The truth revisited: Bayesian analysis of individual differences in the truth effect

Martin Schnuerch[1], Lena Nadarevic[1], & Jeffrey N. Rouder[2]

[1] University of Mannheim

[2] University of California, Irvine

Author Note

Correspondence concerning this article should be addressed to Martin Schnuerch, B6, 30-32, 68159 Mannheim. E-mail: martin.schnuerch@psychologie.uni-mannheim.de

Abstract

The repetition-induced truth effect refers to a phenomenon where people rate repeated statements as more likely true than novel statements. In this paper we document *qualitative* individual differences in the effect. While the overwhelming majority of participants display the usual *positive* truth effect, a minority are the opposite—they reliably discount the validity of repeated statements, what we refer to as *negative* truth effect. We examine 8 truth-effect data sets where individual-level data are curated. These sets are composed of 1,105 individuals performing 38,904 judgments. Through Bayes factor model comparison, we show that reliable negative truth effects occur in 5 of the 8 data sets. The negative truth effect is informative because it seems unreasonable that the mechanisms mediating the positive truth effect are the same that lead to a discounting of repeated statements' validity. Moreover, the presence of qualitative differences motivates a different type of analysis of individual differences based on ordinal (i.e., Which sign does the effect have?) rather than metric measures. To our knowledge, this paper reports the first such reliable qualitative differences in a cognitive task.

Keywords: Individual differences, qualitative differences, truth effect, hierarchical models, Bayesian model comparison

The truth revisited: Bayesian analysis of individual differences in the truth effect

In the usual course of experimental psychology, we often understand phenomena by computing the mean effect. This mean effect may be used to compute effect sizes or statistical tests, and the resulting inferences are about the mean level in the population. In our view, this focus on the mean makes sense when all people experience a phenomenon in a qualitatively similar way. For example, suppose we ask people to identify a briefly-presented and subsequently masked letter. In this case, increasing the stimulus duration of the letter should affect every individual in the same direction, namely that longer durations correspond to better performance. It seems implausible in fact for any person's true performance to decreases with increasing stimulus duration. And it is in this sense, where we can be almost sure that a phenomenon affects people in a qualitatively similar manner, that recourse to the mean seems judicious.

What happens if a treatment affects different people differently? A good example might be the effect of aspirin. For most people, the drug aspirin safely relieves pain. Yet, a minority of the population are allergic to aspirin, and for these people, the allergic reaction may be serious. In this case, questions about the mean response seem unimportant. Instead, the important questions are what proportion of the population is allergic and what are the seperate mechanisms of pain relief and allergic reactions.

The question of whether the overall mean is useful hinges on whether an effect is qualitatively consistent across individuals. In the first example, it seems implausible that increasing the stimulus duration of a briefly flashed and subsequently masked object could decrease identification for anyone. For this example, the average performance gain as a function of stimulus duration seems a reasonable target for inquiry. For the aspirin example, however, average gain in pain relief seems far less helpful.

The key methodological question is how to tell if an effect is qualitatively consistent

across a set of participants. Progress on this question has been made by Haaf and Rouder and colleagues (Haaf & Rouder, 2017, 2019; Thiele et al., 2017). We will review their approach subsequently, but for now, their research has yielded a startling finding. When it comes to performance tasks, it seems that people don't qualitatively differ. For example, nobody truly responds quicker to incongruent items in Stroop, Simon, or flanker tasks. Indeed, we are previously unaware of any case where substantial qualitative differences appear in a performance task. And this paper is the first we are aware of that shows such differences.

In this paper, we explore individual differences in a truth-judgment task. More precisely, the target of inquiry is a popular psychological effect, the repetition-based *truth effect*. In a typical truth-effect task, participants rate how likely it is that a particular statement is true. The critical manipulation is repeating some statements, and these repeated statements are more likely rated as true than novel ones. The real world impact of the truth effect is obvious: If a lie is repeated, it is more likely to be believed.

The question we ask is whether all people are susceptible to what we call a *positive* truth effect where repeated statements are judged as more valid than novel ones. The alternative is that some people have a true *negative* truth effect where they tend to discount the validity of repeated statements. Here, we analyze data from eight previous experiments spanning 1,105 participants and 38,904 trials. We show to our surprise that in all the data sets where variation is detectable, there are some people who have a reliably negative truth effect. Most people show a positive truth effect, but a small minority truly discount the validity of repeated information.

The Truth Effect

Repeatedly encountering a piece of information is likely to increase the subjective belief in its validity. This phenomenon has been known – and used – for thousands of years.

Around 150 BC, the Roman politician Cato reportedly concluded each of his speeches with the same sentence: "Carthago delenda est" (*Carthage must be destroyed*). It seems that he succeeded in convincing the Roman Senate to approve of his proposition: In 146 BC, Carthage was destroyed. Cato's strategy might have resonated well with French military leader Napoleon Bonaparte, who conquered large parts of Europe in the early 19th century. According to Bonaparte, "there is only one figure in rhetoric of serious importance, namely, repetition" (Le Bon, 1895, p. 125). Roughly 100 years later, Nazi German Minister of Propaganda Joseph Goebbels made use of this rhetorical device; "Repeat a lie often enough and it becomes the truth", is typically attributed to Goebbels (Stafford, 2016).

The effect of repetition on the perception of a proposition's truth is a well-documented phenomenon in experimental psychology. In a seminal study, Hasher et al. (1977) asked participants to provide truth ratings for trivia statements in successive sessions. Critically, some of the statements were repeated across sessions while others were novel. The authors found validity ratings for repeated statements to increase, independent of the statements' actual validity, while ratings for novel statements did not change.

Over the past 40 years, the truth effect has been replicated numerous times (Dechene et al., 2010; Unkelbach et al., 2019). It has been shown to be robust across different experimental designs, material, and instructions. Even explicit warnings about the effect do not eliminate it, but only reduce it (Nadarevic & Aßfalg, 2017). The consequences of such a robust cognitive bias are evident: If used strategically, repeated dissemination increases belief even in false information (Lazer et al., 2018; Pennycook et al., 2018; Unkelbach et al., 2019).

Meta-analytic results indicate that the truth effect is stable across studies, and is medium in size (Cohen's $d \approx 0.50$; Dechene et al., 2010). Consequently, the truth effect has been well-established for the average of individuals. In contrast, we are aware of only a small number of published studies on individual differences (Arkes et al., 1991; Boehm, 1994; Brashier et al., 2017; De keersmaecker et al., 2020; Newman et al., 2020; Parks & Toth,

2006). All of these studies assessed the covariation of individual truth effects and certain person-specific variables (e.g., age, need for cognition, and cognitive style). Yet, correlational analyses do not address the main question here, namely: Are individual differences only quantitative, or qualitative?

Quantitative differences occur if all participants provide somewhat higher truth ratings for repeated than for novel statements. We might reasonably assume a common process underlying the effects in this case, and the mean might even be an adequate representation for understanding this process. The assumption is less reasonable, however, if differences are *qualitative*, that is, if some participants were to depreciate the validity of repeated statements. Indeed, qualitative individual differences are precedented in the domain of truth judgments (i.e., belief polarization; Cook & Lewandowsky, 2016). Is it still the same process that leads some people to increase their belief in repeated statements and others to decrease it? If qualitative differences can be shown, this has theoretical implications. A theory of the truth effect would have to account for both the increase and the decrease in beliefs due to repetition. Therefore, to gain constraint on theory, a fundamental question in the analysis of the repetition-based truth effect is what Haaf and Rouder (2017) coined the "Does everybody?" question: Does everybody show a positive truth effect?

This fundamental question comes with a methodological challenge: Even if we observe a negative truth effect for some individuals, this observation might reflect sampling noise rather than true qualitative differences. How do we assess whether people truly differ and, if so, whether these differences are qualitative? To answer these questions, we follow the strategy proposed by Haaf and Rouder (2017, 2019). We develop a set of hierarchical Bayesian models that represent different structures of individual differences. By means of model comparison, we then directly assess the evidence from data about the nature of individual differences.

In the following, we provide a brief description of the eight data sets that we reanalyze

in this study. We then develop the statistical models of individual-differences structures and outline the procedure to quantify evidence for these models. With model comparison, we find a surprising result: Across many of the data sets, there is a small proportion of individuals that show a negative truth effect.

