Original Article

Topical Issue: Psychological Perspectives on Science Communication

Time for an Update

Belief Updating Based on Ambiguous Scientific Evidence

Marcel R. Schreiner[1], Julian Quevedo Pütter[1], and Tobias R. Rebholz[2]

[1]Department of Psychology, University of Mannheim, Germany

[2]Department of Psychology, Eberhard Karls University of Tübingen, Germany

Abstract: Scientific evidence for many effects tends to be ambiguous. Here we investigated how psychological novices update their preexisting beliefs about psychological effects based on ambiguous scientific evidence. Specifically, we investigated various predictors and systematic patterns of belief updating. Participants were presented a series of fictitious hypotheses, accompanied by a series of fictitious study outcomes. For each hypothesis, we assessed participants' preexisting beliefs and subjective expertise regarding the topic, as well as their posterior beliefs after presentation of scientific evidence. We found a negative effect of subjective expertise and positive effects of trust in psychological science and number of studies investigating an effect on belief updating. We further found evidence for a belief updating pattern according to which participants weight the outcome of the most recent study stronger than that of previous studies. The results advance our understanding of evidence-based belief updating and provide practical implications for science communication.

Keywords: belief updating, science communication, metascience, advice taking, statistical modeling

Science provides us with information based on a reliable body of knowledge (Scheufele & Krause, 2019). It can thus be considered a suitable foundation on which we can form and adapt our beliefs about the world. However, scientific evidence for many effects tends to be ambiguous, that is, some studies show an effect and others do not. For example, the Open Science Collaboration (2015) conducted a replication project of 100 studies. Out of 97 original findings with significant results, only 36% of replications yielded significant results. In a many labs replication project, Klein et al. (2014) found a large range of replication rates for 13 psychological effects. Scientists may be used to ambiguity of scientific evidence and are, at least to some extent, trained to evaluate it (e.g., some replication failures can be explained simply by statistical reasons; e.g., Amrhein et al., 2019). This may, however, not be the case for the general public, as suggested by public reactions to the "replication crisis" in psychology (e.g., Anderson & Maxwell, 2017) and to scientific evidence concerning the COVID-19 pandemic (e.g., Simonovic & Taber, 2022). In the present study, we investigate how people update their beliefs about psychological phenomena based on ambiguous scientific evidence.

Past research has used different manipulations of scientific ambiguity or uncertainty to evaluate its effects on evaluations of scientific evidence. For example, providing numerical ranges and verbal explanations in news article-like texts had a relatively small effect on participants' trust in the presented outcomes (van der Bles et al., 2020). Similarly, verbally introducing uncertainty in news articles about climate change did not influence participants' decision-making concerning climate-friendly behaviors (Hendriks & Jucks, 2020). Using a quasi-experimental approach, study outcomes originating from research fields perceived as generally more (forensics) or less trustworthy (psychology) were not evaluated more or less positively (Broomell & Kane, 2017). These results suggest that effects of perceived ambiguity may be rather limited. In the current research, we adopted a different strategy by manipulating ambiguity in terms of the relative number of individual study outcomes confirming versus rejecting a specific hypothesis. Thereby, ambiguity in our study was operationalized as a continuous instead of dichotomous construct.

People typically hold some prior beliefs about various effects (e.g., Greenhoot et al., 2004). When being confronted with new information, such as scientific evidence, this information may be integrated into peoples' belief system and trigger an updating process that leads to altered beliefs (posterior beliefs). In the advice taking literature, the "weight of advice" is used to measure the degree of integration for shifting from prior to posterior beliefs in the light of external information provided by, for instance, experts or peers serving as advisors (Harvey & Fischer, 1997). On an aggregated level, people typically integrate external evidence provided by others less than normatively justified (Bailey et al., 2023; Soll & Larrick, 2009), which might not be true – at least not to the same extent – for scientific evidence that represents objective external information. Essentially, the conceptual idea of measuring external influences on internal beliefs can be extended to evidence from any external source of information, such as unrelated anchors (e.g., Tversky & Kahneman, 1974), algorithmic output (e.g., Logg et al., 2019), or, indeed, scientific evidence (e.g., Fiedler et al., 2019). Accordingly, we will rely on the mixed-effects regression weights (of advice) of Rebholz et al. (2024) to investigate various influences on and potential strategies of belief updating based on ambiguous scientific evidence.

First, we assess the impact of people's subjective expertise regarding an effect in question. People that ascribe themselves high subjective expertise may hold stronger prior beliefs and hence integrate scientific evidence to a smaller extent (cf. Li & Wagner, 2020; Perales et al., 2007). We thus expect that people with higher subjective expertise exhibit less belief updating (Hypothesis 1).

Second, we assess the effect of trust in (psychological) science on belief updating. While much research has focused on the effect of failed replications on trust in psychological research, suggesting that low replicability reduces trust (e.g., Anvari & Lakens, 2018; Chopik et al., 2018; Wingen et al., 2020), less research has focused on the evaluation of the effects in question themselves. People with low trust in science may discredit scientific evidence and rather adhere to their prior beliefs (cf. Landrum et al., 2015; Pilditch et al., 2020). We thus expect that people with lower trust in (psychological) science exhibit less belief updating (Hypothesis 2).

Third, we assess the effect of the breadth of information on belief updating. Regarding scientific evidence, we consider the breadth of information to be proportional to the number of studies investigating an effect. Evidence may consist of two components: strength (i.e., the proportion of instances in which the evidence favors particular hypotheses) and credence (i.e., the total amount or reliability of data; Griffin & Tversky, 1992; Kvam & Pleskac, 2016). A larger breadth of information should therefore increase the credence of evidence. We thus expect that people exhibit more belief updating the higher the number of studies investigating an effect (Hypothesis 3).

Finally, we aim to detect general patterns of belief updating. Here, we exploratory probe three potential strategies and a mixture of these strategies. First, people may adopt a kind of Bayesian reasoning and take the uncertainty, or degree of ambiguity, of evidence into account and exhibit less belief updating the higher the uncertainty of evidence (Behrens et al., 2007; Hogarth & Einhorn, 1992). We term this an *uncertainty-weighting strategy*. Regarding scientific evidence, ambiguity, and thus uncertainty, is highest when the same number of studies find confirmatory and contradictory evidence for an effect, respectively (i.e., for an evidence ratio of 0.5). An uncertainty-weighting strategy predicts a U-shaped relation between belief updating and the evidence ratio, with an evidence ratio of 0.5 as the inflection point.

Second, people may consider the deviation of the evidence from their prior beliefs and exhibit more belief updating the more the evidence contradicts their prior beliefs (Nassar et al., 2010). We term this a *unidimensional strategy*. This strategy predicts a linear relation between the absolute deviation of the evidence ratio from a person's prior belief, with no belief updating for a deviation of zero (i.e., the evidence confirms the prior belief) and increasing belief updating with increasing deviation.

