finding

# Graphs with logarithmic axes distort lay judgments

William H. Ryan & Ellen R. K. Evers

## Methods & Analysis

All code, materials, and data necessary to replicate all of our studies and results can be found on our OSF page at https://osf.io/zqut5/?view_only=3aa66d592dd2495ca508b4fa8729381a. Please refer to the scales within the surveys presented there for full details on the measures mentioned here.

Additionally, we did not conduct extensive analysis using a number of different moderators that were collected, including a range of demographics variables, due to time constraints. We encourage interested researchers to investigate possible moderation effects within our existing data set.

### General analysis notes:

All data cleaning, visualization, and analysis was carried out in R. For analysis, the R package lme4 was used to run mixed-effects regressions, and the package lmerTest was used to test the significance of mixed-effects model coefficients. The package coin was used to carry out all permutation tests using its "independence_test()" function. The base R package stats was used to carry out linear regressions, *t* tests, and logistic regressions. The R package MASS was used to carry out ordered logistic regressions. The R package brms was used to carry out Bayesian mixed-effects models, and the R package rstanarm was used for other Bayesian analyses. To ensure those replicating the code use the same versions of the packages, the R package checkpoint was used for each script.

## Main Article Studies 1A, 1B, 2, & 3

### Study 1A: Predictions from real case data

*Methods.*

We presented participants with graphs of the total number of coronavirus cases up until the present for four countries per participant. The United States was always presented first, then Singapore, Japan, and Poland in random order.

Conditions: There was one between-subjects condition: whether the graph had a logarithmic or linear *y*-axis.

*Materials.*

Full materials can be found on our OSF page.

All measures were asked for every country.

Predictions of future cases: The predicted number of cases one, three, five, and 10 days from the present.

Change in growth rate over the next week: A multiple-choice question asking if they believed that the growth rate of coronavirus cases would increase, decrease, or stay the same over the next week.

We collected the following moderators:

- Amount of time spent learning about the coronavirus in the past week (1–5, on a scale ranging from *No time at all to an extremely large amount of time)*
- Partisan lean (1–6, on a scale ranging from *Extremely conservative to Extremely liberal; o*ne scale point was missing due to experimenter error)
- Gender
- Age
- Size of town they live in
- Ability to work from home
- Highest level of education completed

*Data.*

All data collected, as well as the data on actual number of cases used to determine accuracy, can be found on our OSF page.

Two hundred sixty-six Mechanical Turk workers (*M* age = 38 years; 40% were female) completed the survey in exchange for monetary payment.

*Analysis & results.*

*F*or the analysis reported in the article, we first winsorized the predicted cases data at the 5th and 95th percentiles for each prediction question (that is, Poland 1 Day Forward, Poland 3 Day Forward, and so on). We then calculated the mean absolute error of each winsorized prediction for each participant relative to the actual number of cases on that day. We then ran a linear mixed-effects regression with a dependent variable of the absolute error of the predictions, independent variable of log condition, and a covariate of the number of days in the future the prediction was being made, with random intercepts by country. Log condition was found to significantly increase the absolute error, $\beta = 13976$, $t(3.2) = 3.2$, $p < .002$.

Using the preregistered analysis, we winsorized predictions 2.5 standard deviations above and below the mean for each prediction question. We then carried out *t* tests and permutation tests comparing the prediction questions between conditions. Using this analysis, there was no significant relationship between predictions and the condition variable (all *ps* > .05). However, we do not believe this analysis is the most appropriate given the data. The standard deviation of the data was large enough that winsorization excluded almost no data, including clear outliers. Additionally, carrying out individual tests on each question treats them as between-subject measures, reducing power compared with the mixed-effects model we report in the main text, which takes full advantage of the repeated within-subject measures.

In both cases, there were no significant relationships between the variables when the squared error, instead of absolute error, was used (all *ps* > .05). This transformation significantly magnified the effects of outlier data points, and we believe provides a less accurate test of the hypothesis.

Across both conditions, participants generally underpredict growth, as is shown in Figure

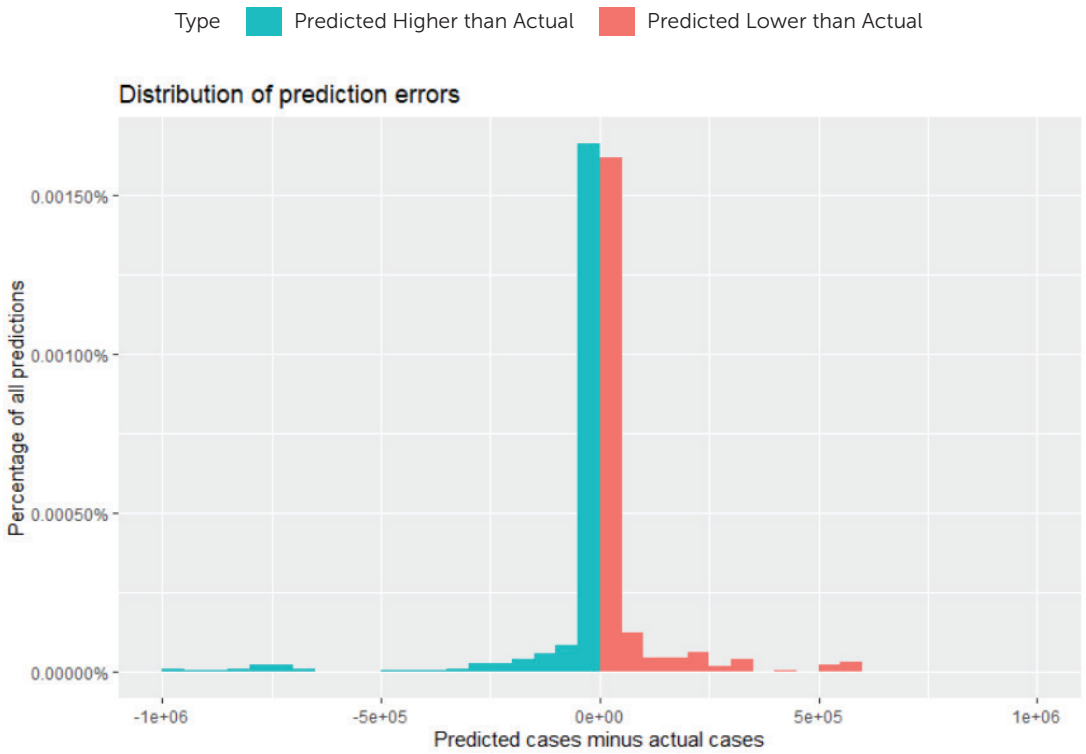
S1, which shows a histogram of participants' predictions minus the actual case counts: bars to the left of zero are underpredictions, those to the right are overpredictions.[1] We find that across both conditions, participants are much more likely to underestimate the number of future cases (77% of participants' predictions) than overestimate it (23% of participant's predictions).[2] Averages of all participants' estimates also tend to be underestimates, although these data are noisier, as shown in Table S1. The fact that most participants' predictions tend to be underestimates rather than overestimates implies that if people are making judgments of threat and growth of COVID-19 on the basis of these estimates, it is likely that individuals generally underrate the threat of COVID-19 across both conditions.