Data Sets

We reanalyzed eight data sets from previous truth-effect experiments, all of which are publicly available from the Open Science Framework (OSF).[1] Six of the sets have been published in peer-reviewed articles; two have been published only on OSF. Detailed information about the data and the experiments can be obtained from Appendix A.

All data sets are based on a common experimental design with three phases. Phase 1 is the *exposure phase*: Participants see a number of trivia statements and, typically, assign each statement to a semantic domain (e.g., biology, geography, sports) or rate each for interest. Phase 2 is a retention interval, in wich participants may perform an unrelated task. This phase can range from a few minutes to several days. Phase 3 is the critical *judgment phase* where participants rate the validity of statements. Ratings are typically given on a Likert scale, for example, from 1 ("definitely false") to 6 ("definitely true"). Critically, half of the statements have been presented during the exposure phase and half are new. The truth effect is measured as the difference between mean truth ratings for repeated and for new statements, $M_{rep} - M_{new}$.

For an overview of sample characteristics and results of the eight data sets, see Table 1. For the sake of comparison, we rescaled truth ratings to range from -1 ("definitely false") to 1 ("definitely true"). As a consequence, the truth effect can range from -2 to 2. If all repeated statements received truth ratings of 1 while all new were rated as -1, the resulting

---

[1] Set 1: https://osf.io/6wv4z/; Set 2: https://osf.io/3uaj7/; Sets 3 & 4: https://osf.io/5pfa2/; Sets 5 & 6: https://osf.io/eut35/; Set 7: https://osf.io/b4szp/; Set 8: https://osf.io/txf46/

Table 1

Summary of the Data Sets.

Set	Source	N	I	Mean Effect*	t Test	Cohen's d
1	Nadarevic et al. (2012)	267	20	0.20 (0.34)	$t(266) = 9.47$	0.58
2	Nadarevic and Rinnewitz (2011)	139	20	0.28 (0.36)	$t(138) = 9.16$	0.78
3	Nadarevic and Aßfalg (2017), Exp. 1	33	88	0.11 (0.10)	$t(32) = 6.54$	1.14
4	Nadarevic and Aßfalg (2017), Exp. 2	98	80	0.20 (0.23)	$t(97) = 8.61$	0.87
5	Nadarevic and Erdfelder (2014), Exp. 1	85	88	0.05 (0.10)	$t(84) = 5.09$	0.55
6	Nadarevic and Erdfelder (2014), Exp. 2	35	88	0.07 (0.17)	$t(34) = 2.51$	0.42
7	Brashier et al. (2020), Exp. 1	52	120	0.09 (0.12)	$t(51) = 5.52$	0.77
8	Pennycook et al. (2018), Exp. 1	396	20	0.14 (0.27)	$t(395) = 10.08$	0.51

Note. N = Number of participants; I = Number of statements rated per participant; t values and degrees of freedom are based on paired t tests. *Standard deviation is given in parentheses.

truth effect would be 2. An effect of -2, in contrast, would indicate a perfect reversal of the truth effect. Zero represents the absence of any effect.

Figures 1 and 2 (left columns) show the individual truth effects in all eight data sets (black line). Individuals are sorted by the size of their effect, going from the most negative to the most positive. The red line indicates if the individual effect is below 0, that is, negative observed truth effects. The grey-shaded area surrounding the line denotes 95% confidence intervals. The average effect across all people is given by the dashed horizontal line. In all data sets, we observe considerable differences between individual participants.
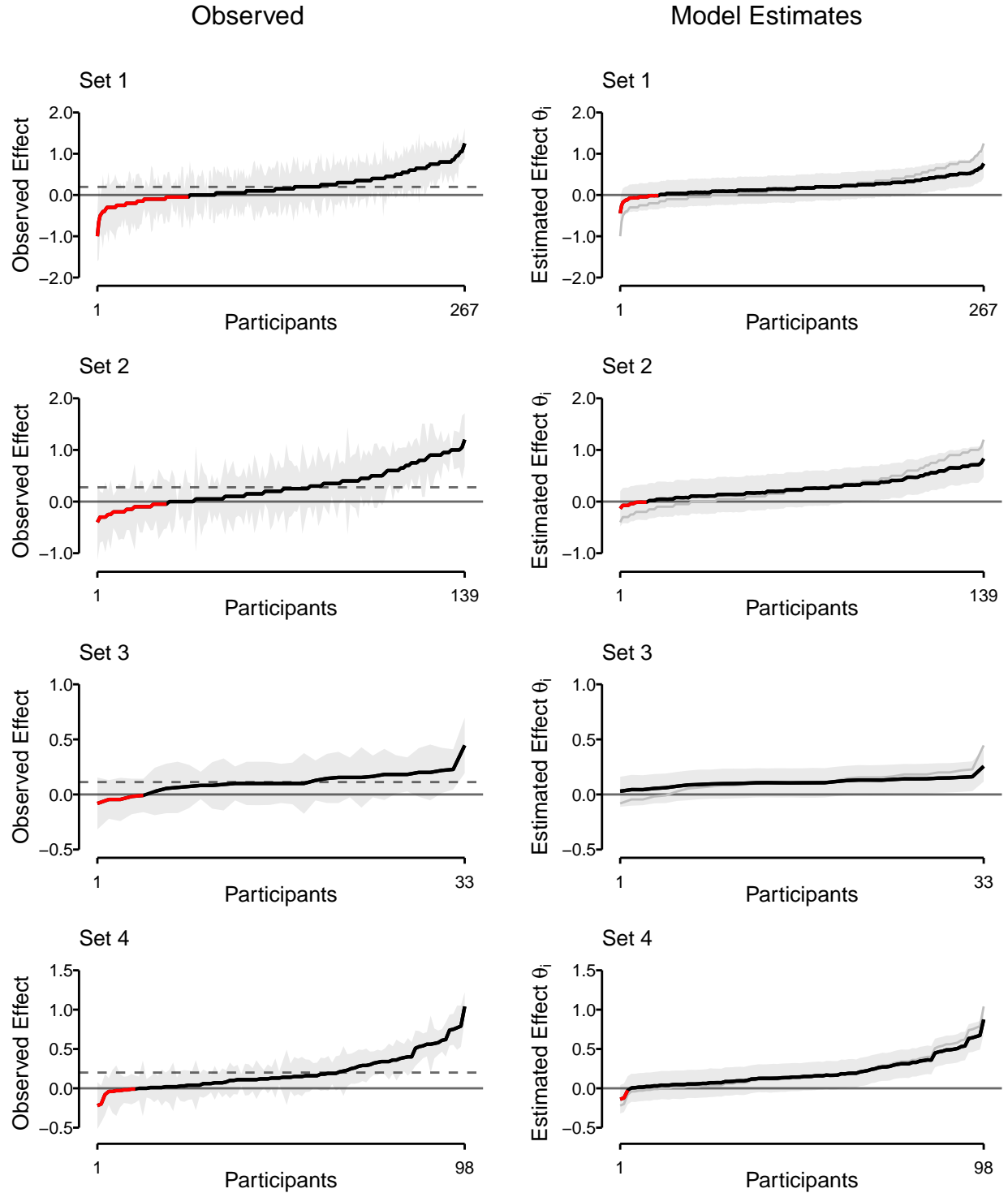
Figure 1. Observed (left column) and estimated (right column) individual truth effects for Data Sets 1–4, ordered by observed effect size. On the left side, the shaded area denotes individual 95% confidence intervals. The dashed line represents average observed effects. On the right side, the shaded area denotes the 95% credible interval. The grey line represents observed truth effects. Negative observed and estimated effects (i.e., higher truth ratings for novel than for repeated statements) are denoted by red color on both sides.
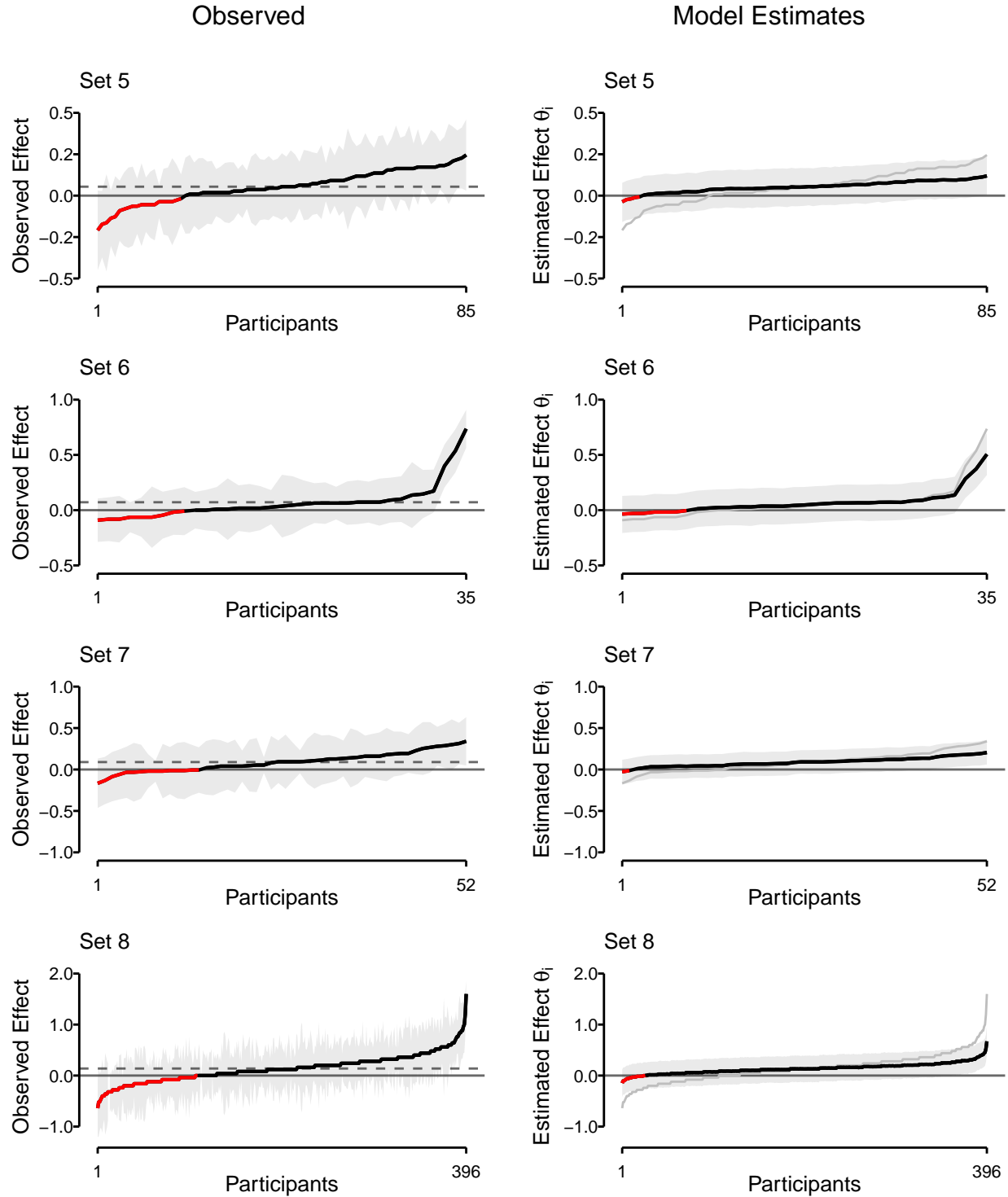
Figure 2. Observed (left column) and estimated (right column) individual truth effects for Data Sets 5–8, ordered by observed effect size. On the left side, the shaded area denotes individual 95% confidence intervals. The dashed line represents average observed effects. On the right side, the shaded area denotes the 95% credible interval. The grey line represents observed truth effects. Negative observed and estimated effects (i.e., higher truth ratings for novel than for repeated statements) are denoted by red color on both sides.