Third, people may weight the evidence of the latest study in a sequence stronger than that of previous studies (e.g., Hogarth & Einhorn, 1992). We term this a *weight-last-stronger strategy*. Such a strategy may be a sensible heuristic when evaluating a series of studies, since one may expect that the most recent study incorporates the knowledge gained from previous studies, in particular if scientific progress is construed as knowledge accumulation (e.g., Bird, 2007; Rabinovich & Morton, 2012). A weight-last-stronger strategy predicts higher posterior beliefs if the last study in a sequence found confirmatory evidence for an effect than if it found contradictory evidence for an effect.

We found confirmatory evidence for all three hypotheses (i.e., a negative effect of subjective expertise and positive effects of trust in psychological science and number of studies on belief updating) and evidence for a weight-last-stronger updating strategy. However, we did not find evidence for an uncertainty-weighting or a unidimensional strategy in our main analysis.

Methods

The design, hypotheses, and analysis plan of the experiment were preregistered (http://dx.doi.org/10.23668/psycharchives.6911). The experiment was approved by the Ethics Committee of the University of Mannheim.

Participants

Participants were recruited via Prolific (https://www.prolific.com) and received a compensation of £2.67. They were prescreened to be native German speakers that do or did not study psychology. An a priori power simulation (P. Green & MacLeod, 2016) for detecting the three hypotheses and the different updating strategies with 80%

power yielded a desired sample size of 250 participants. The script for the power simulation can be found online in the supplementary materials. We gathered data from 300 participants to account for the potential necessity to exclude some participants from the analyses. All participants provided online informed consent for their participation and publication of their data. However, one participant revoked their consent, and we excluded the data of this participant from the analyses and from the data we made publicly available. As preregistered, we excluded participants who reported that their data should not be used at the end of the study. This applied to two participants who indicated not having properly responded to some questions or confusing some scale anchors. Thus, the final sample consisted of 297 participants (146 female, 146 male, five nonbinary) with a mean age of 32.64 years (SD = 11.61, range = 18–72). Information on the education of the sample (based on the CASMIN classification, Brauns et al., 2003; adapted for the German education system; see Schneider, 2016) and study fields (based on the Fields of Science and Technology; OECD, 2007) is available online in the supplementary materials.

Design

Our experiment had a 5 (number of studies) × 2 (most recent study outcome) within-subjects design. A total of 20 hypotheses were presented to participants. Each hypothesis was accompanied by fictitious study outcomes. Each outcome could be either positive (i.e., the hypothesized effect was found) or negative (i.e., the hypothesized effect was not found).

For each hypothesis, 4, 6, 8, 10, or 12 study outcomes were presented. Evidence ratios were chosen for each level of the "number of studies" factor such that a large and evenly distributed range of ratios would be present. Thus, for 4 and 8 studies, evidence ratios were 1/4, 2/4, and 3/4, respectively. For 6 and 12 studies, evidence ratios were 1/6, 2/6, 3/6, 4/6, and 5/6, respectively. For 10 studies, evidence ratios were 1/10, 3/10, 5/10, 7/10, and 9/10. As a consequence, the number of possible evidence ratios varied across the levels of the number of studies factor. Therefore, one evidence ratio was randomly selected per participant to be presented twice for four and eight studies, respectively. Similarly, one randomly selected evidence ratio per participant was excluded for 6, 10, and 12 studies, respectively. As a result, each level of the number of studies factor was repeated four times for each participant.

The most recent study outcome was manipulated such that half of the hypotheses (i.e., 10) would be associated with a positive last-study outcome and the other half with a negative last-study outcome. Positive and negative last-study outcomes were evenly distributed across the levels of the number of studies factor. Similarly, each of the four hypotheses within levels of the number of studies factor belonged to a different subdiscipline of psychology.

Material

Stimuli

For generating the material, we conducted a pilot study (N = 15) in which we asked psychology students at the University of Mannheim to generate one to three research questions or hypotheses for five subdisciplines of psychology (learning and memory, motivation and emotion, social psychology, differential psychology, and developmental psychology) for which they do not know any associated scientific studies. Participants could earn course credit for their participation in the pilot study. From the resulting data, we generated 39 candidate hypotheses by using or adapting the hypotheses suggested by the participants and by excluding unsuitable ones. Because we wanted to use hypotheses that have not yet been scientifically investigated, two of the authors independently generated keywords for each hypothesis and conducted a literature search using the GoogleScholar (https://scholar.google.com/) and PsycInfo (https://www.apa.org/pubs/databases/psycinfo) databases (considering the first 20 results each) to remove hypotheses with associated research. Disagreements were resolved by discussion. Finally, we jointly selected 20 hypotheses from four subdisciplines of psychology that served as the study material (e.g., "Hungry persons are more cooperative than full persons." and "People with more social contacts engage in less bullying." in German). Differential psychology was excluded because all candidate hypotheses had associated research. All hypotheses were specific enough to allow for a dichotomous evaluation as either *true* or *false*.

Trust in Psychological Science

We assessed participants' trust in psychological science using a scale adapted from Nisbet et al. (2015; cf. Wingen et al., 2020), translated into German. The scale consists of five items with a 7-point Likert scale ranging from 1 (=*strongly disagree*) to 7 (=*strongly agree*) and showed good reliability (cf. Kline, 2000; Taber, 2018) according to an internal consistency estimate (categorical ω = .88; S. B. Green & Yang, 2009) and an empirical reliability estimate ($r_{xx'}$ = .86; see, e.g., Maydeu-Olivares & Brown, 2010) derived from an item response theory (IRT; Lord, 1980; Lord & Novick, 1968) model (see the Data Analysis section). These estimates are comparable to previously reported ones (Nisbet et al., 2015; Wingen et al., 2020).

Scientific Literacy

We assessed participants' scientific literacy using 10 items from the Test of Scientific Literacy Skills (TOSLS; Gormally et al., 2012), translated into German. We slightly adapted one question of the TOSLS (Question 23) because in its original form the question was lacking some detail for being answered correctly. The test consists of items with a four-alternative forced-choice format. The selected items relate to the skills *identify a scientific argument*, *understand elements of research design and how they impact scientific findings/conclusions*, and *solve problems using quantitative skills, including probability and statistics*. We only used a subset of test items because of feasibility and test economy concerns and selected skills based on theoretical considerations regarding their relevance for the setting of the present study. The test showed questionable to acceptable reliability (cf. Kline, 2000; Taber, 2018) according to an internal consistency estimate (categorical ω = .63; S. B. Green & Yang, 2009) and an empirical reliability estimate ($r_{xx'}$ = .63; see, e.g., Maydeu-Olivares & Brown, 2010) derived from an IRT model (see the Data Analysis section). These estimates are lower than previously reported ones (Gormally et al., 2012), but note that we only used a subset of test items. For the use of scientific literacy as a control variable in the present study, we deemed the reliability of the test to be sufficient.