In addition, we carried out a number of analyses of between-conditions differences in accuracy using alternative operationalizations of accuracy other than mean absolute error.[3] These were intended to more directly test for biases in individual's predictions, as opposed to variance, which is what was primarily captured in our mean absolute error measure. For example, do individuals in one condition overestimate the number of cases more frequently than in another condition? The dependent variables tested here, such as the prediction error of individual's predictions (that is, their predicted number of cases minus the actual number of

Table S1. This table shows the mean of the participant's predicted number of cases minus the actual number for each day, as well as the standard deviation in parentheses for Study 1A.

| Days forward | Mean predicted case count – actual case count (*SD*) |
|---|---|
| 1 | –31,200 (125,265) |
| 3 | –28,667 (136,194) |
| 5 | –32,281 (153,298) |
| 10 | –45,635 (297,560) |

Figure S1. This plot is a histogram of predicted cases minus actual cases for Study 1A




*Note.* Underpredictions are colored blue, overpredictions are red.

cases), generally had higher variance than the mean absolute error measure used previously, meaning this experiment may be underpowered to detect an effect on the variables.

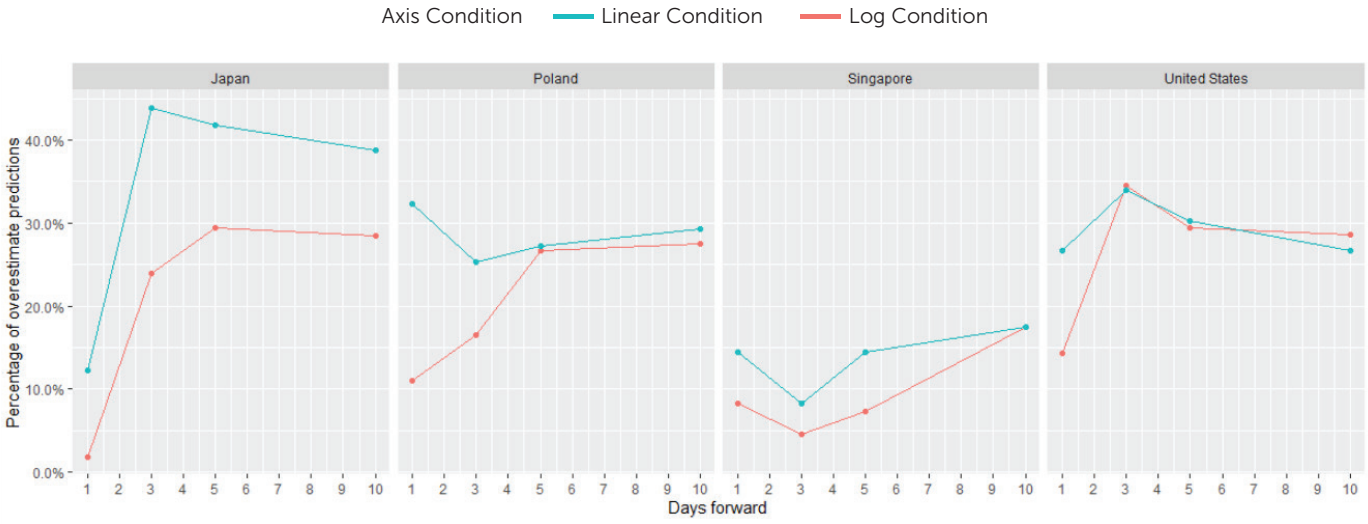
The first analysis we carried out was investigating how frequently participants' predictions were under- or overestimates in each condition. We coded each participant's prediction as either an under- or overestimate and then plotted the proportion of overestimates by condition, country, and number of days into the future the prediction was being made. We found a pattern wherein the log condition was more likely to be an overestimate for predictions closer to the present, but as predictions began to be about days farther into the future, the number of overestimates became much more equal between conditions, with the linear condition number of overestimates increasing to be more equal to that of the logarithmic condition. This pattern is shown in Figure S2. We fit a logistic regression with a dependent variable of whether each prediction was an under- or overestimate and independent variables of axis condition, number of days in the future of the prediction (days forward), the interaction between those two terms, and the target country. This regression finds that the logarithmic axis condition leads to a higher chance of overestimates, an increased number of days forward leads to a higher chance of overestimates, and the interaction between the logarithmic axis condition and an increased number of days forward leads to a higher chance of underestimates (see Table S2).

Table S2. Logistic regression results, predicting whether a participant's prediction is an overestimate, using independent variables of logarithmic axis condition and days forward, as well as their interaction. Not reported here, but included in the regression, are country-level fixed effects.

| Variable | β | *z* | *p* |
|---|---|---|---|
| Logarithmic axis | 0.77 | 5.13 | < .0001 |
| Days forward | 0.09 | 5.66 | < .0001 |
| Interaction of Logarithmic Axis and Days forward | −0.07 | −2.9 | .004 |

Figure S2. This figure shows the percentage of all participant predictions that are overestimates by the number of days forward from the present, split by axis condition and plotted separately for each county.

Second, we tested whether one condition generally leads to higher or lower prediction errors than the other—that is to say,



does one condition lead to greater under- or overestimates on average? To test this, we ran a mixed-effects model with a dependent variable of the participant's prediction minus the actual case count, and independent variables of the axis condition and day, as well as their interaction, with varying random intercepts by country. We also fit this model without the interaction term. In both cases, the effect of condition was not significant (all *ps* > .05).

Third, we compared the average predictions in each condition to determine whether one condition generally led to higher predictions than the other. Our primary test of this difference was with a mixed-effects model with a dependent variable of participants' predictions and independent variables of the axis condition, number of days into the future the prediction was being made, and their interaction, and we allowed random intercepts by country. We found no significant effect of condition or significant interaction between condition and day (all *ps* > .05), although, as expected, predictions were generally higher for predictions made later on, $\beta$ = 6,325, $t(3389) = 8.078$, $p < .0001$, reflecting participants' very reasonable belief that total cases were increasing over time. The effects of condition and the interaction between condition and day remained nonsignificant when a nonhierarchical model was used, and the individual effect of condition was not significant when either mixed-effects or nonhierarchical models were fit without an interaction term between condition and the number of days forward (all *ps* > .05).

Fourth, we examined the differences in changes in growth rate between each condition: essentially, do participants in one condition appear to believe that the growth rate is larger than participants in the other condition believe it to be? To do this, we calculated the daily growth rate in cases implied by each of the participant's predictions by interpolating between each of their predictions after the first. For example, to calculate the growth rate implied by their prediction one day forward versus three days forward, we subtract the predicted number of cases one day forward from the number of cases three days forward and then multiply it by 2 (the number of days between these predictions) to get the inferred daily growth rate. We then ran a mixed-effects model with a dependent variable of daily growth rate and independent variables of condition and the number of days in the future the prediction was being made, with random intercepts by country. We found no significant effect of condition ($p > .05$), although we did find a significant positive effect of the number of days in the future, implying that participants believed the growth rate would be increasing over time, $\beta$ = 16,942, $t(2541) = 7.654$, $p < .0001$. We also ran this mixed-effects model with an interaction term between condition and the number of days in the future the prediction was being made, allowing us to test for differences in beliefs about changes *in* growth rate. We again found that neither the condition nor the interaction term was significant (all *ps* > .05).