Statistical Models

The main substantive question is whether individual differences are quantitative or qualitative. Our strategy in answering this question is to implement each of these positions in statistical models, and then compare the models in light of data with Bayes factors. The specific models come from Haaf and Rouder (2017), and, consequently, we provide only a brief overview here. Let Y_{ijk} denote the truth judgment of the ith person $(i = 1, \ldots, I)$ for the jth statemt $(j = 1, \ldots, J)$ in condition k $(k = 1, 2$ for new versus repeated, respectively). Note that not every statement j is necessarily seen by each participant i. Consequently, the data sets do not contain Y_{ijk} for every possible combination of i, j, and k. This fact presents no problem in analysis.

We specify the following linear model on the dependent variable:

$$Y_{ijk} \overset{\text{ind}}{\sim} \text{Normal}(\mu + \alpha_i + t_j\beta + x_k\theta_i, \ \sigma^2). \tag{1}$$

In this model, μ denotes the grand mean intercept and α_i is a person-specific deviation from this grand mean. The term t_j codes the truth status of the jth statement, which can either be 0 if it is false, or 1 if it is true. Hence, β denotes the effect of a statement's factual truth on the judgment. The term x_k codes the repetition condition, which can either be 0 if the statement is new or 1 if it is repeated. Consequently, θ_i denotes the ith individual's truth effect, and this parameter is the main target of inquiry. The last term, σ^2, denotes the sampling variance of observed values. The main theoretical positions about individual differences motivate the following four models on θ_i:

The Unconstrained Model

The unconstrained model, $\mathcal{M}_u$, does not impose any constraints on the individual effects. It may be used to capture qualitative individual differences:

$$\mathcal{M}_u: \theta_i \overset{\text{iid}}{\sim} \text{Normal}(\nu, \ \delta^2). \tag{2}$$

In this model, ν and δ^2 denote the mean and variance of individual effects. These group-level parameters are estimated from the data.

The Positive-Effects Model

The positive-effects model, $\mathcal{M}_+$, is less flexible. It only allows for positive individual effects:

$$\mathcal{M}_+: \ \theta_i \overset{\text{iid}}{\sim} \text{Normal}_+(\nu, \ \delta^2). \tag{3}$$

The distribution denoted by Normal_+ is a truncated normal with a lower bound at zero. Thus, the model naturally incorporates the constraint that individuals may differ but they are all in the same predicted direction. Substantively, this model implies that differences are quantitative, but not qualitative.

Common-Effect Model

The critical specification in the common-effect model, $\mathcal{M}_1$, is that all individuals share one common effect:

$$\mathcal{M}_1: \ \theta_i = \nu. \tag{4}$$

Accordingly, there are no true individual differences in the truth effect. Any observed variation would thus be due to sampling noise.

Null Model

The final model is a null model, $\mathcal{M}_0$, where there is no truth effect at all:

$$\mathcal{M}_0: \ \theta_i = 0. \tag{5}$$

Accordingly, any observed effects of statement repetition are due to sampling noise.

Figure 3 (left column) illustrates the four models. On the x-axis, the true effect of a hypothetical participant is shown, θ_1. On the y-axis, the true effect of a second participant is shown, θ_2. The null model specifies that both effects are 0, thus, the model is represented by

a point at the origin. The common-effect model does not restrict true effects to one value, but specifies that all individual effects are identical. This is represented by the diagonal line. No equality constraints are imposed in the positive-effects model, but all true individual effects are defined as larger than zero. Accordingly, θ_1 and θ_2 are free to vary in the upper-right quadrant. Finally, the unconstrained model puts no restrictions on individual effects; the model for the two hypothetical participants is represented by a bivariate normal centered at the origin.

Prior Specification

We implement these models in the Bayesian framework, and, as such, priors are needed on parameters. For some parameters, those that are common in all four models, the priors may be set without undue influence on the posteriors or model comparison statistics. These specifications are provided in Appendix B. For other parameters, however, those that vary across models (θ, ν, δ^2), prior settings are important and are discussed here.

