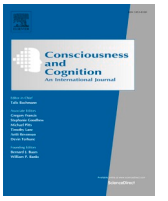
Consciousness and Cognition 135 (2025) 103923

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/yccog



Full Length Article

Assessing the truth effect's reliability and test–retest stability☆

Frank Calio [a,*], Lena Nadarevic [b,c], Jochen Musch [a]

[a] University of Düsseldorf, Germany
[b] University of Mannheim, Germany
[c] Charlotte Fresenius Hochschule, Germany

ARTICLE INFO

Keywords:
Truth effect
Illusory truth
Validity rating
Individual differences
Test–retest stability
Split-half reliability

ABSTRACT

The finding that repeating a statement typically increases its perceived validity is referred to as the truth effect. Research on individual differences in the magnitude of the effect and its correlates is scarce and has yielded rather mixed results. However, any search for replicable relations between the truth effect and other cognitive or personality variables is bound to fail if the truth effect cannot be measured reliably at the individual level and if the effect is not a stable phenomenon. We conducted two experiments investigating the split-half reliability and test–retest stability of the truth effect. To operationalize the magnitude of the effect, Experiment 1 used the between-items criterion and Experiment 2 used the within-items criterion of the truth effect (Dechêne et al., 2010). In both experiments, the truth effect's test–retest stability was found to be very low, probably due to a highly insufficient reliability of the measures that were used. While there may be meaningful and stable individual differences in the truth effect, our findings raise concerns about the usefulness of established indices and standard measures of the truth effect for personality and individual difference research.

1. Introduction

A large body of research accumulated over the past 40 years has shown that repeating a statement typically increases its perceived validity. This phenomenon, which is now referred to as the *truth effect*, was first described by Hasher et al. (1977) and has been observed for a wide variety of stimuli, including true and false trivia statements, opinion statements, and product claims (for recent reviews, see Henderson et al., 2022, and Nadarevic, 2022). The truth effect is typically explained in terms of *processing fluency*, defined as the "metacognitive experience of ease during information processing" (Dechêne et al., 2010 p. 240). According to this account, the truth effect occurs because repeated statements are processed more fluently than novel stimuli and feelings of processing ease are typically used as a cue for validity in judgments of truth (Unkelbach, 2007). Empirical support for the role of processing fluency in truth judgments was provided by Reber and Schwarz (1999), who manipulated processing fluency by varying the color contrast of statements that were presented to a group of participants. Statements which were easier to read and process due to better color contrast were more likely to be judged as true.

In their *referential theory*, Unkelbach and Rom (2017) reconsidered the role of processing fluency in the repetition-based truth effect. The theory assumes that people evaluate the truth of a statement based on the number and the coherence of corresponding

☆ Author note: Frank Calio and Jochen Musch, Department of Experimental Psychology, University of Düsseldorf, Germany; Lena Nadarevic, Department of Psychology, Charlotte Fresenius Hochschule, Wiesbaden, Germany.

* Corresponding author at: Department of Experimental Psychology, University of Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany.
E-mail address: frank.calio@uni-duesseldorf.de (F. Calio).

https://doi.org/10.1016/j.concog.2025.103923
Received 21 April 2025; Received in revised form 1 August 2025; Accepted 15 August 2025
Available online 27 August 2025

references in memory. Specifically, the theory posits that presenting a statement like "Manfred von Richthofen, also known as the 'Red Baron', was a German fighter pilot during World War I" either activates corresponding references (e.g., "Red Baron", "World War I") and their links within a localized network, or prompts the construction of new references and links if the statement contains new information. A statement that has been presented before therefore arguably has more corresponding and coherently linked references than a new statement. This results in more fluent processing of the repeated statement, while also increasing the probability that it will be judged as true.

Although the truth effect has been extensively studied and is considered to be a robust phenomenon that occurs under a wide variety of conditions (Dechêne et al., 2010), there are still important open questions. For instance, there is a lack of research on individual differences in the truth effect and its correlates which has only recently begun to be addressed. In a re-analysis of eight data sets, Schnuerch et al. (2021) found that despite the robustness of the truth effect at the aggregate level, the effect varied notably at the individual level across participants in terms of its magnitude *and* its direction. Furthermore, using a Bayesian modeling approach, Schnuerch et al. (2021) showed that the finding that most people assign higher truth judgments to repeated statements than to novel ones, while others show an opposite pattern, does reflect the existence of true qualitative individual differences in the truth effect rather than sampling noise or measurement error.