We also repeated our primary accuracy analyses using Bayesian statistics.[4] For this analysis, we used the R package brms to run a Bayesian mixed-effects model, using the default weakly informative priors given by the package. All code and details on models can be found on the OSF page. We ran a Bayesian mixed-effects regression with a dependent variable of the absolute error of the predictions, an independent variable of log condition, and a covariate of the number of days in the future the prediction was being made, with intercepts allowed to vary by country. The posterior distribution of the logarithm condition indicated that this condition increased the absolute error of predictions (posterior median = 14,119 [5,856 , 22,710]). There were two divergent transitions after warm-up. To help double-check model validity, we also ran a nonhierarchical Bayesian regression using the R package rstanarm with absolute error as a dependent variable, log condition as an independent variable, and country and day as covariates. This posterior distribution of the log condition coefficient in this regression also indicated that those in the log condition had greater absolute error in their predictions.

## Study 1B: Predictions from hypothetical data

### *Methods.*

We presented participants with graphs of the total number of coronavirus cases up until the present for one hypothetical country, whose number of cases was generated by a known exponential function. In the hypothetical country, the outbreak had been going on for 25 days.

Conditions: There was one between-subjects condition, if the graph had a logarithmic or linear *y*-axis.

### *Materials.*

Predictions of future cases: The predicted number of cases one, three, five, and 10 days from the present, as well as an estimate of the current number of cases.

Change in growth rate over the next week: A multiple-choice question asking if they believed that the growth rate of coronavirus cases would increase, decrease, or stay the same over the next week.

Growth rate: A 7-point Likert-type scale ranging from *Stay the same* to *Grow extremely quickly.*

Perceived threat: A 7-point Likert-type scale ranging from *Not at all dangerous* to *Extremely dangerous.*

Government action: 7-point Likert-type scale from *Disagree strongly* to *Agree strongly* asking agreement with "The Country should ban public gatherings, close non-essential businesses, and ask all citizens stay at home unless they are going to work or carrying out necessary errands."

Personal action: A 7-point Likert-type scale asking how they believe citizens of this country should adjust their level of effort taken to combat the virus, ranging from *Do much less than they are now* to *Do much more than they are now.*

Objective Numeracy Scale: We used a 10-question scale derived from that of Lipkus, Samsa, and Rimer (2001). Scale questions can be found in the materials. Each correct answer was scored as 1 point. The total objective numeracy score for an individual was their total number of points.

Subjective Numeracy Scale: We used an eight-question scale derived from that of Zikmund-Fisher, Smith, Ubel, and Fagerlin (2007). Question 7 was reverse-coded, and then the mean of answers Likert-type scale questions was used to calculate a subjective numeracy score for each individual. Higher scores indicated higher subjective numeracy.

We collected the following additional moderators:

- Amount of time spent learning about the coronavirus in the past week (1–5)
- Partisan lean (1–7)
- Gender
- Age
- Size of town they live in
- Ability to work from home
- Highest level of education completed

### *Data.*

All data collected, as well as the data used to create the hypothetical disease graph and validate predictions, can be found on our OSF page.

Four hundred three Mechanical Turk workers (*M* age = 36.7 years; 37% were female) completed the survey in exchange for monetary payment.

### *Analysis & results.*

For the analysis reported in the article, we first winsorized the predicted cases data at the 5th and 95th percentiles for each prediction question (that is, 1 Day Forward, 3 Days Forward, and so on). We then calculated the mean absolute error of each winsorized prediction for each participant relative to the actual number of cases on that day as dictated by the mathematical function that created the underlying data. We then ran a linear regression with a dependent variable of the absolute error of the predictions, an independent variable of log condition, and a covariate of the number of days in the future

the prediction was being made. Results of the regression indicated a significant collective relationship between the variables and absolute error, $F(2, 1956) = 522$, $p < .00001$. Log condition was found to significantly increase the absolute error, $\beta = 901$, $t(1956) = 2.098$, $p = .0361$.

When analysis was completed consistent with the preregistered plan of analysis, which was the same exclusions and tests as in Study 1A, no individual *t* test or permutation test was significant (all *ps* > .07). However, we believe that these exclusion criteria and the plan of analysis were incorrect for similar reasons as in Study 1A. The initial winsorization at 2.5 standard deviations above or below the mean excluded almost no observations, including clear outliers, and the planned analysis again was lower powered because it analyzed the partially within-subjects design as though it were fully between-subjects.

Numeracy analysis was conducted for objective and subjective numeracy measures. These analyses were exploratory. A linear regression of absolute error on the number of days forward for the prediction and objective numeracy scores found that objective numeracy predicted significant decreases in absolute error as objective numeracy scores increased, $\beta = -1585$, $t(1943) = -27.95$, $p = < .00001$. Linear regressions were conducted predicting the exploratory Likert measures of growth, danger, policy, and effort ratings using the interaction of objective numeracy and log condition and their simple effects. Objective numeracy had a consistent relationship with scale ratings. Danger, policy, and effort ratings were higher with higher objective numeracy (all *ps* < .01). Danger and policy further had an interaction between objective numeracy score and the log condition such that higher objective numeracy scores resulted in higher relative judgments in the log condition versus linear condition (all interaction *ps* < .01)

A linear regression of absolute error on the number of days forward for the prediction and subjective numeracy scores found that subjective numeracy predicted significant decreases in absolute error as subjective numeracy scores increased, $\beta = -647$, $t(1946) = -3.17$, $p = .001$. Similar linear regression models predicting Likert ratings on the basis of the interaction between subjective numeracy and log condition, as well as the simple effects of each, found that subjective numeracy score increases predicted higher growth, effort, and danger ratings (all *ps* > .02). However, there were no significant interaction terms (all *ps* > .05).

We additionally tested for overall under- or overestimates across both conditions and found that more of participant's predictions were underestimates (72%) than overestimates (28%) across both conditions.[5] Their average estimates were also lower than the actual predicted number of cases based on the graph's growth, as shown in Table S3. This replicates our finding in our appendix analysis of Study 1A.

In addition, we carried out a number of analyses of between-condition differences in accuracy using alternate specifications other than mean absolute error, which were intended to directly test for bias (that is, over- or underestimation) as opposed to variance, which our mean absolute error measure tested for. We carried out four analyses similar to those in Study 1A. First, we found a pattern wherein the logarithmic axis generally led to fewer overestimates and thus more underestimates. However, as the predictions went farther into the future, the differences between conditions appeared to decrease. We ran a logistic regression with a binary dependent variable (if a given prediction was an overestimate) and independent variables of the number of days into the future the prediction was being made, the logarithmic axis condition, and the interaction of those two terms. This regression found that the logarithmic axis condition

Table S3. This table shows the mean of the participants' predicted number of cases minus the actual number for each day, as well as the standard deviation in parentheses for Study 1B.

| Days forward | Mean predicted case count – actual case count (*SD*) |
|---|---|
| 1 | –3,406 (6,442) |
| 3 | –3,219 (7,824) |
| 5 | –4,439 (10,894) |
| 10 | –12,725 (24,623) |

Figure S3. This figure shows the percentage of all participant predictions that are overestimates by the number of days forward from the present, split by axis condition for Study 1B.