We follow the common g-prior specification approach (Zellner, 1986), which is based on placing priors on effect sizes. The setup is described in detail in Haaf and Rouder (2017) and Rouder et al. (2012). Let g_θ be a signal-to-noise ratio defined as $g_\theta = \delta^2/\sigma^2$. This is an effect-size description of θ; it describes how much true variability there is across people relative to the variability in observations. With this parameter, we may write $\theta_i \sim \text{Normal}(\nu, g_\theta \sigma^2)$. Priors are needed on ν and g_θ. The prior on ν is also scaled to the variability in observations: $\nu \sim \text{Normal}(0, g_\nu \sigma^2)$, and there is a new parameter g_ν. Priors on these g parameters are Inverse-χ^2 distributions with one degree of freedom and a scale parameter r^2:

$$g_\nu \sim \text{Inverse-}\chi^2(r_\nu^2),$$
$$g_\theta \sim \text{Inverse-}\chi^2(r_\theta^2). \tag{6}$$

Researchers need to set the scales of these priors before the analysis. We advocate that
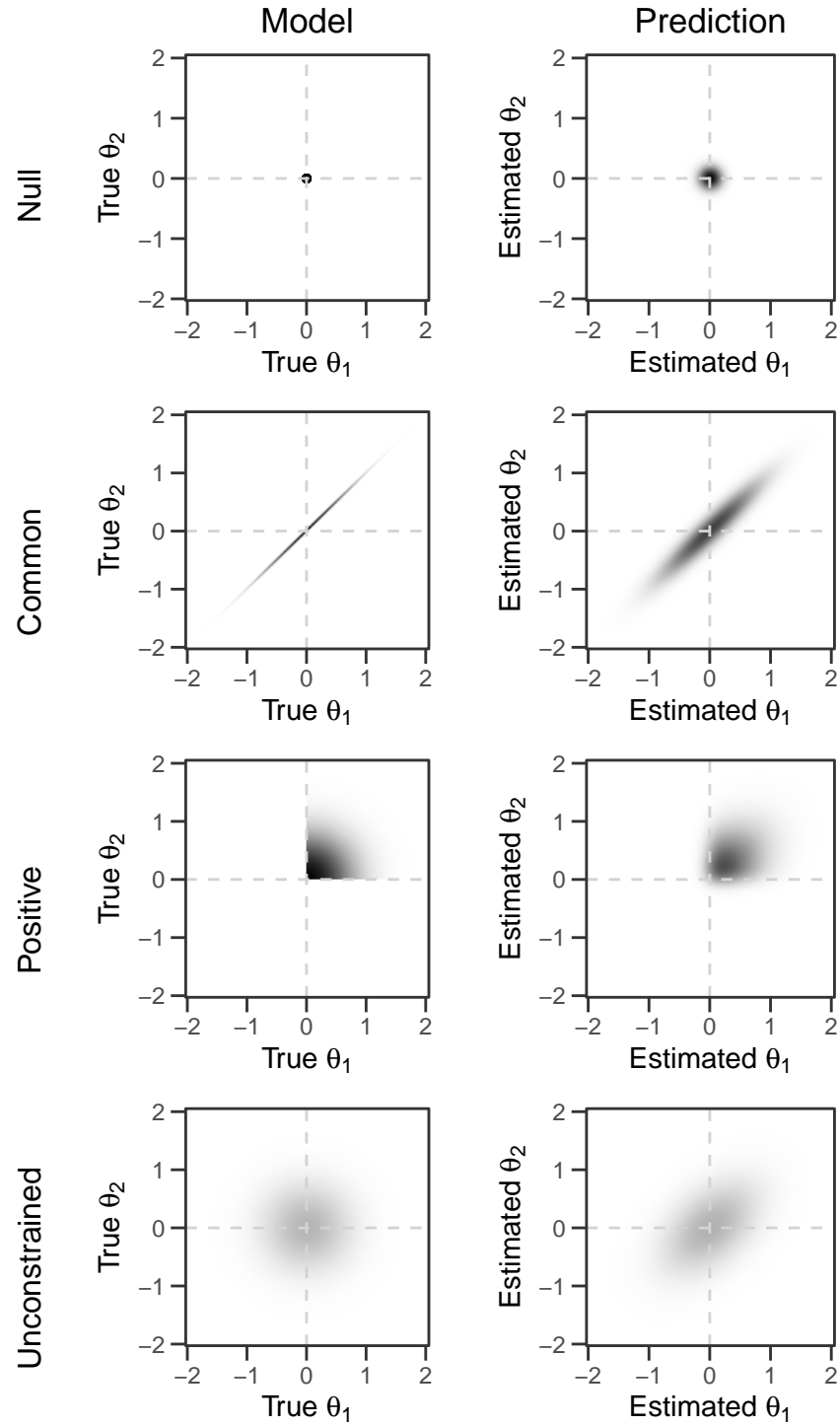
Figure 3. Illustration of the four models (left column) on θ_i and corresponding predictions (right column) for two hypothetical participants.

doing so should rely on substantive considerations rather than statistical arguments. Here is our line of thought: In our experience, on a standardized scale of -1 to 1, truth judgments' trial-by-trial variability covers about a quarter of the scale, that is, $\sigma = 0.50$. As a reference, on a scale from 1 to 6, this corresponds to a standard deviation of $\sigma = 1.50$. When specifying r, it is helpful to consider that it represents an expectation about the variability of the parameter relative to σ. For example, a value of $r_\theta = 1$ encodes the belief that the variability of person-specific truth effects (i.e., δ) is comparable to the trial-by-trial variability. Likewise, $r_\theta = 1/2$ or $r_\theta = 2$ represent the expectation that δ scales about half or about twice as large as σ, respectively. Figure 4 illustrates the effect of different choices of r_θ. It shows the resulting prior distributions on the variability of θ_i conditional on a trial-by-trial variability of $\sigma = 0.50$.
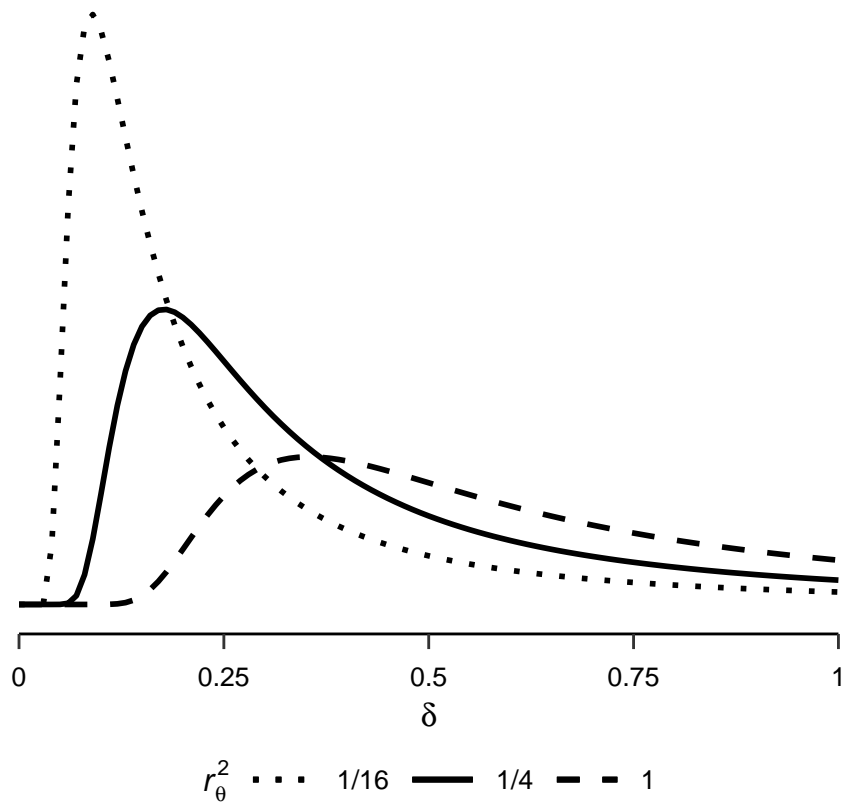

Figure 4. Prior distributions on δ, the variability of θ_i, for different scale settings. A trial-by-trial variation of $\sigma = 0.50$ is assumed.

Based on this information, how do we choose the scale parameters? The average truth effect is of medium size, Cohen's $d = 0.50$ (Dechene et al., 2010). Assuming $\sigma = 0.50$, the expected observed effect on a rating scale of -1 to 1 is thus 0.25. We find it reasonable to expect that the variability of this effect is of a comparable magnitude. Therefore, we define the scales of our g priors on θ_i and ν such that they place most weight on values around half the trial-by-trial variability, that is, $r_\nu^2 = r_\theta^2 = (1/2)^2$.