While it seems plausible to assume that individual differences in the truth effect are related to certain personality traits or cognitive variables, there is only a limited number of studies that have investigated possible correlates of the truth effect. In two early studies by Arkes et al. (1991) and Boehm (1994), the magnitude of the truth effect was found not to be associated with participants' need for cognition, defined as "an individual's tendency to engage in and enjoy effortful cognitive endeavors" (Cacioppo et al., 1984, p. 306). In contrast to these findings, results by Newman et al. (2020) suggest that people high in need for cognition are more susceptible to the truth effect, but only if they are not warned about the existence of false claims when they first encounter the statements. However, it should be noted that need for cognition did not significantly moderate the truth effect in a second experiment, although participants high in need for cognition again showed a slightly more pronounced truth effect than people scoring low on the Need for Cognition scale. In an advertising context, Sundar et al. (2015) observed a truth effect only for participants scoring high on the Need for Affect scale by Maio and Esses (2001), presumably because consumers who are sensitive to their feelings are more likely to pay attention to and rely on metacognitive cues such as processing ease when judging the truth of a product claim. In another study, DiFonzo et al. (2016) reported a small effect of dispositional skepticism on the magnitude of the truth effect, indicating that participants high in skepticism were less influenced by repetition.

Arguably, the most comprehensive study on the relationship between the truth effect and cognitive as well as personality variables comes from De keersmaecker et al. (2020). In their experiments, the truth effect was not related to participants' cognitive ability or their need for cognitive closure, defined as the "desire for a firm answer to a question and an aversion toward ambiguity" (Kruglanski & Webster, 1996, p. 264). The truth effect was also unaffected by individual differences in the Cognitive Reflection Test, a test that assesses an individual's tendency to engage in "miserly information processing" (Toplak et al., 2014, p. 147). Using the Rational-Experiential Inventory (REI) by Pacini and Epstein (1999) as well as Betsch's (2004) Preference for Intuition and Deliberation scale, De keersmaecker et al. (2020) also investigated the influence of an individual's cognitive style on the truth effect. Whereas there were no significant results with regard to Betsch's scale (2004), De keersmaecker et al. (2020) observed a small positive relationship between experiential thinking as measured by the Rational-Experiential Inventory and the magnitude of the truth effect in one out of four experiments. Furthermore, the REI's measure of rational thinking was significantly associated with the size of the truth effect in two out of four experiments. However, the direction of the latter relationship was inconsistent and meta-analytic techniques summarizing the results across the four relevant experiments showed that neither intuitive thinking nor rational thinking was significantly related to the magnitude of the truth effect. In contrast to the findings by De keersmaecker et al. (2020), a study by Stump et al. (2022) reported a moderating effect of need for cognitive closure on the truth effect, suggesting that the effect is stronger for participants with a high need for cognitive closure. However, it should be noted that in the experiments by Stump et al. (2022), this relationship was only found in one out of two retention interval conditions, i.e., when the exposure phase and the judgment phase were separated by only a few minutes. Preference for intuition and deliberation, which were also assessed in the study by Stump et al. (2022), either failed to show a significant relationship to the truth effect or produced inconsistent results.