Procedure

After providing informed consent and being screened for exclusion criteria (see the Participants section), participants received detailed instructions regarding the main part of the experiment. First, the terms "Hypothesis" and "Scientific study" were explained in simple language. Hypotheses were explained as statements about reality that can be either true or false, whereas scientific studies were explained as a means to determine the validity of a given hypothesis. Participants were told that they would be presented with a series of hypotheses and associated study outcomes from genuine psychological research. All studies were described to be of comparably high scientific quality.

For each of the 20 hypotheses, participants were first asked to provide their prior belief by answering the question "How likely do you think it is that this hypothesis accurately describes reality?" on a visual analog scale (VAS). Anchors of the VAS were "Extremely unlikely," "Unsure," and "Extremely likely." On the next page, participants were asked to provide their subjective expertise ("How much expertise do you have in this area?") on a VAS ranging from "Extremely low" over "Moderate" to "Extremely high." Next, study outcomes were presented (see Figure 1). Outcomes were shown in chronological order along an arrow pointing to the right, with the most recent study at the right end of the arrow. The outcome of each study was represented by a green check (i.e., the hypothesized effect was found) or a red cross (i.e., the hypothesized effect was not found). Study outcomes were named as S1, S2, and so forth. Participants were given unlimited viewing time. On a separate page, participants were asked for their posterior belief ("How likely do you now think it is that this hypothesis accurately describes reality?"), again on a VAS ranging from "Extremely unlikely" over "Unsure" to "Extremely likely." This procedure was repeated in the same manner for all 20 hypotheses.

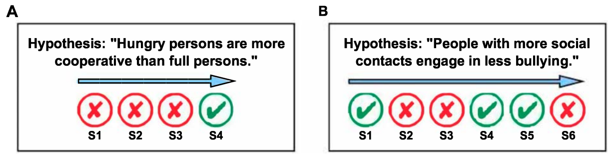

Figure 1. Exemplary depiction of the presentation of study outcomes. (A) Four studies with an evidence ratio of 1/4 and a positive last-study outcome. (B) Six studies with an evidence ratio of 3/6 and a negative last-study outcome. Green checks show that the effect has been found, and red crosses show that the effect has not been found.

Finally, participants were asked to answer the 10 items of the TOSLS and the five items of the trust in psychological science scale. After providing demographic information, they were thanked for their participation and debriefed. The study had a total duration of about 20 min.

Data Analysis

All analyses were conducted in R (R Core Team, 2022), and we used the packages *papaja* (version 0.1.1; Aust & Barth, 2022) and *tinylabels* (version 0.2.3; Barth, 2022) for reporting. We used a significance level of α = 5% for all analyses.

Measurement Model for Trust and Scientific Literacy

To obtain trait estimates for trust in psychological science and scientific literacy, while taking the categorical nature of the items into account, we fit a two-dimensional generalized partial credit IRT model (cf. Kelderman, 1996; Muraki, 1992) with independent latent traits[1] to the

[1] We also compared this model with a model with correlated latent traits, but the models showed almost identical fit, so we chose the more parsimonious model.

respective item responses. The model was fit using the package *mirt* (version 1.36.1; Chalmers, 2012). We then derived latent trait estimates from the model and used them as predictors in the subsequently described mixed-effects regression model.

Mixed-Effects Regression Model

We used participants' posterior beliefs about the stimulus items as a dependent variable for a mixed-effects regression model based on Rebholz et al. (2024). The model distinguishes between the judgment and weighting levels of belief formation. At the judgment level, participants' posterior beliefs are modeled as a function of their prior beliefs and the presented evidence ratio. Moreover, for testing the weight-last-stronger strategy, the evidence valence of the last study was included as a contrast-coded fixed effect at the judgment level. At the weighting level, we specified mutually independent random intercepts of participants and stimulus items. We also included fixed effects of the remaining strategies (unidimensional and uncertainty-weighting), the hypothesized predictors (subjective expertise, trust in science, and number of studies), and control variables (scientific literacy and education level) at the weighting level. More formal details on the specification of the mixed-effects regression model can be found in the Appendix. The model was fitted using the package *lme4* (version 1.1.31; Bates et al., 2015), and the statistical significance of the fixed effects estimates was assessed using the package *lmerTest* (version 3.1.3; Kuznetsova et al., 2017). We conducted directed tests of our main hypotheses and for testing the different updating strategies but still report two-tailed rather than one-tailed p values to follow common reporting standards. We highlight cases where the outcome of the directed test differs from that of an undirected test.

Results

The full model as specified in Equations 1 and 2 in the Appendix, including all hypothesized predictors, updating strategies, and control variables, can be found in Table 1. According to this model, the mean weighting of the presented evidence ratios when all explanatory variables are zero, that is, $\widehat{b}_0$, was higher than the equal weighting threshold of 0.50. However, there were also large inter- and intraindividual differences, as indicated by $\widehat{\tau}_S$ and $\widehat{\tau}_T$, respectively. For the control variables, we found that the level of education did not affect participants' weighting of the presented evidence ratios. By contrast, scientifically more literate participants weighted the evidence significantly more strongly than scientifically less literate participants.

Hypothesis 1 states that participants who ascribe themselves higher subjective expertise for assessing a certain hypothesis integrate scientific evidence to a smaller extent. The significantly negative coefficient estimate $\widehat{b}_{\mathrm{SE}}$ provided confirmatory evidence for this hypothesis (see also Figure 2, panel H1). According to Hypothesis 2, people with lower trust in psychological science exhibit less belief updating than participants with more trust. As indicated by the significantly positive coefficient estimate $\widehat{b}_{\mathrm{TP}}$, trust indeed had a positive effect on participants' evidence weighting (see also Figure 2, panel

Table 1. Full multilevel model as specified in Equations 1 and 2 including all hypothesized predictors, updating strategies, and control variables

	Estimate	95% CI	*SE*	*t*	*df*	*p*
b_0	0.6303	[0.5253, 0.7354]	0.0536	11.76	465.55	<.001
b_{WLS}	0.0165	[0.0084, 0.0247]	0.0041	3.99	5,719.64	<.001
b_{Edu}	0.0072	[−0.0087, 0.0231]	0.0081	0.89	294.43	.37
b_{SL}	0.0454	[0.0077, 0.0830]	0.0192	2.36	295.50	.02
b_{SE}	−0.3101	[−0.3637, −0.2565]	0.0274	−11.33	4,772.31	<.001
b_{TP}	0.0404	[0.0085, 0.0723]	0.0163	2.48	293.45	.01
b_{NS}	0.0125	[0.0085, 0.0166]	0.0021	6.03	5,723.71	<.001
b_{UWS}	0.1414	[−0.1232, 0.4061]	0.1350	1.05	5,829.27	.29
b_{uni}	−0.0094	[−0.0848, 0.0660]	0.0385	−0.24	5,793.27	.81
τ_S	0.2402	[0.2181, 0.2644]				
τ_T	0.0469	[0.0263, 0.0651]				
σ	0.1558	[0.1530, 0.1587]				
ICC	0.25					
$R^2_{marg.}$	0.64					
$R^2_{cond.}$	0.73					

Note. Wald 95% CI for fixed and bootstrap 95% CI (with 1,000 iterations) for random effects are shown.

