87.9)	91.2 (87.8–94.6)	97.2 (96.2–98.3)	91.8 (90.0–93.6)	93.2 (91.4–95.0)	89.8 (86.6–93.1)	94.6 (93.0–96.3)
TOTAL		45.7 (43.6–47.9)	49.2 (46.5–51.8)	90.5 (89.3–91.6)	93.7 (92.7–94.6)	81.1 (79.0–83.2)	88.5 (87.0–90.1)	91.9 (90.2–93.5)	97.6 (97.1–98.1)	89.7 (88.3–91.0)	94.4 (93.6–95.2)	91.5 (90.2–92.8)	95.9 (95.2–96.5)