Apart from these findings, some researchers have investigated the relationship between the magnitude of the truth effect and several demographic, neuropsychological and psychiatric variables. A meta-analysis by Dechêne et al. (2010), for example, has summarized the results of a small number of studies in which the size of the truth effect was compared between young and old participants (e.g., Law et al., 1998; Mutter et al., 1995). The analysis revealed that the truth effect is not moderated by participants' age. This finding was corroborated in a study by Henderson et al. (2021), who did not find a significant correlation between participants' age and the truth effect either. However, while manipulating the time interval between the initial study phase and the truth judgment phase, Henderson et al. (2021) observed that older participants showed a larger truth effect than younger participants when there was no retention interval between phases, but a slightly smaller truth effect when the study phase and the truth judgment phase were spaced one month apart. To further complicate matters, Brashier et al. (2017) showed that younger adults exhibited a truth effect irrespective of whether they held relevant knowledge about a claim or not, whereas stored knowledge tended to protect older adults from the credibility enhancing effect of repetition. Regarding the relationship between the truth effect and psychiatric variables, Moritz et al. (2012) found that for emotional material, the truth effect was positively correlated with self-reports of positive symptoms typically present in people with schizophrenia. This correlation was observed for both healthy participants and participants with a probable diagnosis of schizophrenia. Ladowsky-Brooks (2010) examined the truth effect in a sample of individuals with traumatic brain injury. In this study, there were no consistent patterns of correlations between scores in several neuropsychological tests (e.g., memory tests, tests of executive functioning) and two different operationalizations of the truth effect's magnitude.

In summary, findings regarding the relationship between the truth effect and cognitive or personality variables have been inconsistent and most significant associations that were discovered have not turned out to be replicable across different experiments or studies. However, a successful search for correlates of the truth effect requires certain preconditions to be met: Replicable relations with cognitive or personality traits can only be found for phenomena that are sufficiently reliable and stable at the individual level. Michalkiewicz and Erdfelder (2016), for instance, who investigated individual differences in the use of the recognition heuristic (Goldstein & Gigerenzer, 2002), argued that before trying to explain variability in the use of the recognition heuristic in terms of personality variables or cognitive traits, it should first be investigated whether the use of the recognition heuristic is stable at the individual level, because no replicable relations with cognitive or personality variables are to be expected if use of the recognition heuristic varies haphazardly within individuals across time and situations. Similarly, no replicable correlation between the magnitude of the truth effect and any cognitive or personality trait is to be expected if the truth effect is not consistent over time or if measures of the truth effect are not sufficiently reliable.

From a theoretical standpoint, stable individual differences in the truth effect are at least conceivable. For instance, as noted earlier, the fluency account suggests that the truth effect arises from the fact that repeated statements are processed more easily, and that the resulting metacognitive feeling of processing ease is used as a cue for truth. As studies have suggested before, it is possible that the magnitude of the truth effect could vary based on characteristics influencing individuals' experience of and response to processing fluency. For instance, individuals may exhibit consistent differences in their sensitivity to fluency or bring different naïve theories to bear when they interpret the metacognitive experience of fluency (cf. Schwarz, 2004). To the extent that such stable individual differences exist, stable individual differences in the magnitude of the truth effect would be expected, provided that the truth effect can be measured with sufficient reliability.

A few studies have already investigated the stability of various constructs in the area of judgment and decision making. Kantner and Lindsay (2012, 2014), for example, found evidence for the temporal and cross-situational stability of response bias in recognition tasks. Michalkiewicz and Erdfelder (2016) demonstrated that the use of the recognition heuristic is stable across time and situations. Other studies (e.g., Beck & Triplett, 2009; Kirby, 2009) have reported evidence for the temporal stability of delay discounting, the effect that an outcome which is remote in time tends to have less value than a more immediate outcome. We sought to extend this research by investigating whether traditional measures of the truth effect are stable at the individual level and correlate across two or more points in time. Because test–retest correlations critically depend on both the temporal stability of a construct as well as the reliability of its measures, we additionally conducted reliability analyses for standard truth effect indices that are based on difference scores. The data from both experiments we conducted are available via the Open Science Framework: https://osf.io/mpaz7/.

2. Experiment 1

The first study was conducted as a computer-based online experiment and consisted of two sessions that were separated by a time interval of approximately one week. Each session consisted of three phases: In an initial exposure phase, participants were asked to assign a set of trivia statements to different categories of knowledge. Participants then completed a distractor task, and finally judged the validity of two sets of statements, one of which had already been presented during the initial category-sorting task. Based on this experimental procedure, it was possible to assess the truth effect by comparing validity ratings between the set of repeated statements and the set of statements that participants had not encountered before. This comparison is known as the *between-items criterion* (Dechêne et al., 2010). Conducting two sessions allowed us to obtain *two* separate indices of the truth effect per participant. The test–retest stability of the truth effect was then calculated as the correlation between these two indices.