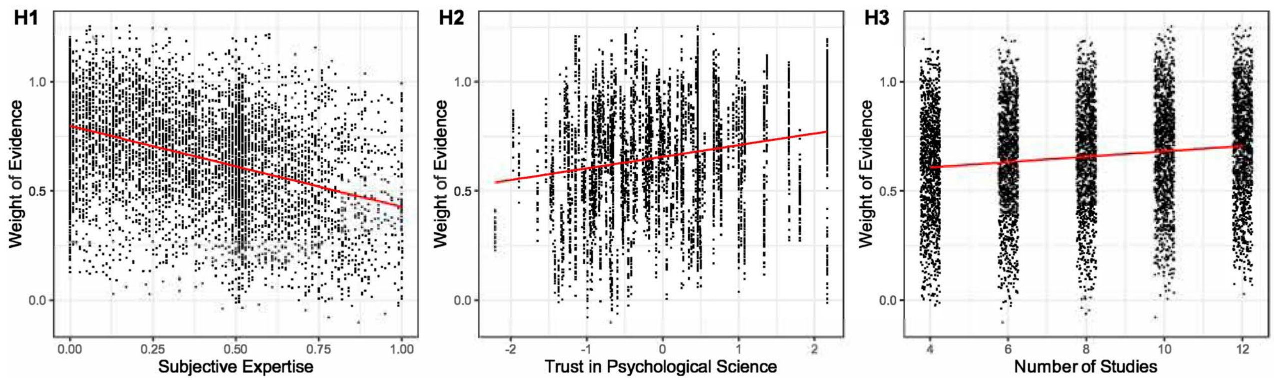

Figure 2. Weight of evidence as functions of the interaction of prior beliefs and subjective expertise (H1), trust in psychological science (H2), and number of presented studies (H3). The red solid lines show the linear regression fits. In panel H3, the points are scattered across the *x*-axis for better visualization.

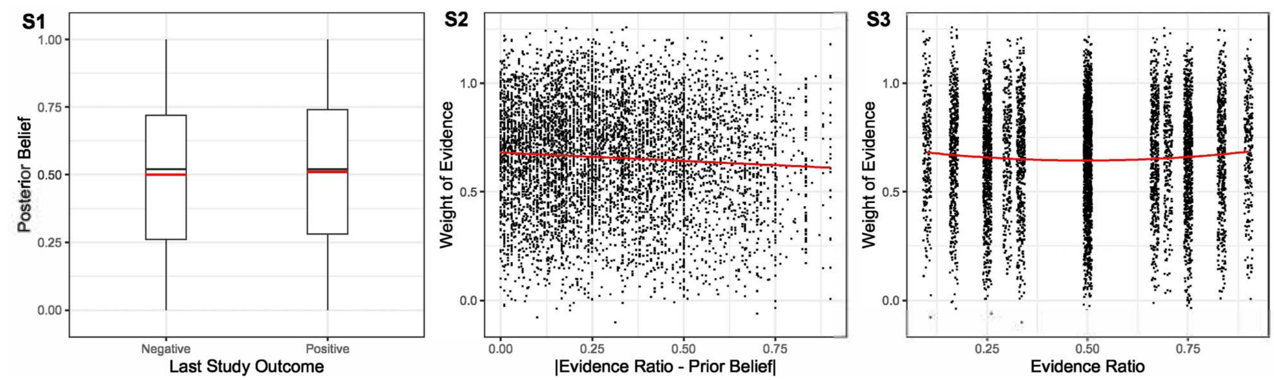

Figure 3. Posterior beliefs by last-study outcome (S1) and weight of evidence as functions of the absolute difference between the evidence ratio and prior beliefs (S2) and the raw evidence ratio (S3). The red bars in the boxplots display the means. The red solid lines in the scatter plots show the linear and second-order polynomial regression fits, respectively. In panel S3, the points are scattered across the *x*-axis for better visualization.

H2). Following from Hypothesis 3, participants should weight the evidence ratios more strongly if those are derived from more as compared to less number of studies investigating an effect. Indeed, the significantly positive coefficient estimate $\widehat{b}_{\mathrm{NS}}$ is in line with this hypothesis, that is, a larger breadth of information being weighted relatively more strongly (see also Figure 2, panel H3).

For participants' belief updating, we found evidence only for one of the preregistered updating strategies. Specifically, the significantly positive coefficient estimate $\widehat{b}_{\mathrm{WLS}}$ indicates that participants' final beliefs were more than one percentage point higher (i.e., in support of a certain hypothesis) if the last study found confirmatory evidence for an effect than if it found contradictory evidence for an effect (see also Figure 3, panel S1). That is, participants indeed applied the weight-last-stronger strategy by placing additional weight on the outcome of the last study in a sequence of multiple studies. By contrast, according to the coefficient estimates $\widehat{b}_{\mathrm{uni}}$ and $\widehat{b}_{\mathrm{UWS}}$, there was evidence for neither a unidimensional nor uncertainty-weighting strategy in participants' belief updating (see also Figure 3, panels S2 and S3, respectively).[2]

In the advice taking literature, research typically finds an inverse-U-shaped relation between advice distance and weighting for quantitative judgment tasks (e.g., the airline distance between two cities), where advice of "intermediate" absolute distance is weighted relatively more strongly than both closer and more distant advice (e.g., Moussaïd et al., 2013; Rebholz & Hütter, 2022). Therefore, we extended Equation 2 by the log of the absolute distance,

[2] In the power simulations, we conducted likelihood ratio testing of the models including the respective updating strategies against the null model including only the control variables. The statistical conclusions are the same for this alternative testing procedure, the results of which can be found online in the supplementary materials.