Model Comparison

The models may be used to answer the main substantive question about qualitative individual differences. One approach is simply to estimate individual effects (θ_i) in the full model. Yet, we think this approach is ultimately unhelpful. The problem is that whether there are qualitative differences between individuals is a global property rather than an individual one. One may know that someone must be negative without the ability to identify who. Consequently, individual estimates are not helpful in themselves. And population-level mean effects are not helpful either as they do not address the distinction between qualitative and quantitative individual differences. Hence, the key to assessment here may be made with model comparison and not with estimation.

A leading approach to model comparison in Bayesian analysis is the *Bayes factor* (Jeffreys, 1935, 1961). Bayes factors measure the relativ strength of evidence for models by comparing how well these models predict the data (Kass & Raftery, 1995; Rouder & Morey, 2017). Figure 3 (right column) illustrates the predictions that the different models make for observed data of the two hypothetical participants. These predictions are noisy versions of the structure on true values. The more flexible a model is, that is, the fewer restrictions it imposes on the structure of individual differences, the more diffuse are its predictions. Hence, models are penalized for flexibility.

Analysis for the null model, the common-effect model, and the unconstrained model

have been developed by Rouder et al. (2012). Their approach is implemented in the *BayesFactor* package (Morey & Rouder, 2015) in R (R Core Team, 2019), which allows for fast and accurate calculation of Bayes factors for three of the four models. This development, however, does not apply to the positive-effects model. Therefore, we calculated the Bayes factors between the positive-effects and the unconstrained model using the *encompassing prior* method proposed by Klugkist and colleagues (Klugkist & Hoijtink, 2007; Klugkist et al., 2005). The combination of these two approaches is straightforward (e.g., Haaf & Rouder, 2017, 2019).

<div align="center">

Evidence for Qualitative Differences

</div>

Model Convergence

Posterior distributions for all parameters in the unconstrained model are obtained by Markov chain Monte Carlo (MCMC) sampling within the BayesFactor package. We checked model convergence by inspecting MCMC chains and computing autocorrelations for critical parameters (i.e., ν, θ, g_ν, and g_θ). As in previous applications, the models converged fast and the chains mixed well. The autocorrelations for even the slowest converging parameters were inconsequential compared to the large number of posterior samples (10,000).

Estimation

Individual truth effect estimates from the unconstrained model for all eight data sets are shown in the right columns of Figures 1 and 2. The black line denotes the posterior means of θ_i for each participant. The grey band around this line is the 95% credible interval, that is, an interval that contains 95% of the posterior samples. The grey line represents observed individual truth effects and the ordering is obtained from these observed values (see left columns).

The first aspect to note is the effect of the hierarchical model specification on individual estimates. The effect is called shrinkage: Individual estimates inform each other

and thus, outliers are pulled (*shrunk*) toward the mean. This shrinkage is clearly visible in the estimates; there is less variability than in the observed effects. Note that the credible intervals are much smoother than the individual confidence intervals of observed effects (left column), reflecting regularization from the homogenous variance specification. The second aspect to note is that even with the shrunk estimates, considerable true individual differences remain. And the third aspect, perhaps the most consequential, is that some of these shrunk estimates are negative.

Model Comparison

Table 2 summarizes the results of the Bayes factor model comparison for the eight data sets. In each column, an asterisk marks the *preferred* model, the one for which the data provided the most evidence. The other cells in each column show the Bayes factors between the remaining models and this preferred model. Because these remaining models are less preferred, the Bayes factors are always less than one, mostly by many orders of magnitude.

In five data sets, we find strong evidence for qualitative differences: In Sets 1, 2, 4, 6, and 8, the Bayes factors in favor of the unconstrained model $\mathcal{M}_u$ are at least 1000-to-1 over the next leading competitor. As $\mathcal{M}_u$ is the only model that allows for qualitative individual differences, the Bayes factors provide compelling evidence for them. In these data sets, there must be some individuals with a true negative truth effect.

In the remaining three data sets, in contrast, we do not find evidence for this negativity. Interestingly, the preferred model in Sets 3, 5, and 7 is not $\mathcal{M}_+$, which allows for quantitative individual differences. Instead, it is $\mathcal{M}_1$, which specifies a common effect without individual differences. In Sets 3 and 5, there is strong evidence for the common-effect model; the Bayes factors in favor of $\mathcal{M}_1$ are at least three orders of magnitude over the next competitor. In Set 7, in contrast, the evidence is fairly ambiguous, indicating that the data do not contain sufficient resolution to adjudicate among the different models. In summary, whenever we find

individual differences, they are qualitative in nature rather than quantitative.

We find strong evidence for these differences in five data sets. In at least two, however, we find evidence against them. How do these two data sets, Sets 3 and 5, differ from the others? Looking for a psychological explanation, we note that in both sets the judgment phase was administered after a one-week retention interval, whereas it took place within one experimental session in all other sets. If this difference was systematic, the influence of retention-interval length could tell us something about the nature of individual differences. It is possible that differences in cognitive performance (e.g., source recollection; Begg et al., 1992) rather than personality underly qualitative differences in truth effects, and that these cognitive differences are affected for example by the length of the retention interval. Alternatively, however, there could be simple statistical reasons for this result: Data Set 3 is rather small, thus allowing for strong influence of shrinkage and making it difficult to evidence true individual differences should they exist. Therefore, we should be careful not to overinterpret the results. Any post-hoc explanation should be addressed and critically tested in future experiments.

Classifying Individuals

Who are these individuals that depreciate the validity of repeated statements? Posterior means, such as those in Figures 1 and 2, do not provide enough information for classification because classification should depend on the underlying variability. A better approach to classifying individuals with truly negative truth effects is to assess the posterior probability that an individual's estimate is less than 0.

Figure 5 shows the posterior probability of a *positive* truth effect for each individual, that is, $P(\theta_i > 0 | \text{Data})$. The red color denotes individuals with negative posterior means. As a means to classify people for the purposes of this article, we assign individuals with a probability of at least 3-to-1 to one of two groups: We define individuals with

Table 2

Bayes Factor Model Comparison

| | Data Set | | | | | | | |
Model	1	2	3	4	5	6	7	8
$\mathcal{M}_0$	3.2e-58	5.1e-57	4.1e-08	1.7e-124	8.9e-05	2.8e-07	3.8e-07	4.9e-41
$\mathcal{M}_1$	1.4e-20	8.8e-17	*	1.0e-49	*	5.9e-05	*	3.1e-09
$\mathcal{M}_+$	1.2e-03	1.1e-03	2.0e-03	1.1e-03	1.7e-07	8.7e-04	2.7e-01	1.2e-03
$\mathcal{M}_u$	*	*	1.9e-03	*	1.6e-04	*	2.0e-03	*

Note.   The preferred model for each data set is indicated by an asterisk. Remaining cells contain Bayes factors for each model against the preferred model. $\mathcal{M}_0$ = null model; $\mathcal{M}_1$ = common-effect model; $\mathcal{M}_+$ = positive-effects model; $\mathcal{M}_u$ = unconstrained model.

$P(\theta_i > 0|\text{Data}) \geq .75$ as *positive truthers*. In contrast, individuals with $P(\theta_i > 0|\text{Data}) \leq .25$ are *negative truthers*. We classify participants with $.25 \leq P(\theta_i > 0|\text{Data}) \leq .75$ as undecided. We acknowledge that this way to classify people is somewhat arbitrary. That notwithstanding, across all data sets, 70.05% of all participants are classified as positive truthers; only 1.09% are classified as negative truthers and 28.87% are classified as undecided.

In Sets 3 and 5, we found strong evidence for the absence of individual differences. Yet, while no individuals classify as negative truthers in these data sets, we still find 22.03% to be undecided. This result shows that classification may find differences even when a more global approach—model comparison—indicates that these differences are unwarranted. A somewhat complimentary state-of-affairs occurs for Data Set 6. Here, model comparison showed strong evidence for qualitative individual differences. Yet, we cannot classify anyone with certainty as a negative truther.

These discrepencies illustrate the difficulty with classification: When we apply a classification approach, we may classify individuals as different even when there are no true individual differences. This state occurs because classification is local to the individual, and as such, it is more susceptible to noise than Bayes factor assessment of global patterns. Conversely, we may know from the Bayes factor global assessment that a set may have at least one individual with a true negative effect. Yet, based on individual posterior probabilities, it may be difficult to know which one that is.