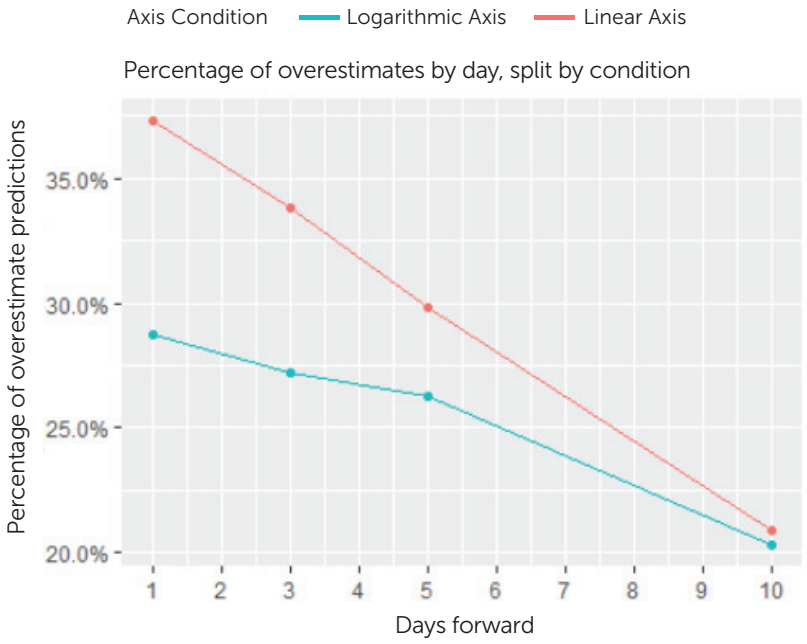



Table S4. Logistic regression results, predicting if a participant's prediction is an overestimate, using independent variables of logarithmic axis condition and days forward, as well as their interaction.

| Variable | β | z | p |
|---|---|---|---|
| Logarithmic axis | −0.41 | −3.15 | .001 |
| Days forward | −0.09 | −3.74 | .027 |
| Interaction of Logarithmic Axis and Days forward | 0.039 | 1.148 | .25 |

reduced the chances of an overestimate, as did increasing days in the future. However, the interaction between the two terms was not significant. This regression is summarized in Table S4, and overall overestimation is visualized in Figure S3. Notably, these results fairly directly contradict the results of a similar analysis carried out in Study 1A, where overestimates were more common in the logarithmic axis condition. This seems to suggest that some differences in the stimuli presented beyond axis condition matter for participants' predicted estimates—more research is needed to fully identify these factors. It is also possible that because Study 1A used real data instead of data inferred from the exact function presented to participants, some part of participant error in Study 1A was due to things like changes in how countries counted cases or new actions taken by countries, which made their case counts diverge from previous trends.

Second, we examined differences in prediction errors between conditions, essentially comparing the predicted values minus actual values for each prediction between the different axis conditions. We ran a linear regression with an independent variable of the prediction error (predicted case count minus actual case count) and dependent variables of logarithmic axis condition and the number of days in the future the prediction was made for. We found no significant effect of axis condition ($p > .05$), although number of days in the future did predict generally lower prediction errors, that is, greater underestimates, $\beta = -1{,}079$, $t(1609) = 15.01$, $p < .0001$.

Third, we compared the average predictions in each condition to determine if one condition generally led to higher predictions than the other. We ran a linear model with a dependent variable of the prediction and an independent variable of axis condition and day and found no significant effect of axis condition ($p > .05$), although number of days in the future did predict generally higher predictions, $b = 1{,}769$, $t(1609) = 15.01$, $p < .0001$.

Fourth, we examined the differences in changes in growth rate between each condition, as in Study 1A. We ran two linear models, with a dependent variable of the inferred daily growth rate at the time of the prediction, and independent variables of the axis condition and the number of days in the future the prediction was being made. In one model we also included the interaction term between these two independent variables, to test for differences in beliefs about changes in growth rate. As in Study 1A, in both regressions, the logarithmic axis condition did not significantly predict growth rates (all *ps* > .05), and the number of days forward predicted higher growth rates (all *ps* < .0001).

Finally, this study allowed us to test whether differences in predictions being over- or underestimates predicted participants' Likert scale ratings of growth, threat, need for policy action, and need for changes in individual efforts, as both sets of data were collected. We calculated the mean prediction error (predicted – actual case count) across all predictions for each individual and then used this error to predict their answers to the Likert questions. We found that growth was not significantly predicted by error, b = 0.000004, $t(401) = 0.871$, $p = .38$. Higher mean prediction errors (that is, errors that indicated more overestimation) were marginally related to higher threat ratings, b = 0.000009, $t(401) = 1.82$, $p = .06$, and significantly predicted higher judgments of need for policy action, b = 0.000013, $t(401) = 2.727$, $p = .007$, and need for increased individual action, b = 0.00001, $t(401) = 2.208$, $p = .027$. Taken together, these results appear to imply that individuals who judge the number of cases as increasing more rapidly tend to view there as being a greater need for policy response. However, this analysis does not allow us to test additional moderators of this effect, such as the differing effects of predictions of increasing growth rates versus predictions of higher case counts in general.

We also repeated our primary accuracy analyses using Bayesian statistics. For this analysis, we used the R package rstanarm. We ran a linear Bayesian regression with absolute error as a dependent variable, log condition as an independent variable, and day as a covariate. This posterior distribution of the log condition coefficient in this regression also indicated that those in the log condition had greater absolute error in their predictions.

## Study 2: Beliefs & attitudes

### *Methods.*

We presented participants with graphs of the number of coronavirus cases in a number of countries from the time they first hit 100 cases up to the present.

Participants were shown four of these graphs. The first was always the United States; the remainder were three additional countries with subjectively similar-looking logarithmic graphs presented in random order.

There were two between-subjects conditions:

Log vs. linear condition: We varied whether the axis of the graph was logarithmic or linear.

Context condition versus single: We varied whether the target country individuals made judgments about was presented alone or on a graph that also showed the data for the ten countries with the highest total coronavirus case counts in addition to the data for the target country.

### *Materials.*

We elicited four main dependent variables using seven questions.

Growth rate: A 7-point Likert-type scale asking how participants expect growth rates to change, from *decrease significantly* to *increase significantly.* We also elicited a point estimate of the number of cases three days from the time the survey was taken.

Perceived threat: A 7-point Likert-type scale ranging from *Not at all dangerous* to *Extremely dangerous.*

Government action: A 7-point Likert-type scale from *Disagree strongly* to *Agree strongly* asking agreement with the statement "The [COUNTRY] should ban public gatherings, close non-essential businesses, and ask all citizens to stay at home unless they are going to work or carrying out necessary errands."

Government action, United States only: A list of policies to implement in the United States. Coded as the total number of policies implemented for a supplementary analysis.

Personal action, United States only: Two 7-point agreement Likert-type scales asking whether they will increase or decrease (a) their mask use and (b) their adherence to social distancing relative to now based on the graph.

Personal action, all countries: A 7-point Likert-type scale asking participants to consider all the efforts people in [COUNTRY] are currently taking to combat coronavirus and asking whether they should significantly decrease (–3) to significantly increase (3) those efforts.

*Moderators.*

We collected the following moderators:

- An Objective Numeracy Scale, as used in Study 1B
- A Subjective Numeracy Scale, as used in Study 1B
- Partisan lean, education, gender, age, size of city they reside in, ability to work from home, all as in Study 1B
- Self-reported confidence in their knowledge about coronavirus on a 1–7 scale, not previously collected

*Data.*

Nine hundred twenty-one participants completed this study on Amazon Mechanical Turk in exchange for payment. Of those, 4% failed an attention check asking them to correctly identify which of four countries they did not see in the survey, leaving 891 for analysis (*M age* =37.9 years; 48% were female). These exclusions do not change our main results below if we choose to include these participants.