Table 2. Full multilevel model as specified in Equations 1 and 2 including all hypothesized predictors, updating strategies, control variables, and a logistic trend of absolute advice distance

	Estimate	95% CI	*SE*	*t*	*df*	*p*
b_0	0.5096	[0.3738, 0.6455]	0.0693	7.36	1,210.39	<.001
b_{WLS}	0.0165	[0.0083, 0.0246]	0.0041	3.97	5,718.93	<.001
b_{Edu}	0.0071	[−0.0088, 0.0230]	0.0081	0.88	294.45	.38
b_{SL}	0.0453	[0.0076, 0.0830]	0.0192	2.36	295.51	.02
b_{SE}	−0.3037	[−0.3575, −0.2499]	0.0274	−11.08	4,682.90	<.001
b_{TP}	0.0411	[0.0092, 0.0730]	0.0163	2.53	293.61	.01
b_{NS}	0.0127	[0.0086, 0.0167]	0.0021	6.09	5,723.41	<.001
b_{UWS}	0.2520	[−0.0237, 0.5277]	0.1406	1.79	5,824.65	.07
b_{uni}	−1.2681	[−2.1702, −0.3659]	0.4602	−2.76	5,835.44	.006
$b_{\text{uni- log}}$	1.8595	[0.5305, 3.1885]	0.6779	2.74	5,817.55	.006
τ_{S}	0.2402	[0.2163, 0.2621]				
τ_{T}	0.0451	[0.0229, 0.0634]				
σ	0.1557	[0.1529, 0.1584]				
ICC	0.25					
$R^2_{marg.}$	0.64					
$R^2_{cond.}$	0.73					

Note. Wald 95% CI for fixed and bootstrap 95% CI (with 1,000 iterations) for random effects are shown.

specifically $\log(|E_{ij} - Pr_{ij}| + 1)$,[3] to estimate nonlinear unidimensional belief updating in an exploratory post hoc analysis (Schultze et al., 2015). The model summary can be found in Table 2. Qualitatively and quantitatively, the results did not differ much from the ones as reported in Table 1, which indicated that the main findings from above were replicated for implementing nonlinear unidimensional weighting. More importantly, however, both the linear effect $\widehat{b}_{\text{uni}}$ and the logistic effect $\widehat{b}_{\text{uni-log}}$ of the absolute distance on weighting were significant in this extended model. That is, there was evidence for an inverse-U-shaped pattern of weighting scientific evidence that resembled the results as traditionally found in advice taking for quantitative judgments. Additionally, the coefficient estimate $\widehat{b}_{\text{UWS}}$ was significantly larger than zero for one-tailed testing in this extended model. Thus, by accounting for nonlinear unidimensional weighting, there was also post hoc evidence for an uncertainty-weighting strategy.

Discussion

In the current research, we investigated how people update their beliefs when being faced with ambiguous scientific evidence. Confirming all our hypotheses, we found that people with higher subjective expertise and people with lower trust in (psychological) science exhibit less belief updating and that people exhibit more belief updating the higher the number of studies investigating an effect. We further investigated general patterns of belief updating in an exploratory manner. Here, we found evidence for a weight-last-stronger strategy, indicating that peoples' final beliefs were higher if they were presented with a last study that found confirmatory evidence for an effect, but no evidence for a unidimensional or uncertainty-weighting strategy in our main analysis.

Predictors of Belief Updating

In the current research, we found subjective expertise to be negatively related to belief updating, being the strongest predictor of the ones investigated.[4] This was the case although high subjective expertise was likely unwarranted in the study, as participants were psychological novices and we used fictitious hypotheses that have, to our knowledge, not been investigated in prior research. Still, subjective expertise and actual expertise are not necessarily independent. For example, participants may have some expertise in areas

[3] Adding one to the absolute distance before taking the natural logarithm is necessary to avoid undefined values for cases in which the presented evidence ratio exactly matched participants' prior beliefs.

[4] To test this claim, we also estimated the model in Equations 1 and 2 using z-standardized predictors. The results, which can be found online in the supplementary materials, confirm that the effect of subjective expertise on weighting is the strongest among all investigated predictors.

related to the topic of the presented hypotheses or science in general. Given that we only used psychological hypotheses though, this should result in rather stable ratings of subjective expertise across hypotheses for a given participant. However, there was substantial within-person variability in ratings of subjective expertise, with an average *SD* of 0.20, and there was only one participant who exhibited no within-person variability in expertise ratings. Thus, subjective expertise and actual expertise were likely rather independent in the current study.

Although high subjective expertise was likely unwarranted, people who indicated higher subjective expertise regarding the topic of a hypothesis incorporated the available evidence to a much smaller extent than people who indicated lower subjective expertise (see also Li & Wagner, 2020; Perales et al., 2007). This finding suggests that it may be especially difficult to convince people who feel they are well-informed but gather information from unreliable sources, which frequently consumed also promote conspiracy beliefs (Schemer et al., 2022). This does not mean that expertise is maladaptive. Indeed, experts can draw from a larger set of information and may thus update their beliefs in a more informed manner (see, e.g., Mayer et al., 2023). In the current research however, high subjective expertise likely reflected unwarranted overconfidence in one's knowledge and may thus also be considered an indicator of the certainty of peoples' prior beliefs. The negative relationship between subjective expertise and belief updating suggests that belief updating can be expected to be weaker for more polarizing topics than the rather neutral ones investigated in the current research, as people may more strongly adhere to their prior beliefs. This expectation has been confirmed by other research (Su, 2022). In addition, participants' extremity of prior beliefs (i.e., the difference between prior belief and scale midpoint) was significantly but weakly correlated with subjective expertise ($r = .16$, 95% CI $[.05, .27]$, $t(295) = 2.79$, $p = .006$). Thus, belief extremity and the certainty with which beliefs are held do not seem to be strongly related.

We further found trust in (psychological) science to be positively related to belief updating, which suggests that people with lower trust tend to discount scientific evidence when adjusting their beliefs (cf. Landrum et al., 2015; Pilditch et al., 2020). This highlights a severe practical consequence of reduced trust in psychological science due to, for example, failed replications (Anvari & Lakens, 2018; Chopik et al., 2018; Wingen et al., 2020). Our findings suggest that effects of failed replications or low replicability are not limited to trust but also result in people discounting scientific evidence.

Moreover, we found that the number of studies investigating an effect was positively related to belief updating. A higher number of studies provide a larger breadth of information and therefore increase the credence of the evidence (cf. Griffin & Tversky, 1992; Kvam & Pleskac, 2016). Our results suggest that participants take the breadth of available information into account. Therefore, conducting and properly communicating replication studies may at least partially counteract the negative consequences of low trust in science. This highlights the importance of replication studies not just for scientific needs but also for the purpose of science communication. Of course, the present study itself provides only one piece of information for the investigated effects. While some findings are in line with previous ones, the replicability and generalizability of our findings requires further scrutiny and should be investigated in future research. This would then increase the number of studies that investigated the same effects we did and enable the same benefits we observed from resulting in the communication of an increased number of studies.

Finally, it is noteworthy that out of the two control variables we considered only scientific literacy, but not education level, was significantly related to belief updating.[5] This suggests that, to foster belief updating, the communication of study outcomes need not necessarily be adjusted to different levels of education, but the consideration of recipients' scientific literacy is more important. The finding further suggests that the promotion of scientific literacy in the general population may prove beneficial in increasing peoples' utilization of scientific evidence in the adjustment of their beliefs about the world. However, the results regarding scientific literacy should be interpreted with some caution, as the reliability of the subset of items from the TOSLS (Gormally et al., 2012) we used in the present study was limited.