The aim of this paper is the global assessment of individual difference patterns in truth-effect experiments. If our overriding goal was to classify people, we could construct a latent-class classification model. In such a model, the normal in the unconstrained model could be replaced with a mixture of two states. One state would cover the positive truthers; and the distribution would be limited to positive true values. The other would cover the negative truthers, and the distribution would be limited to negative true values. If there is little mass toward zero, the model would have the effect of cleaving people clearly into two groups. A good example here is Houpt and Fifić (2017), who used this latent-class approach to classify people as using either serial or parallel processing in a systems factorial setting. The development of such a model is beyond the scope of this paper, but may prove useful in understanding the relationship between individual differences in the truth effect and other variables.

Sensitivity to Prior Settings

The Bayesian analysis presented herein requires the analyst to set a prior scale, denoted r^2, on the signal-to-noise ratio g. The dependence of Bayesian analysis on prior settings is frequently criticized as posing a threat as it provides for uncounted researcher degrees of freedom (Simmons et al., 2011). Indeed, it seems reasonable to require that for the same data set, different researchers should reach the same conclusions. Yet, almost all Bayesians note that priors have effects on inference. To align Bayesian inference with the
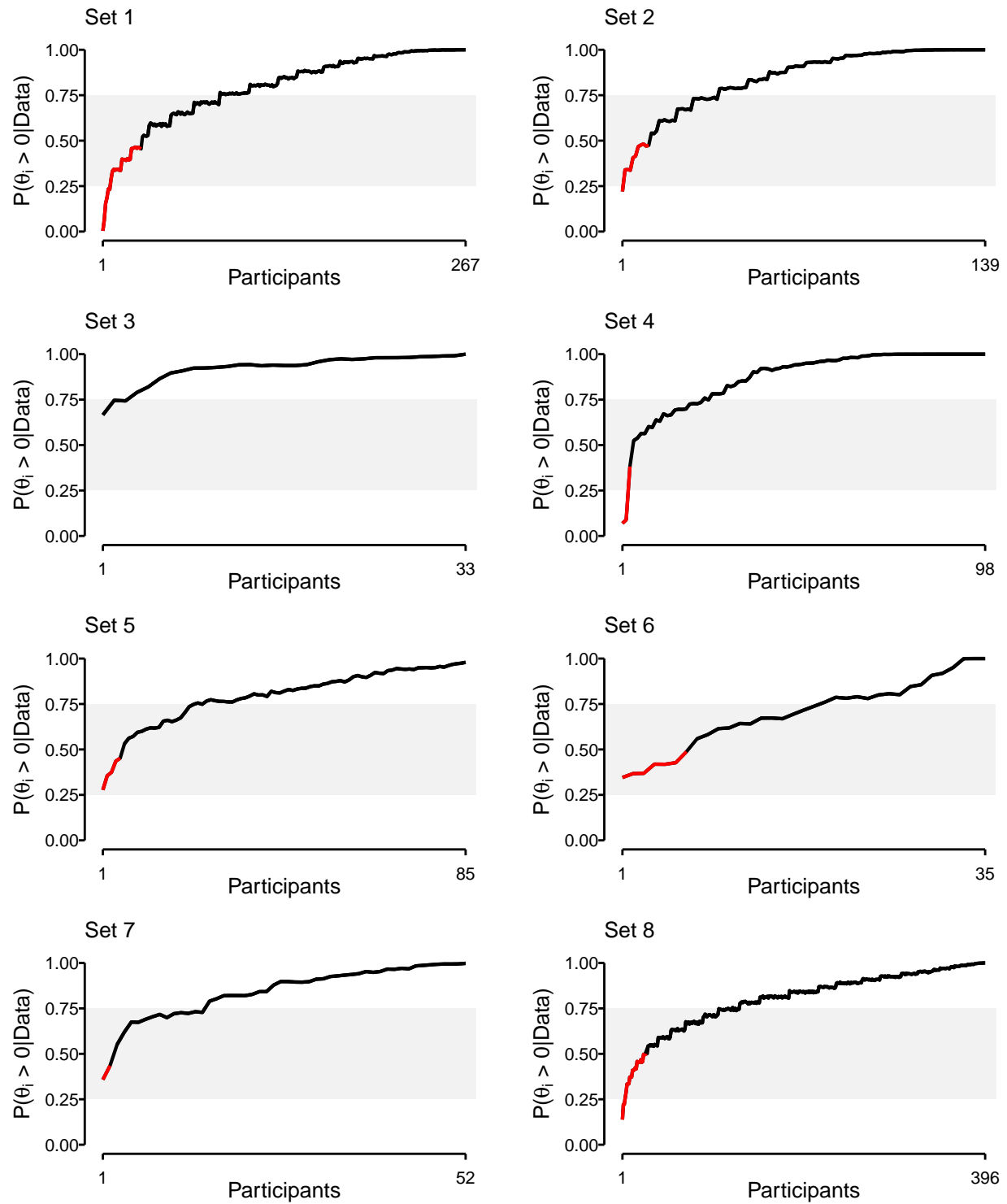
Figure 5. Individual posterior probabilities of a positive truth effect. Participants are ordered by observed effect size. The red line denotes individuals with negative posterior means (see Figures 1 and 2).
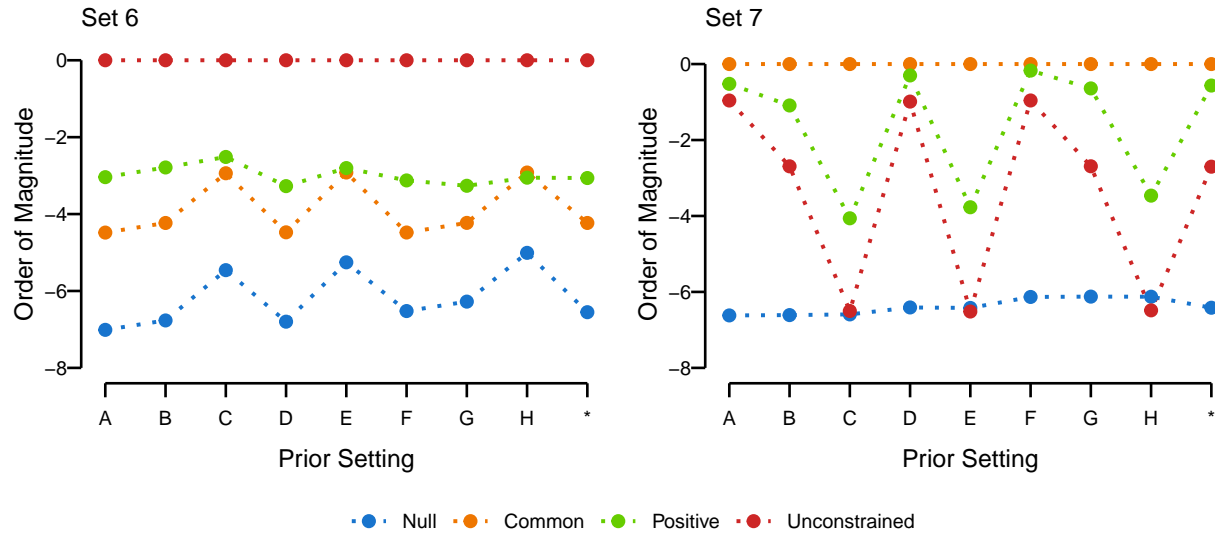
Figure 6. Sensitivity of Bayes factor model comparison to different prior settings. Shown are Bayes factors for each model against the preferred model. The asterisk denotes the prior setting used in the previous analysis. Details for each prior setting are shown in Table 3.

above desideratum, many Bayesian analysts actively seek to minimize the effects of prior settings (e.g., Aitkin, 1991; Gelman et al., 2004; Kruschke, 2013; Spiegelhalter et al., 2002).

We do not subscribe to the view that minimization of prior effects is necessary or even laudable. In fact, all reasonable statistical procedures that we are aware of require the researcher to make decisions that will affect the inference (e.g., choosing the sample size). The choice of prior settings is important because it affects the predictions that models make about data. Therefore, these settings that affect the predictive accuracy of a model *should* affect our opinions about it in light of data.

Thus, when different researchers use different priors, they may reach different opinions about the data. Rouder et al. (2016) argue that so long as various prior settings are *justifiable,* the variation in results should be embraced as the legitimate diversity of opinion. When reasonable prior settings result in conflicting conclusions, we may infer that the data do not afford the precision to adjudicate among competing positions.