*Analysis & results.*

Data for each of the Likert scale questions asked for multiple questions were analyzed using mixed-effects linear regressions, where the Likert question response is the dependent variable, the independent variables are the interaction of the two condition variables and the variables alone, and there are random intercepts by country the judgment was made on.

Growth judgments: Log condition was found to significantly decrease growth judgments, $\beta = -0.31$, $t(3553) = -13.78$, $p < .00001$. Being in the single countries condition was found to significantly increase growth judgments, $\beta = 0.05$, $t(3553) = 2.28$, $p = .02$. The interaction between conditions was not significant for growth judgments ($p > .05$).

Danger judgments: Log condition was found to significantly decrease danger judgments, $\beta = -0.17$, $t(3552) = -7.23$, $p < .00001$. Being in the single countries condition was not found to significantly increase danger judgments, $\beta = 0.04$, $t(3552) = 1.7$, $p = .07$. The interaction between conditions was significant for danger judgments; the interaction effect of being in the log condition and single country condition, relative to baseline, was $\beta = 0.069$, $t(3552) = 2.9$, $p = .003$.

Policy need judgments: Log condition was found to significantly decrease policy judgments, $\beta = -0.17$, $t(3553) = -7.2$, $p < .00001$. Being in the single countries condition was found to significantly increase policy judgments, $\beta = 0.06$, $t(3553) = 2.54$, $p = .01$. The interaction between conditions was significant for policy judgments; the interaction effect of being in the log condition and single country condition, relative to baseline, was $\beta = 0.065$, $t(3553) = 2.72$, $p = .006$.

Need for individual action judgments: Log condition was found to significantly decrease individual action judgments, $\beta = -0.19$, $t(3553) = -9.49$, $p < .00001$. Being in the single countries condition was found to significantly increase individual action judgments, $\beta = 0.05$, $t(3553) = 2.46$, $p = .01$. The interaction between conditions was significant for individual action judgments; the interaction effect of being in the log condition and single country condition, relative to baseline, was $\beta = 0.079$, $t(3553) = 3.9$, $p < .0001$.

For the Likert questions only asked for the United States—namely, mask use and social distancing—linear regressions were carried out predicting the Likert response with the interaction of the two conditions.

All of the mixed-effects models were repeated with simple linear models of the interaction between conditions per our preregistration, and in each case discussed next, the direction and significance of the effect of the log condition were the same with this alternate specification. The same is true of the direction and significance of interactions between the conditions.

Unlike the models discussed next, the context condition was not significant in these simple linear models of the interaction (all *ps* > .05).

For the two questions only asked about the United States, we preregistered that we would carry out a *t* test for each comparing ratings in the two log conditions. This found that the log condition significantly decreased future frequency of mask use ratings, $t(888) = -3.6$, $p = .0003$, and decreased future commitment to social distancing ratings, $t(888) = -3.9$, $p < .00001$. In the interest of fully reporting results, we also present the results of a linear regression predicting each of these ratings with the interactions of both conditions below, and the effects of the log condition remain significant and in the same direction under this alternate specification.

**Future frequency of mask use:** Log condition was found to significantly decrease frequency of mask use ratings, $\beta = -0.49$, $t(886) = -3.74$, $p = .0002$. Being in the single countries condition was not significant ($p > .05$), nor was the interaction between conditions ($p > .05$).

**Future commitment to social distancing:** Log condition was found to significantly decrease future commitment to social distancing ratings, $\beta = -0.54$, $t(886) = -4.407$, $p < .0001$. Being in the single countries condition was not significant ($p > .05$) The interaction between conditions was significant for future commitment to social distancing ratings; the interaction effect of being in the log condition and single country condition, relative to baseline, was $\beta = 0.4$, $t(886) = 2.235$, $p = .025$.

**Replication of past results:** We also preregistered that we would report whether our basic effect of log condition decreasing growth, danger, policy, and need for individual action judgments replicated using a mixed-effects regression predicting ratings based on log condition with random intercepts for country and subset to only those participants in the single country condition. We ran this analysis, and in every case, the log condition was found to significantly decrease these Likert ratings when data were subset only to the single country condition (all *ps* < .00001). Per our preregistration, we also ran this analysis with a simple linear regression predicting the ratings with the log condition, and all results remained significant under this alternate specification (all *ps* < .0001).

Numeracy analyses were conducted for objective and subjective numeracy measures. These analyses were exploratory. Mixed-effects linear regressions were conducted for each growth, danger, policy, and effort rating predicting the rating using the interaction between the log condition and objective numeracy or subjective numeracy scores, respectively, with country being viewed as a random intercept.

**Objective numeracy:** For the objective numeracy scores, higher objective numeracy scores predicted lower growth, $\beta = -0.04$, $t(3541) = -3.7$, $p < .001$, and danger ratings, $\beta = -0.03$, $t(3541) = -2.6$, $p < .01$; and higher policy ratings, $\beta = 0.03$, $t(3541) = 2.7$, $p < .01$; and they had no significant relationship with effort ($p > .05$).

**Subjective numeracy:** Subjective numeracy predicted higher ratings of danger, $\beta = 0.07$, $t(3548) = -3.5$, $p < .01$; effort, $\beta = 0.07$, $t(3548) = 3.9$, $p < .01$; and policy, $\beta = 0.1$, $t(3548) = 4.96$, $p < .01$, but had no significant relationship with growth ($p > .05$). This is the only case in which any numeracy score predicts lower ratings on these values, making this a puzzling exception to the general relationship that deserves further research.

We also collected a single prediction judgment for each of the countries, asking participants to predict the number of cases three days from the present. We determined the mean absolute error of predictions for each country. We ran a linear regression for each individual country, predicting absolute error with logarithmic axis condition, and found that logarithmic axis condition significantly predicted higher mean absolute error in each (all *ps* < .0001). This fully replicated our pattern of results from Studies 1A, 1B, and Appendix Study 3. Additionally, we looked for differences in over- or underestimation between conditions, running four logistic

regressions (one for each country) regressing condition on a binary variable indicating if a prediction was an overestimate or not. These regressions found that logarithmic axes increased overestimate chances in two countries, decreased in one, and had no significant relationship in the last. We checked for higher or lower predictions in one condition or the other with four linear models, one for each country, predicting the predictions with the condition variable. Results were inconsistent here: in three countries, log condition led to higher predictions, whereas in one, it lead to lower predictions (all *ps* < .05). Finally, we compared prediction errors between conditions using individual linear regressions for each country with error predicted by log condition and again found mixed results. For two countries, the log condition predicted higher prediction error, and for two it predicted lower error (all *ps* < .01). Again, the overall pattern was not consistent.

In addition, we conducted a set of Bayesian analyses checking the primary effects in the study. We ran Bayesian mixed-effects models using the R package brms. In all cases, we used the default priors, which are intended to be only weakly informative. For all models, we fit linear models using Markov chain Monte Carlo with a dependent variable of the relevant Likert judgment (growth ratings, danger ratings, and so on), independent variables of the two conditions, and their interaction, and we allowed the intercept to vary by country the judgment was being made on. This can be considered roughly equivalent to the frequentist mixed-effects models fit previously. In several cases, there were divergent transitions after warm-up that could not be eliminated with changes to fitting parameters checked by experimenters. Encouragingly, $R^{hat}$ for each coefficient of the models was 1, indicating convergence. Nonetheless, this is a concern for validity. We include the diagnostic plots for each model as well as the coefficient estimates so that readers can make their own judgments here, and full code and data can be found on our OSF page. Additionally, we also fit nonhierarchical models, which did not have these potential validity issues to confirm our result.