Patterns of Belief Updating

In our exploratory investigation of different patterns of belief updating, we found that people place additional weight on the outcome of the last study in a sequence of studies (i.e., a weight-last-stronger strategy; cf. Hogarth & Einhorn, 1992). This may indeed be a sensible heuristic in practice, since the last study in a sequence may incorporate knowledge from previous studies. However, the effect was rather small: There was only an increase of 1.65 percentage points in posterior beliefs if the outcome of the last study reflected confirmatory evidence for an effect compared to when it reflected contradictory evidence (see $\hat{b}_{\mathrm{WLS}}$ in Table 1). From this

5 The two variables were also only weakly and nonsignificantly related as indicated by their Spearman's rank correlation ($r_s = .10$, $S = 3,918,148.21$, $p = .077$).

perspective, such a weight-last-stronger strategy may have limited practical consequences. On the other hand, it seems noteworthy that we found evidence for this strategy although all study outcomes were presented simultaneously on the same screen. In the context of scientific communication, meta-analyses can be considered an example for such a simultaneous presentation format (e.g., in forest plots). However, in many more applied settings, study outcomes may instead be presented sequentially, which may further increase the additional weight placed on the outcome of the last (i.e., most recently encountered) study outcome.

We further found that people seem to be insensitive to the distance of the presented evidence from their prior beliefs, at least if considering the linear trend proposed by the unidimensional strategy that we have investigated (cf. Nassar et al., 2010). This suggests the possibility that people update their beliefs by a fixed amount, potentially based on an individual threshold. However, instead of a linear unidimensional strategy, in a post hoc analysis, we found that people may actually be sensitive to the distance of the presented evidence from their prior beliefs. There is evidence that the actual relation is inverse-U-shaped such that people update their beliefs most for evidence of intermediate (absolute) distance from their prior beliefs and less for evidence that is similar to or very conflicting with their prior beliefs. This trend closely corresponds to the typical inverse-U-shaped relation between advice distance and weighting in the advice taking literature, where highest weighting is found for advice of intermediate absolute distance (Moussaïd et al., 2013; Schultze et al., 2015). However, it remains unclear how this distance relation for quantitative judgments that range from zero to infinity translates to probabilistic judgments as used in the current research, which are restricted to the $[0, 1]$ interval by definition.

Finally, we found that people seem to not take the uncertainty or ambiguity of available evidence into account, as we did not find evidence for an uncertainty-weighting strategy (cf. Behrens et al., 2007; Hogarth & Einhorn, 1992) in our main analysis. However, this finding should be interpreted with some caution, as there was a descriptive trend for a U-shaped relation between belief updating and evidence ratio proposed by this strategy in the main analysis, and a significant effect in the exploratory analysis when taking the inverse U-shaped relation between evidence weighting and distance into account. Further research is needed to determine whether, and to what extent, people incorporate uncertainty or ambiguity into their weighting of scientific evidence. Note that the current study does not allow to compare participants' judgment accuracy to that of a normative model (e.g., a model of Bayesian reasoning), as the actual validity of the hypotheses used as study material is unknown.

Limitations

There are some potential limitations concerning the results of our current research. First, we used rather neutral psychological hypotheses, in which participants had little knowledge and were unlikely to have a strong personal investment. We did so because we intended to investigate belief updating under a kind of baseline condition. While such neutral scientific topics are indeed often reported in the media (cf. Schäfer, 2009, 2011), more polarizing topics such as climate change or COVID-19 are also frequently reported and may more strongly enter public discourse. Belief updating regarding such more polarizing topics may differ, at least to some extent, from belief updating regarding more neutral topics (see also Su, 2022). For instance, people likely hold stronger prior beliefs for more polarizing topics that may, for example, be influenced by political partisanship (Li & Wagner, 2020; Van Bavel & Pereira, 2018). However, we believe that our findings also have implications for more polarizing topics. For example, the extremity of beliefs and the certainty with which they were held (using subjective expertise as an indicator) were only weakly correlated in the current research. In addition, despite the neutrality of topics, an effect of trust in science on belief updating emerged. Thus, we expect that effects similar to those observed in the current study would also be observed when more polarizing topics are used, although the magnitude of the effects may differ. For example, a stronger effect of trust in science might be expected when using more polarizing topics.

Second, we kept the information provided about each presented study limited to its dichotomous outcome and presented all outcomes simultaneously on a single screen. While this methodological approach allowed us to specifically investigate our hypothesized predictors and patterns of belief updating under controlled conditions, it should be acknowledged that such a highly internally valid approach might reduce the external validity of our results. Indeed, in many applied settings, more information about each study is provided (e.g., sample size), and outcomes are sometimes presented as more ambiguous (e.g., only partially confirmed hypotheses) and in a sequential format (e.g., in television and radio formats). Nevertheless, the minimalistic setup of our experiment is approximated in certain applied contexts, such as forest plots in meta-analyses or short news headlines in social media feeds. Moreover, it could be argued that by using a more realistic setup involving more complex information, some of the patterns we found should even be more pronounced, such as the effect of scientific literacy. Future research should therefore explicitly manipulate these factors and compare their effects to our baseline results.

Third, overall evidence ratios rather than specific study outcomes were the main focus of our current research. Despite trying to suggest a temporal order of study outcomes, given the simultaneous presentation format, considerations about temporal properties of the belief updating process were omitted. For example, whereas we found a weight-last-stronger or recency effect, an additional primacy effect might be expected to emerge when people have more time to integrate the results of the first study into their pre-existing belief system and discount subsequently encountered information (Kunda, 1990). Future research might allow for such processes to occur by presenting study outcomes sequentially and manipulating the duration of time intervals between presentations.

Finally, our current research was focused on short-term belief updating immediately following the presentation of study outcomes. For practical purposes, it might be interesting to look at longer time intervals to assess the persistence of the effects. Recent evidence in the context of COVID-19 misconceptions suggests that initial belief shifts might not persist over time (Carey et al., 2022).

Conclusion

Overall, our findings emphasize the relevance of psychological research to the general public. More specifically, psychological novices seem to recognize psychological research output as a suitable foundation on which to form and adapt their beliefs about the world. However, the impact of psychological research on individuals' belief updating varies considerably and is influenced not only by factors inherent to the research itself (e.g., the number of studies that so far have investigated an effect) but also by factors that are outside of researchers' direct sphere of influence (e.g., people's subjective expertise and trust in psychological science). Thus, there might be more to science communication than merely conveying specific scientific findings to interested nonscientists. Instead, a more holistic approach to science communication could benefit from considering alternative sources of knowledge that people may use to form and adapt their beliefs (e.g., *life experience* or practical experience in the case of psychological phenomena), and acknowledging concerns that people may have regarding psychological research (e.g., the replication crisis).