With this argument in mind, we may assess whether reasonable variation in prior settings affects Bayes factor conclusions about the nature of individual differences in the truth effect for the current data. To that end, we repeated the above analysis with a number of different prior settings. The critical settings are r_ν and r_θ, which code the scale of effects. We set r_ν and r_θ to 0.50 in value, meaning that we expect the variation in ν and θ to be about half the variation in repeated observations. Here, we allow each of these settings to be this value, half this value, and twice this value; and the factoral combination yields nine possible settings. We computed the Bayes factors for all models for all nine settings for all data sets to understand how reasonable variation in prior settings affects inference.

For seven of eight data sets, model comparison was unaffected by reasonable variation in prior settings. As an illustration, the results for two data sets are depicted in Figure 6. The figure shows the Bayes factors for all models relative to the preferred one in the previous analysis. On the right is Data Set 7, the most concerning. Here, the common-effect model is preferred only by a negligible amount depending on the prior setting, indicating a lack of a clear verdict between the models. This lack of resolution holds only for this data set. The left panel shows the case for Data Set 6, and we chose this set because, outside of Data Set 7, Bayes factors were most dependent on prior settings. Even so, the unconstrained model is preferred over all other models by at least a factor of 100 across the range of reasonable prior settings. In the remaining data sets (not shown), there is even more stability of Bayes factors across the ranges. Hence, across reasonable variation in prior settings, Data Sets 1, 2, 4, 6, and 8 show strong evidence for qualitative individual differences. In a similar fashion, the common-effect model is clearly preferred for all prior settings in Data Sets 3 and 5. Overall, the results presented here are robust to a wide range of reasonable prior opinion.

General Discussion

In this paper, we show a surprising finding. Although the truth effect is reliably obtained across many data sets, the effect itself is inconsistent across people. We are

Table 3

Prior Settings for Sensitivity Analysis

	Prior Setting								
Parameter	A	B	C	D	E	F	G	H	*
r_ν	1/4	1/4	1/4	1/2	1/2	1	1	1	1/2
r_θ	1/4	1/2	1	1/4	1	1/4	1/2	1	1/2

Note.   The asterisk codes the prior setting used in the previous analysis.

confident that in most experiments some people truly judge repeated statements as more valid than novel ones, while others truly judge them as less so. This effect is not just noise—the models indicate that this inconsistency occurs above and beyond trial-by-trial variation. What makes the finding surprising to us is that the result is in contrast to previous work with these individual-difference models. The modal result is that "everybody does", that is, there are no qualitative individual differences in common cognitive effects such as Stroop and Flanker effects (Haaf & Rouder, 2017, 2019). In the repetition-based truth effect, these differences exist, and they occur consistently across several data sets.

Does the presence of qualitative individual differences inform current cognitive theories of the truth effect? We think it should. A number of theoretical explanations have been proposed for the repetition-based truth effect, for example, the recognition account (Bacon, 1979), the source-dissociation hypothesis (Arkes et al., 1991), the familiarity account (Begg et al., 1992), processing fluency (Reber & Schwarz, 1999), or the referential theory (Unkelbach & Rom, 2017). These accounts assume different underlying cognitive mechanisms, yet, they all make the same core prediction: repetition increases perceived validity. Unkelbach et al. (2019) summarize thusly: "No matter which mental processes may underly the repetition-induced truth effect, on a functional level, repetition increases

subjective truth" (p. 5). We argue, based on our analysis, that this statement is too general. In fact, we show what Davis-Stober and Regenwetter (2019) call the *paradox of converging evidence*: Across data sets, we find converging evidence that the statement holds on the mean level—yet, at the same time, we accumulate strong evidence that it doesn't hold for everybody. Consequently, our results present converging evidence against theoretical positions that do not account for negative truthers.

Open Practices Statement

All data used in this study are publicly available from their original OSF repositories (links have been provided in the text). The reanalysis was not preregistered. This paper is written in RMarkdown which includes all code for analyses and figures. The RMarkdown and all supporting files are curated at https://github.com/PerceptionAndCognitionLab/hc-truth/tree/public.

References

Aitkin, M. (1991). Posterior Bayes Factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(1), 111–128. https://doi.org/10.1111/j.2517-6161.1991.tb01812.x

Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, *27*(6), 576–605. https://doi.org/10.1016/0022-1031(91)90026-3

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(3), 241. https://doi.org/10.1037/0278-7393.5.3.241

Begg, I., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*(4), 446.

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin*, *20*(3), 285–293. https://doi.org/10.1177/0146167294203006

Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, *194*, 104054. https://doi.org/10.1016/j.cognition.2019.104054

Brashier, N. M., Umanath, S., Cabeza, R., & Marsh, E. J. (2017). Competing cues: Older adults rely on knowledge in the face of fluency. *Psychology and Aging*, *32*(4), 331–337. https://doi.org/10.1037/pag0000156

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186

Davis-Stober, C. P., & Regenwetter, M. (2019). The 'paradox' of converging evidence. *Psychological Review*, *126*(6), 865–879. https://doi.org/10.1037/rev0000156

De keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C.,
    & Roets, A. (2020). Investigating the robustness of the illusory truth effect across
    individual differences in cognitive ability, need for cognitive closure, and cognitive
    style. *Personality and Social Psychology Bulletin*, *46*(2), 204–215.
    https://doi.org/10.1177/0146167219853844

Dechene, A., Stahl, C., Hansen, J., & Wanke, M. (2010). The truth about the truth: A
    meta-analytic review of the truth effect. *Personality and Social Psychology Review*,
    14(2), 238–257. https://doi.org/10.1177/1088868309352251

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd
    edition)*. London, Chapman; Hall.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in bayesian mixed models.
    Psychological Methods, *22*(4), 779–798. https://doi.org/10.1037/met0000156

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of
    individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789.
    https://doi.org/10.3758/s13423-018-1522-x

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential
    validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112.
    https://doi.org/10.1016/S0022-5371(77)80012-1

Houpt, J. W., & Fifić, M. (2017). A hierarchical Bayesian approach to distinguishing serial
    and parallel processing. *Journal of Mathematical Psychology*, *79*, 13–22.
    https://doi.org/10.1016/j.jmp.2017.05.001

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability.
    Mathematical Proceedings of the Cambridge Philosophical Society, *31*(02), 203–222.
    https://doi.org/10.1017/S030500410001330X

Jeffreys, H. (1961). *Theory of probability*. New York, Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical
    Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379. https://doi.org/10.1016/j.csda.2007.01.024

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Le Bon, G. (1895). *The crowd. A study of the popular mind.* New York, Viking.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs.*

Nadarevic, L., & Aßfalg, A. (2017). Unveiling the truth: Warnings reduce the repetition-based truth effect. *Psychological Research*, *81*(4), 814–826. https://doi.org/10.1007/s00426-016-0777-y

Nadarevic, L., & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, *23*, 74–84. https://doi.org/10.1016/j.concog.2013.12.002

Nadarevic, L., Meckler, D., & Schmidt, A. (2012). [An investigation of the truth effect and different personality traits]. Unpublished raw data. https://doi.org/10.17605/OSF.IO/6WV4Z

Nadarevic, L., & Rinnewitz, L. (2011). [Judgment mode instructions do not moderate the truth effect]. Unpublished raw data. https://doi.org/10.17605/OSF.IO/3UAJ7

Newman, E. J., Jalbert, M. C., Schwarz, N., & Ly, D. P. (2020). Truthiness, the illusory truth effect, and the role of need for cognition. *Consciousness and Cognition*, *78*, 102866. https://doi.org/10.1016/j.concog.2019.102866

Parks, C. M., & Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition*, *13*(2), 225–253. https://doi.org/10.1080/138255890968691

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

R Core Team. (2019). R: A language and environment for statistical computing. *Vienna, Austria.*

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*(3), 338–342. https://doi.org/10.1006/ccog.1999.0386

Rouder, J. N., & Morey, R. D. (2017). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, *73*, 186–190. https://doi.org/10.1080/00031305.2017.1341334

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 1–12. https://doi.org/10.1525/collabra.28

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stafford, T. (2016). *How liars create the 'illusion of truth'.* BBC.