Growth judgment: The posterior distribution of the logarithmic axis coefficient indicates that the logarithmic axis condition likely leads to decreased judgments of growth as the 90% confidence interval did not include 0 (posterior median = −0.31 [−0.36, −0.27]). Being in the single countries condition likely increased judgments of growth (posterior median = 0.05 [0.01, 0.1]). The interaction between the two conditions had a posterior distribution whose 90% confidence interval included zero (posterior median = 0.03 [−0.01, 0.08]).

Danger judgment: The posterior distribution of the logarithmic axis coefficient indicates that the logarithmic axis condition decreased judgments of danger (posterior median = −0.17 [−0.22, −0.12]). The posterior distribution of the single countries condition indicates that that it is more likely than not that the condition likely increased judgments of danger, but the 90% confidence interval did overlap zero (posterior median = 0.04 [−0.01, 0.09]. The posterior distribution of the interaction term between the two conditions indicated that it increased judgments of danger (posterior median = 0.07 [0.02, 0.12]).

Need for policy action judgment: The posterior distribution of the logarithmic axis coefficient indicates that the condition decreased judgments of the need for policy action (posterior median = −0.17 [−0.22, −0.13]). The posterior distribution of the single country coefficient indicates that this condition increased judgments of the need for policy action (posterior median = 0.06 [0.02, 0.11]). The posterior distribution of the interaction term between the two conditions indicated that it increased judgments of the need for policy action (posterior median = 0.06 [0.02, 0.11]).

Need for individual action judgment: The posterior distribution of the logarithmic axis coefficient indicates that the condition decreased judgments of need for individual action (posterior median = −0.19 [−0.23, −0.15]). The posterior distribution of the single country coefficient indicates that this condition increased judgments of need for individual action (posterior median = 0.05 [0.01, 0.09]). The posterior distribution of the interaction

term between the two conditions indicated that it increased judgments of need for individual action (posterior median = 0.08 [0.04, 0.12]).

Because of the concerns with model validity when fitting the Bayesian mixed-effects models, we replicated these findings by fitting Bayesian linear models for each of these four judgments. To fit these models, we used the R package rstanarm, again with default priors. For these models, we had a dependent variable of the relevant Likert judgment and independent variables of each of the conditions, their interaction, and the country the judgment was being made on.

Growth judgment: Model results indicate that logarithmic axis condition decreased growth judgments, single country condition increased growth judgments, and the interaction of the two conditions had no effect to a small increasing effect on growth judgments.

Danger judgment: Model results indicate that the logarithmic axis condition decreased danger judgments, the single country condition had no effect to a small increasing effect on danger judgments, and the interaction had no effect to a small increasing effect on danger judgments.

Need for policy action judgment: Model results indicate that the logarithmic axis condition decreased judgments of the need for policy action, the single country condition increased such judgments slightly, and the interaction increased such judgments slightly.

Need for individual action judgment: Model results indicate that the logarithmic axis condition decreased judgments of the need for individual action, the single country condition increased such judgments slightly, and the interaction increased such judgments slightly.

## Study 3: Debiasing

### *Methods.*

This study tested the influence of a debiasing manipulation on the effect of presenting coronavirus data using a logarithmic versus linear axis.

#### *Conditions:*

The experiment had a 2 (logarithmic versus linear scale graphs) × 2 (debiasing manipulation versus control) design, for a total of four between-subjects conditions.

Participants were first shown either the debiasing manipulation or a control video of equivalent length and answered an attention check.

Participants then saw graphs for four countries and were asked questions about each. The United States was shown first, followed by three other countries in random order. Participants then answered a second attention check.

Finally, participants answered demographic and scale questions.

#### *Materials.*

The debiasing intervention was the following video, a 1 minute 45 second clip from the second section of this Vox Media video: https://www.youtube.com/watch?v=O-3Mlj3MQ_Q.

The control video was an equivalent length and had content unrelated to the topic – a painter describing how to paint a nature scene.

There were two main dependent variables:

Growth rate: a 7-point Likert-type scale asking how participants expect growth rates to change, from *decrease significantly* to *increase significantly.*

Perceived threat: a 7-point Likert-type scale ranging from *Not at all dangerous* to *Extremely dangerous.*

We also collected several moderators:

- Age
- *G*ender

#### *Data.*

Nine hundred ten Mechanical Turk workers completed the study for payments. After excluding all of those who did not pass attention checks, 739 (82%, *M* age= 40.3 years; 50.2% were female) remained for analysis.

*Analysis & results.*

We ran a mixed-effects regression predicting each of the Likert ratings with the interaction of the two condition variables. We preregistered that we would run these both with and without exclusions; given the relatively high number of participants failing our attention checks, we report both measures here in full. However, the basic pattern of results remains the same both with and without exclusions.

The results without any exclusions using the full data set:

Growth judgments: Log condition was found to significantly decrease growth judgments, $\beta = -0.11$, $t(3615) = -10.7$, $p < .00001$. Being in the debiasing condition was not significantly related to growth judgments ($p > .05$). The interaction between conditions was significant for growth judgments. The interaction effect of being in the log condition and the debiasing condition, relative to baseline, was $\beta = 0.06$, $t(3615) = 2.78$ $p = .006$.

Danger judgments: Log condition was found to significantly decrease danger judgments, $\beta = -0.11$, $t(3616) = -4.75$, $p < .00001$. Being in the debiasing condition was not found to significantly increase danger judgments ($p > .05$). The interaction between conditions was significant for danger judgments; the interaction effect of being in the log condition and the debiasing condition, relative to baseline, was $\beta = 0.06$, $t(3616) = 2.48$, $p = .01$.

The results with exclusions:

Growth judgments: Log condition was found to significantly decrease growth judgments, $\beta = -0.25$, $t(2945) = -10.7$, $p < .00001$. Being in the debiasing condition was not significantly related to growth judgments ($p > .05$). The interaction between conditions was significant for growth judgments; the interaction effect of being in the log condition and the debiasing condition, relative to baseline, was $\beta = 0.08$, $t(2945) = 3.263$, $p = .001$.

Danger judgments: Log condition was found to significantly decrease danger judgments, $\beta = -0.11$, $t(2945) = -4.25$, $p < .00001$. Being in the debiasing condition was not found to significantly increase danger judgments ($p > .05$). The interaction between conditions was significant for danger judgments; the interaction effect of being in the log condition and the debiasing condition, relative to baseline, was $\beta = 0.06$, $t(2945) = 2.61$, $p = .009$.

Directions and significance of effects remain the same when either of these analyses are repeated with a simple linear regression instead.

In addition, we ran a mixed-effects regression with data subset to those in the control condition predicting each Likert rating using the log condition alone, and we replicated our past results (all $ps < .05$). These results also replicated with a simple linear regression (all $ps < .05$).