References

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(S1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, *3*(3), 266–286. https://doi.org/10.1080/23743603.2019.1684822

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. R package version 0.1.1. https://cran.r-project.org/package=papaja

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2023). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, *42*, 24516–24541. https://doi.org/10.1007/s12144-022-03573-2

Barth, M. (2022). *tinylabels: Lightweight variable labels*. R package version 0.2.3. https://cran.r-project.org/package=tinylabels

Bates, D. M., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bird, A. (2007). What is scientific progress? *Noûs*, *41*(1), 64–89. https://doi.org/10.1111/j.1468-0068.2007.00638.x

Brauns, H., Scherer, S., & Steinmann, S. (2003). The CASMIN educational classification in international comparative research. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in cross-national comparison: A European working book for demographic and socio-economic variables* (pp. 221–244). Springer. https://doi.org/10.1007/978-1-4419-9186-7_11

Broomell, S. B., & Kane, P. B. (2017). Public perception and communication of scientific uncertainty. *Journal of Experimental Psychology: General*, *146*(2), 286–304. https://doi.org/10.1037/xge0000260

Brown, L. D., Mukherjee, G., & Weinstein, A. (2018). Empirical Bayes estimates for a two-way cross-classified model. *The Annals of Statistics*, *46*(4), 1693–1720. https://doi.org/10.1214/17-AOS1599

Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour*, *6*(2), 236–243. https://doi.org/10.1038/s41562-021-01278-3

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology*, *45*(2), 158–163. https://doi.org/10.1177/0098628318762900

Fiedler, K., Hütter, M., Schott, M., & Kutzner, F. (2019). Metacognitive myopia and the overutilization of misleading advice. *Journal of Behavioral Decision Making*, *32*(3), 317–333. https://doi.org/10.1002/bdm.2109

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a Test of Scientific Literacy Skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBELife Sciences Education*, *11*(4), 364–377. https://doi.org/10.1187/cbe.12-03-0026

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

Greenhoot, A. F., Semb, G., Colombo, J., & Schreiber, T. (2004). Prior beliefs and methodological concepts in scientific reasoning. *Applied Cognitive Psychology*, *18*(2), 203–221. https://doi.org/10.1002/acp.959

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-r

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Hendriks, F., & Jucks, R. (2020). Does scientific uncertainty in news articles affect readers' trust and decision-making? *Media and Communication*, *8*(2), 401–412. https://doi.org/10.17645/mac.v8i2.2824

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement*, *20*(2), 155–168. https://doi.org/10.1177/014662169602000205

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180. https://doi.org/10.1016/j.cognition.2016.04.008

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109–111. https://doi.org/10.1016/j.tics.2014.12.007

Li, J., & Wagner, M. W. (2020). The value of not knowing: Partisan cue-taking and belief updating of the uninformed, the ambiguous, and the misinformed. *Journal of Communication*, *70*(5), 646–669. https://doi.org/10.1093/joc/jqaa022

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*(6), 935–974. https://doi.org/10.1080/00273171.2010.531231

Mayer, M., Broß, M., & Heck, D. W. (2023). Expertise determines frequency and accuracy of contributions in sequential collaboration. *Judgment and Decision Making*, *18*, e2. https://doi.org/10.1017/jdm.2023.3

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PLoS ONE*, *8*(11), Article e78433. https://doi.org/10.1371/journal.pone.0078433

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

Nisbet, E. C., Cooper, K. E., & Garrett, R. K. (2015). The partisan brain: How dissonant science messages lead conservatives and liberals to (dis)trust science. *The ANNALS of the American Academy of Political and Social Science*, *658*(1), 36–66. https://doi.org/10.1177/0002716214555474

OECD. (2007). *Revised field of science and technology (FOS) classification in the Frascati manual*. https://www.oecd.org/sti/inno/38235147.pdf

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Perales, J. C., Catena, A., Maldonado, A., & Cándido, A. (2007). The role of mechanism and covariation information in causal belief updating. *Cognition*, *105*(3), 704–714. https://doi.org/10.1016/j.cognition.2006.12.003

Pilditch, T. D., Madsen, J. K., & Custers, R. (2020). False prophets and Cassandra's curse: The role of credibility in belief updating. *Acta Psychologica*, *202*, Article 102956. https://doi.org/10.1016/j.actpsy.2019.102956

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rabinovich, A., & Morton, T. A. (2012). Unquestioned answers or unanswered questions: Beliefs about science guide responses to uncertainty in climate change risk communication. *Risk Analysis*, *32*(6), 992–1002. https://doi.org/10.1111/j.1539-6924.2012.01771.x

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications.

Rebholz, T. R., Biella, M., & Hütter, M. (2024). Mixed-effects regression weights for advice taking and related phenomena of information sampling and utilization. *Journal of Behavioral Decision Making*, *37*(2), Article e2369. https://doi.org/10.1002/bdm.2369

Rebholz, T. R., & Hütter, M. (2022). The advice less taken: The consequences of receiving unexpected advice. *Judgment and Decision Making*, *17*(4), 816–848. https://doi.org/10.1017/S1930297500008950

Schäfer, M. S. (2009). From public understanding to public engagement: An empirical assessment of changes in science coverage. *Science Communication*, *30*(4), 475–505. https://doi.org/10.1177/1075547008326943

Schäfer, M. S. (2011). Sources, characteristics and effects of mass media communication on science: A review of the literature, current trends and areas for future research. *Sociology Compass*, *5*(6), 399–412. https://doi.org/10.1111/j.1751-9020.2011.00373.x

Schemer, C., Ziegele, M., Schultz, T., Quiring, O., Jackob, N., & Jakobs, I. (2022). Political information use and its relationship to beliefs in conspiracy theories among the German public. *Journalism & Mass Communication Quarterly*, *99*(4), 908–929. https://doi.org/10.1177/10776990211063527

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116*(16), 7662–7669. https://doi.org/10.1073/pnas.1805871115

Schneider, S. L. (2016). *Die Konzeptualisierung, Erhebung und Kodierung von Bildung in nationalen und internationalen Umfragen* [The conceptualisation, assessment, and coding of education in national and international surveys]. GESIS - Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). https://doi.org/10.15465/10.15465/gesis-sg_020

Schreiner, M. R., Quevedo Pütter, J., & Rebholz, T. R. (2022). *Time for an update: How does ambiguous scientific evidence change our beliefs about the world?* [Pre-registration]. PsychArchives. https://doi.org/10.23668/psycharchives.6911

Schreiner, M. R., Quevedo Pütter, J., & Rebholz, T. R. (2023a). *Dataset for: Time for an update: Belief updating based on ambiguous scientific evidence* [Data set]. PsychArchives. https://doi.org/10.23668/psycharchives.12877