Thiele, J. E., Haaf, J. M., & Rouder, J. N. (2017). Is there variation across individuals in processing? Bayesian analysis for systems factorial technology. *Journal of Mathematical Psychology, 81*, 40–54. https://doi.org/10.1016/j.jmp.2017.09.002

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science, 28*(3), 247–253. https://doi.org/10.1177/0963721419827854

Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition, 160*, 110–126. https://doi.org/10.1016/j.cognition.2016.12.016

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honour of Bruno de Finetti* (pp. 233–243). Amsterdam, NL, Elsevier.

Appendix A

Description of Data Sets

Set 1

Set 1 contains data from an unpublished study by Nadarevic et al. (2012) on the association of personality traits and the truth effect. In this web-based study, participants were presented with 20 trivia statements, half of which were true and half were false. Participants were informed that statements could be either true or false, and asked to assign each to one of five knowledge categories. In a subsequent phase, participants completed a number of personality questionnaires. Finally, participants were again presented with 20 trivia statements and asked to provide a truth rating on a scale from 1 to 5 for each statement. Ten of these statements had been presented in Phase 1, the other ten statements were new.

In total, 267 participants completed the study. Mean truth ratings for repeated statements were higher than for new statements, $M_{\text{rep}} - M_{\text{new}} = 0.20$ ($SD = 0.34$). A two-sided t test for paired samples revealed a significant truth effect, $t(266) = 9.47$, $p < .001$, Cohen's $d = 0.58$.

Set 2

Data Set 2 is from Nadarevic and Rinnewitz (2011). In the exposure phase, participants were asked to assign 20 unknown trivia statements to one of five knowledge categories. After completing a personality questionnaire in the retention phase, participants were asked to provide truth ratings for 20 statements on a scale from 1 to 5. Ten of these statements were repeated from the exposure phase, the others were new. As an experimental manipulation, one group of participants was instructed to respond intuitively, while the other was asked to think carefully about each truth judgment. This manipulation did not have any systematic effect on truth judgments, however. Therefore, we did not include the factor

condition in the analysis. The total sample comprised 139 participants. On average, truth jdugments for repeated statements were higher than for new statements, $M_{\text{rep}} - M_{\text{new}} = 0.28$ ($SD = 0.36$). This difference is statistically significant, $t(138) = 9.16$, $p < .001$, Cohen's $d = 0.78$.

Set 3

Data Set 3 is based on Experiment 1 in Nadarevic and Aßfalg (2017) who investigated the influence of warnings on the truth effect. For the analysis, we only included data from the control group, which received standard instructions and no warnings. The sample comprises 33 students who participated in the lab-based study. In Phase 1, participants assigned 98 statements (including ten buffer statements) to knowledge categories. After a one-week retention interval, they returned to the lab and provided truth statements on a scale from 1 to 6 for 44 statements from Phase 1 and 44 new statements. Truth statements for repeated items were significantly higher than for new statements, $M_{\text{rep}} - M_{\text{new}} = 0.11$ ($SD = 0.10$), $t(32) = 6.54$, $p < .001$, Cohen's $d = 1.14$.

Set 4

For Set 4, we used data from Nadarevic and Aßfalg's (2017) Experiment 2. It was identical to the first experiment with the exception that the exposure phase was directly followed by the judgment phase and participants rated 80 statements, half of which were repeated. As in the previous data set, we only included the control condition in the analysis. The set contains 98 participants. On the mean level, there is a significant truth effect: $M_{\text{rep}} - M_{\text{new}} = 0.20$ ($SD = 0.23$), $t(97) = 8.61$, $p < .001$, Cohen's $d = 0.87$.

Set 5

Data Set 5 is based on Experiment 1 reported in Nadarevic and Erdfelder (2014). The authors investigated the influence of Phase-1 task and retention interval between exposure and judgment phase on the truth effect. Unlike in the previous experiments, participants

were asked to rate the statements' validity already in Phase 1. In two subsequent truth-judgment phases, participants rated the statements again: One was administered after ten minutes (Phase 2), another one after one week (Phase 3). Half of the statements from Phase 1 were repeated in Phase 2, the other half was repeated in Phase 3. Truth judgments in Phase 1 and a retention interval of several days is a commonly used setting in truth-effect experiments (e.g., Hasher et al., 1977). In contrast, rating the statements' truth in Phase 1 and again after a short retention interval does not lead to a truth effect. Therefore, we only included truth judgments from Phase 3 in the analysis. The sample comprised 85 participants and 88 truth judgments on a scale from 1 to 6 per participant. Analysis of mean truth ratings for repeated and new statements revealed a significant truth effect: $M_{\text{rep}} - M_{\text{new}} = 0.05$ $(SD = 0.10)$, $t(84) = 5.09$, $p < .001$, Cohen's $d = 0.55$.

Set 6

Set 6 contains data from Experiment 2 in Nadarevic and Erdfelder (2014). Participants were assigned to one of two experimental conditions. In one condition, as in Experiment 1, participants rated the truth of each statement in Phase 1. In the other condition, participants assigned the statements to knowledge categories. After a ten minute retention interval, participants in both groups provided truth statements for 88 statements, half of which were repeated from Phase 1. We only included data from the category-rating condition in the analysis. The sample comprised 35 participants. Average truth ratings for repeated statements were significantly higher than for new statements, $M_{\text{rep}} - M_{\text{new}} = 0.07$ $(SD = 0.17)$, $t(34) = 2.50$, $p = .017$, Cohen's $d = 0.42$.

Set 7

Data Set 7 is based on Experiment 1 reported in Brashier et al. (2020). Similar to Nadarevic and Erdfelder's (2014) Experiment 2, participants initially rated 60 statements either for truthfulness or for interest on a scale from 1 to 6. In the judgment phase, they provided truth ratings for these statements and 60 additional statements. We included the

data from the interest-rating condition in the analysis. The data set contains 52 participants. The average difference in truth ratings was $M_{\text{rep}} - M_{\text{new}} = 0.09$ $(SD = 0.12)$, revealing a significant truth effect: $t(51) = 5.52$, $p < .001$, Cohen's $d = 0.77$.

Set 8

For the last data set, we included data from Experiment 1 reported in Pennycook et al. (2018). Participants rated the interestingness of 14 trivia statements in the first phase. After a short retention interval, participants rated the same statements and 14 additional statements in terms of truthfulness on a scale from 1 to 6. Eight of these 28 statements were likely to be known by all participants, the other 20 statements were largely unknown. In the reanalysis, we included truth ratings for the 20 unknown statements. We excluded 13 participants due to missings that occurred during the judgment phase. The final data set comprises 396 participants. On average, truth judgments for repeated statements were higher than for new statements, $M_{\text{rep}} - M_{\text{new}} = 0.14$ $(SD = 0.27)$, revealing a significant truth effect: $t(395) = 10.08$, $p < .001$, Cohen's $d = 0.51$.

Appendix B

Prior Specifications

The parameters α_i, β, μ, and σ^2 are common to all models. We place a noninformative and scale-invariant prior called *Jeffreys prior* (Jeffreys, 1961) on the non-effect parameters μ and σ^2:

$$\pi(\mu, \ \sigma^2) \propto \frac{1}{\sigma^2}. \tag{7}$$

Priors on the remaining parameters are defined as functions of σ^2, the variance of the observed variable. We specify the following prior for person-specific intercepts

$$\alpha_i \overset{\text{iid}}{\sim} \text{Normal}(0, \ g_\alpha \sigma^2),$$
$$g_\alpha \sim \text{Inverse-}\chi^2(r_\alpha^2). \tag{8}$$

We expected the variability in people's individual baselines to be about twice as large as the inter-trial variability. Therefore, we set $r_\alpha^2 = 2^2$. Note that the inverse-χ^2 distribution has fat tails and, given sufficient data, will be overruled by the data if the observed variability differs strongly from what we expected.

We apply a similar reasoning to the effect of factual truth, β. We define

$$\beta \sim \text{Normal}(0, \ g_\beta \sigma^2),$$
$$g_\beta \sim \text{Inverse-}\chi^2(r_\beta^2), \tag{9}$$

and set $r_\beta^2 = 2^2$. Sensitivity analyses revealed that results did not differ meaningfully for any choice of r_α^2, $r_\beta^2 \in [0.25, 16]$.