We also carried out Bayesian analyses of our experimental results. We used the R package brms to fit Bayesian mixed-effects models for both the growth and danger judgments using Markov chain Monte Carlo. Each model had a dependent variable of either the growth or the danger judgment, respectively, and independent variables of the logarithmic condition, debiasing condition, and their interaction, with intercepts allowed to vary by target country being judged. We used default priors provided by the package, which are intended to be only weakly informative. These analyses replicate our previous results.

Results for our models analyzing the full data without any exclusions:

Growth judgments: We found that the log condition decreased judgments (posterior median = −0.24 [−0.29, −0.20]). The debiasing condition did not have a consistent effect (posterior median = −0.02 [−0.06, 0.03]). The interaction between the two terms appeared to increase judgments, although not enough to cancel out the effect of the log condition (posterior median = 0.06 [0.02, 0.11]).6

Danger judgments: We found that the log condition decreased judgments (posterior median = −0.11 [−0.16, −0.07]). The debiasing

condition did not have a consistent effect (posterior median = 0.03, [−0.01, 0.08]). The interaction between the two conditions appeared to increase judgments, although not enough to cancel out the effect of the log condition (posterior median = 0.06 [0.01, 0.11]).[7]

Results for our models analyzing the full data with exclusions:

Growth judgments: We found that the log condition decreased growth judgments (posterior median = −0.26 [−0.31, −0.21]). The debiasing condition did not have a consistent directional effect (posterior median = 0.01 [−.04, 0.06]). The interaction between the two conditions increased growth judgments but not enough to cancel out the effect of the log condition (posterior median = 0.08 [0.03, 0.13]).[8]

Danger judgments: We found that the log condition decreased danger judgments (posterior median = −0.12 [−0.17, −0.06]. The debiasing condition did not have a consistent directional effect (posterior median = 0.01 [−0.04, 0.07]. The interaction between the two conditions increased danger judgments, but not enough to cancel out the effect of the log condition (posterior median = 0.07 [0.02, 0.13]).[9]

## Appendix Studies 1–3

### Appendix Study 1: Effects of time period & slopes

*Methods.*
We first presented individuals with countries and then had them make growth and danger Likert ratings. Each participant saw four countries in random order. The types of countries participants saw varied by condition.

Conditions: The experiment was a 2 (log versus linear graphs) × 2 (concave down log slope versus linear log slope) design.

The concave down slope condition consisted of countries generally later into their outbreaks where cases where more under control, resulting in a slope of the log graphs that appeared concave down. The linear slope condition consisted of countries usually earlier into their outbreaks, where growth rates had not yet slowed and the slope of the log graphs appeared more linear. Countries in the linear slope condition were generally less developed then those in the concave down slope condition, although we attempted to control for this in stimuli selection, it may still be a confound.

*Materials.*
There were two main dependent variables:

Growth rate: A 7-point Likert-type scale asking how participants expect growth rates to change, from *decrease significantly* to *increase significantly.*

Perceived threat: A 7-point Likert-type scale ranging from *Not at all dangerous* to *Extremely dangerous.*

We also collected several moderators:
- Age
- Gender

*Analysis & results.*
The summary of the results is that for growth and danger, log condition decreases judgments, as does being in the concave down condition. This implies that the effect of logarithm condition is substantially the same regardless of slope. The fact that the concave down condition results in relatively lower ratings is not surprising—those countries are later in their epidemics and have them under better control. There is a significant interaction between conditions for danger ratings, which is again not surprising, as the difference between log and linear conditions can appear subjectively to be increased when the slope of the logarithmic graph is concave down versus linear. Full results are reported below.

For our main analysis, we predicted both Likert ratings with the interaction between the condition variables and both condition variables alone using a simple linear regression. We replicated this analysis with a mixed-effects regression including a random intercept for country, and

the direction and significance of all effects remains the same.

Growth judgments: Log condition was found to significantly decrease growth judgments, $\beta = -0.26$, $t(2960) = -10.34$, $p < .00001$, as was being in the concave down slope condition, $\beta = -0.32$, $t(2960) = -12.64$, $p < .00001$. The interaction between conditions was not significant ($p > .05$).

Danger judgments: Log condition was found to significantly decrease danger judgments, $\beta = -0.10$, $t(2960) = -4.21$, $p < .00001$, as was being in the concave down condition, $\beta = -0.19$, $t(2960) = -7.6$, $p < .00001$. The interaction between conditions was significant for danger judgments; the interaction effect of being in the log condition and concave down condition, relative to baseline, was $\beta = -0.08$, $t(2960) = -3.21$ $p = .001$.

These results hold with the same direction of effect and significance when those who do not pass the attention check are excluded.

We also replicate our previous findings that log condition decreases both judgments when we restrict to only those in the concave down condition and run a mixed effect or linear regression (all coefficients negative, all *ps* < .05).

An additional advantage of this study, beyond confirming that these results hold in different time periods of a pandemic, is helping further elucidate a possible explanation for what specific features of logarithmic scales change growth judgments. Other studies compare logarithmic scale graphs with linear scale graphs, so this study is the only one comparing logarithmic scale graphs to one another. This allows us to control for other differences between logarithmic and linear scales, such as differing granularities (or physical-to-numerical-distance ratios) of measurement. The fact that changing the perceived slope of the graph to be less steep while holding the other features of a logarithmic graph constant lowers growth and danger judgments appears to imply that the slope of these graphs may be one of the drivers of misperceptions. Further research is needed to explore this finding.

## Appendix Study 2: Accuracy of Growth Judgments

### *Methods.*

In this study, participants were presented with graphs showing three countries with varying growth rates; they were then asked questions about the countries' relative growth rates and the effectiveness of their interventions, as inferred from growth rates.

Conditions: There was a log graph condition and a linear graph condition.

### *Materials.*

The main dependent variables are the following:

Growth rate ranking: Rank order the three countries based on which country's cases are most likely to double the soonest to the latest from now based on current trends in the graph. This tests the accuracy of relative growth rates prediction.

Intervention effectiveness ranking: Rank order the effectiveness of each country's interventions based on the graph. This tests the accuracy of inferences about intervention effectiveness prediction.

Intervention report cards: Give a letter grade from A to F for each country's coronavirus response, based on the graph. This is not a dependent variable being used to test the main hypothesis; rather, it is exploratory and designed to see if there are inconsistencies between these letter grades and the rankings and to get a nonordinal ranking of relative effectiveness.

Policy adoption: Likert scale repeated for each country asking if the United States should adopt that country's policies, with answers ranging from *Definitely should adopt their policies* to *Definitely should not adopt their policies.* This tests the connection between believing policies are effective and wanting to adopt them.

*Data.*
Four hundred eight Mechanical Turk workers completed the study. After excluding those who failed to identify an attention check asking them which country they did not see, 384 (94%) remained.

*Analysis & results.*
The growth rate measure was tested with an ordered logistic regression, where the dependent variable is a four-level outcome variable where the highest value is all countries in the correct order and the lowest is no countries in the correct order, and the maximum possible distance between accurate ranking and the participant's ranking. The intervention effectiveness dependent variable was tested with an ordered logistic regression as above. The correct order was defined as one in which the country with the lowest growth rate is ranked as most effective and the one with the highest growth rate ranked as least effective. These analyses were repeated with a simple logistic regression with a dependent variable of 1 (all correct) or 0 (none correct) as a robustness check. Ultimately, all of these regressions found no significant relationship between the logistic regression condition and accuracy of the rankings using either the order metric or the simple all correct versus not binary metric (all *ps* > .157). Analyzing data with or without exclusions does not change this result.