Schreiner, M. R., Quevedo Pütter, J., & Rebholz, T. R. (2023b). *Experiment files for: Time for an update: Belief updating based on ambiguous scientific evidence*. PsychArchives. https://doi.org/10.23668/psycharchives.12878

Schreiner, M. R., Quevedo Püter, J., & Rebholz, T. R. (2023c). *Supplemental materials for: Time for an update: Belief updating based on ambiguous scientific evidence*. PsychArchives. https://doi.org/10.23668/psycharchives.12880

Schreiner, M. R., Quevedo Pütter, J., & Rebholz, T. R. (2024). *R analysis and power simulation code for: Time for an update: belief updating based on ambiguous scientific evidence*. PsychArchives. https://doi.org/10.23668/psycharchives.14106

Schultze, T., Rakotoarisoa, A.-F., & Stefan, S.-H. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision Making*, *10*(2), 144–171. https://doi.org/10.1017/S1930297500003922

Simonovic, N., & Taber, J. M. (2022). Psychological impact of ambiguous health messages about COVID-19. *Journal of Behavioral Medicine*, *45*(2), 159–171. https://doi.org/10.1007/s10865-021-00266-2

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805. https://doi.org/10.1037/a0015145

Su, S. (2022). Updating politicized beliefs: How motivated reasoning contributes to polarization. *Journal of Behavioral and Experimental Economics*, *96*, Article 101799. https://doi.org/10.1016/j.socec.2021.101799

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, *117*(14), 7672–7683. https://doi.org/10.1073/pnas.1913678117

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science*, *11*(4), 454–463. https://doi.org/10.1177/1948550619877412

History
Received May 26, 2023
Revision received March 14, 2024
Accepted March 21, 2024
Published online September 6, 2024

Acknowledgments
We thank Mandy Hütter for helpful comments regarding the study design and research questions. We further thank Alicia Gernand, Bastian Schnieders, Eva Pöhlmann, and Timo Seitz for testing the experiment, and Coralie Notarbartolo Pfeifer, Marie Mundt, Rabea Krone, and Asheley Landrum for assisting in translating the study materials.

Conflict of Interest
We have no conflict of interest to declare.

Publication Ethics
Informed consent was obtained from all participants included in the study. All procedures were performed in accordance with the ethical standards of the Human Research Ethics Committee of the University of Mannheim.

Authorship
Marcel R. Schreiner, conceptualization, methodology, formal analysis, investigation, project administration, software, resources, validation, visualization, data curation, writing - original draft, Writing - review & editing; Julian Quevedo Pütter, conceptualization, investigation, software, resources, validation, writing - original draft, writing - review & editing; Tobias R. Rebholz, conceptualization, methodology, formal analysis, software, validation, visualization, writing - original draft, writing - review & editing. All authors approved the final version of the article.

Open Science
Open Data: The information needed to reproduce all of the reported results is available at https://doi.org/10.23668/psycharchives.12877 (Schreiner et al., 2023a).

Open Materials: The information needed to reproduce all of the reported methodology is available at https://doi.org/10.23668/psycharchives.12878 (Schreiner et al., 2023b), https://doi.org/10.23668/psycharchives.14106 (Schreiner et al., 2024), and https://doi.org/10.23668/psycharchives.12880 (Schreiner et al., 2023c).

Preregistration and Analysis Plan: This study was preregistered at http://dx.doi.org/10.23668/psycharchives.6911 (Schreiner et al., 2022).

The online supplementary materials are available at https://doi.org/10.23668/psycharchives.12880 (Schreiner et al., 2023c).

Funding
This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2277 "Statistical Modeling in Psychology." Open access publication enabled by University of Mannheim.

ORCID
Marcel R. Schreiner
https://orcid.org/0000-0001-7719-6281
Julian Quevedo Pütter
https://orcid.org/0000-0002-7340-9937
Tobias R. Rebholz
https://orcid.org/0000-0001-5436-0253

Marcel R. Schreiner
Department of Psychology
University of Würzburg
Röntgenring 11
97070 Würzburg
Germany
marcel.schreiner@uni-wuerzburg.de

Appendix

Specification of the Mixed-Effects Regression Model

We used the posterior beliefs Po_{ij} of participants $i = 1, \ldots, N$ about items $j = 1, \ldots, M$ as a dependent variable for a mixed-effects regression model based on Rebholz et al. (2024). The judgment level of the model was specified as:

$$Po_{ij} = Pr_{ij} + w_{ij}\left(E_{ij} - Pr_{ij}\right) + b_{WLS}L_{ij} + \varepsilon_{ij}, \qquad (1)$$

where Pr_{ij} denotes participants' prior beliefs and E_{ij} the presented evidence ratio. Moreover, for testing the weight-last-stronger strategy, the valence of the evidence of the last study is included as contrast-coded (i.e., -0.5 for negative and 0.5 for positive evidence) fixed effect b_{WLS} at the judgment level. Sum-to-one constraining implied that the weighting of prior beliefs Pr_{ij} is complementary to the weighting of the presented evidence ratio E_{ij}, that is, equal to $1 - w_{ij}$. By means of multilevel modeling, the residuals of the coefficients can be disentangled from the overall error at the judgment level $\varepsilon_{ij} \sim N(0, \sigma^2)$ (Baayen et al., 2008; Brown et al., 2018; Raudenbush & Bryk, 2002). In other words, mixed-effects regression weights per participant and item for the presented evidence ratio were specified at the weighting level of the model as follows:

$$\begin{aligned} w_{ij} = b_0 + a_i^S + a_j^T + b_{Edu}Edu_i + b_{SL}SL_i \\ + b_{SE}SE_{ij} + b_{TP}TP_i + b_{NS}NS_{ij} \\ + b_{uni}\left|E_{ij} - Pr_{ij}\right| + b_{UWS}\left(E_{ij} - 0.5\right)^2, \end{aligned} \qquad (2)$$

where b_0 denotes the sum-to-one-constrained fixed effect of the presented evidence ratio on participants' posterior beliefs. Moreover, $a^q \sim N(0, \tau_q^2)$ denotes the random effects of participants $q = S$ and stimulus items $q = T$, respectively, with τ_S^2 and τ_T^2 mutually independent. The hypothesized predictors are included as interaction terms at the judgment level, as can be seen by plugging Equation 2 into Equation 1. Thus, the respective fixed effects b_{SE} for subjective expertise SE_{ij}, b_{TP} for trust in (psychological) science TP_i, and b_{NS} for number of studies NS_{ij} measure the hypothesized influences on participants' evidence weighting. Similarly, to test the unidimensional and uncertainty-weighting updating strategies, we additionally included interactions with the absolute distance of a presented evidence ratio from prior beliefs, $\left|E_{ij} - Pr_{ij}\right|$, as well as its squared distance from ambiguity, $(E_{ij} - 0.5)^2$, respectively. Finally, we also included scientific literacy SL_i and education level Edu_i as control variables at the weighting level.