Deviating from our preregistration, we did not carry out two planned analyses of the report card and policy adoption questions, where a ranking would be inferred from these Likert questions. There was relatively little variation on these questions, so it is unlikely this analysis will change results.

## Appendix Study 3: Single country psychological reactions

*Methods.*
Participants viewed graphs for four different countries—first the United States, then three other countries in random order—and answered questions about each based on the graphs.

Conditions: Log versus linear condition, which dictated the type of graph participant's saw.

*Materials.*
Growth rate: A 7-point Likert-type scale ranging from *Stay the same* to *Grow extremely quickly.* We will also elicit a point estimate of the number of cases three days from the time the survey is taken.

Perceived threat: A 7-point Likert-type scale ranging from *Not at all dangerous* to *Extremely dangerous.*

Government action: A 7-point Likert-type scale ranging from *Disagree strongly* to *Agree strongly* asking agreement with the statement "The US should ban public gatherings, close non-essential businesses, and ask all citizens stay at home unless they are going to work or carrying out necessary errands."

Personal action: Two 7-point agreement Likert-type scales as above asking for agreement with statements (a) "I will plan to wear a mask when in public places" and (b) "I will reduce the amount of time that I spend outside and with people who do not live in my household." Notably, these measures were less sensitive than those asked in Study 2 within the main article. Instead of asking for changes in action relative to their current efforts, it asked for their reported overall adherence to these measures on a relatively narrow scale. This makes this dependent variable less sensitive to change than the one reported in Study 2. This was asked only for the United States.

Preference for graphs: Two 3-measure questions where choices are the logarithmic, the linear, or neither graph asking (a) which graph they prefer and (b) which graph they believe is more informative.

*Data.*
Three hundred four Mechanical Turk workers completed the study; after excluding those who failed an attention check, 265 (88%) remained for analysis.

*Analysis & results.*

Likert scale questions asked for more than the United States were analyzed using mixed-effects regressions with condition predicting the Likert rating and random intercepts for countries. This was a deviation from our preregistration, where we registered a mistaken plan of analysis. In the preregistration, we registered that we would have random intercepts for both country and participant. However, because participants only ever were in one condition, this specification was not correct and effectively served to artificially reduce the magnitude of the condition effect.

Likert scale questions asked only about the United States were analyzed using linear regressions identical to the mixed-effects models, except without the random intercepts for country.

Results mentioned here hold with or without exclusions.

We report the main results below with *p* values from the mixed-effects regressions.

Participants in the linear condition consistently judged coronavirus's growth to be faster ($M_{linear}$ = 5.3, $M_{log}$ = 4.48, $p > .0001$) and threat to be greater ($M_{linear}$ = 5.85, $M_{log}$ = 5.67, $p = .034$) than did participants who saw the same data in a logarithmic scale. Accordingly, participants also judged the need for a strong policy response to be greater when they saw the linear scale graph than when they saw the logarithmic scale graph ($M_{linear}$ = 5.42, $M_{log}$ = 5.16, $p = .004$).

However, when asked specifically about the US graph and the individual actions they should take to combat the coronavirus—specifically, wearing masks and social distancing—there was no difference between conditions (all *ps* > .05). We believe the difference between this result and that found in Study 2 is that the dependent variables used here were different. As previously mentioned, they ask for use overall, not relative to people's current efforts. This makes them less sensitive measures, which are more noisy than the dependent variable used in Study 2.

At the end of the study, we explain and show both types of graphs to participants and find that participants consistently report both preferring linear to logarithmic graphs and finding them more informative. However, results indicate they may still have trouble understanding these graphs. Only 43% of participants correctly identified the main feature of logarithmic graphs, and overall, only 75% of participants were able to correctly identify the graphs they saw during the experiment as primarily linear or logarithmic.

As in Study 2, we also collected a single prediction judgment for each of the countries, asking participants to predict the number of cases three days from the present. Replicating our analysis in Studies 1A, 1B, and 2, we find that most participants underpredict the total number of cases in three days (69%, 76% in linear condition, 61% in log condition).

We determined the mean absolute error of predictions in both conditions and found that mean absolute error was generally higher for most countries in the logarithmic axis condition. We ran a mixed-effects model with an independent variable of absolute prediction error, a dependent variable of condition, and random effects by country. This found that the logarithmic axis condition marginally predicted increased absolute error, $\beta$ = 11,298, $t(1040) = 1.916$, $p = .055$. This directionally replicated our pattern of results from Studies 1A, 1B, and 2. Additionally, we looked for differences in over- or underestimation between conditions, finding no difference in overall magnitude of predictions or in over- and underestimates between conditions.

We tested for differences in the chances that any given prediction was an overestimate using a mixed-effects logistic regression predicting whether a given prediction was an overestimate with an independent variable of axis condition and a random intercept by country. The logarithmic axis condition was significantly associated with a higher chance of an overestimate, $\beta$ = 0.75, *z* value = 5.31, $p < .0001$). This is consistent with Study 1A, which found more overestimates in the log condition relative to

the linear condition, particularly early on, but contradictory to Study 1B, which found overestimates in the linear condition throughout. We also checked if there were any differences in predictions or prediction errors between conditions using mixed-effects regressions predicting each respective variable with the condition and a random effect for target country and found no significant relationship (all *ps* > .05).

## Publications Review

For our review of the graphs currently used in existing publications, we had three research assistants collect coronavirus articles that included graphs from three major newspapers: *The New York Times, The Wall Street Journal,* and the *Financial Times*. The first two newspapers are two of the top three most read in the United States, and the *Financial Times* is a United Kingdom–based newspaper with a global audience of approximately 16 million. This makes these newspapers good candidates for our search because graphs attached to their articles will be widely viewed by a large number of people. The search was intended to encompass any article on coronavirus from January to the end of April 2020. Complete data can be found on our OSF page.

## endnotes

1. These charts, as with analysis above, use the predictions winsorized at the 5th and 95th percentiles.
2. This pattern is similar across conditions. In the log axis condition, 77% of participant's predictions are underestimates, in the linear condition 78% of participant's predictions are underestimates.
3. These alternate accuracy analyses, as well as those in Study 1B, Study 2, and Appendix Study 3, were carried out in response to reviewer advice, so they were not preregistered.
4. These and other Bayesian analyses in Studies 1B, 2, and 3 were carried out in response to reviewer advice, so they were not preregistered.
5. This pattern holds in each condition individually as well, with 67% underestimate in the linear axis condition and 77% in the logarithmic axis condition.
6. This model initially did not converge for some parts of the model ($R^{hat}$ values significantly above 1), so we reran the model with 4,000 iterations (instead of the default 2,000) and with an adapt_delta parameter of .87, resulting in acceptable $R^{hat}$ values. This model did still have 14 divergent transitions after warm-up.
7. This model had seven divergent transitions after warm-up.
8. This model had five divergent transitions after warm-up.
9. This model initially did not converge, so we increased the number of iterations from the default 2,000 to 3,000. This resulted in satisfactory $R^{hat}$ values of 1.This model had 17 divergent transitions after warm-up.

## references


Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37–44. https://doi.org/10.1177/0272989x0102100105

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the Subjective Numeracy Scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making, 27,* 663–671. https://doi.org/10.1177/0272989x07303824