2.1. Method

Participants. Participants were recruited by sending email invitations to members of a non-commercial online research panel. The panel consisted of volunteers who had previously expressed an interest in participating in psychological research conducted by the University of Düsseldorf. Out of the 246 participants who started the study, 212 finished the first session. All participants who provided complete data in the first session were contacted again approximately one week after their participation and were invited to take part in the second session. We analyzed only the data of the 169 participants who also finished the second session. Two of these participants had to be excluded because they failed a seriousness check by indicating that they had not participated seriously in at least one of the two experimental sessions (Aust et al., 2013). The data from one additional participant were discarded because this person had completed the first session twice. Therefore, the final sample consisted of 166 participants (90 female, 76 male), all of whom were native speakers of German. Participants' age ranged from 21 to 62 years ($M = 40$, $SD = 12$).

Materials. In order to obtain normative data on the perceived validity of items in a pretest, we collected 404 trivia statements that could be either true or false. The collected statements covered different domains of knowledge, including geography, history, politics and sports. The validity of each statement was evaluated by at least 23 participants on a 6-point rating scale ranging from 1 (*definitely false*) to 6 (*definitely true*). Due to ceiling effects, a truth effect is not to be expected for statements that participants know to be true. Therefore, following common practice in research on the truth effect, we selected 80 statements (40 true, 40 false) that were judged to be ambiguous with regard to their truth status for the main study (e.g., "The actor Keanu Reeves was born in Lebanon"). Validity ratings for these statements varied between $M = 3.04$ and 4.00, with standard deviations ranging from 0.84 to 1.90. We assigned these statements to four stimulus sets A, B, C, and D so that each set contained 10 false and 10 true statements. Special care was taken to match all item sets with regard to their mean perceived validity ($M_A = 3.55$, $M_B = 3.55$, $M_C = 3.55$, $M_D = 3.57$) and their mean standard deviation ($SD_A = 1.36$, $SD_B = 1.37$, $SD_C = 1.36$, $SD_D = 1.36$).

Procedure. Participants took part in two sessions of 20 min each that were separated by approximately one week. Both sessions consisted of an initial category-sorting task, a filler task and a validity-rating task. Within each session, half of the statements in the final validity-rating task had previously been presented in the category-sorting task, whereas the other half of the statements had not been shown before. The truth status of the statements was manipulated orthogonally to their repetition status; half of the statements were true, whereas the other half of the statements were false.

At the beginning of Session 1, participants answered some basic demographic questions. Afterwards, they were informed that they were going to see a collection of statements that could be either true or false, and that it would be their task to assign each statement to one of six different categories of knowledge (geography, flora and fauna, politics and history, science, entertainment, other). Participants were then presented with 32 statements. Twenty statements (10 true, 10 false) from one of the four stimulus sets were presented in random order. These critical statements were preceded by six primacy and followed by six recency buffer items (e.g., "The A in the acronym AIDS stands for the word 'acquired'"), the purpose of which was to make it more difficult for participants to recall the critical items during the validity-rating task at the end of the session. After participants had assigned each statement to one of the six categories of knowledge, a nonverbal filler task followed that lasted for a fixed interval of 10 min. Participants were then introduced to their final task: They were informed that they would again see true and false statements and that this time, all statements should be rated for validity on a 6-point scale ranging from 1 (*definitely false*) to 6 (*definitely true*). Furthermore, participants were told that the statements would be selected randomly from a large collection of statements, and that due to this random sampling they might encounter some statements they were already familiar with. Participants then rated the validity of statements from two sets of items, one of which had already been presented during the initial category-sorting task. The 40 statements were presented in random order. Finally, participants were asked to indicate whether they had participated seriously (Aust et al., 2013).

Invitations to the second session were sent out individually approximately one week after each participant had completed the first part of the study. The distribution of the response interval was rather long-tailed because not every participant responded immediately to this invitation; the median time interval between the two sessions was 7.27 days.

The basic structure of the second session was identical to that of the first session: Participants first performed the category-sorting task for 20 statements (10 true, 10 false) from a new stimulus set. As in Session 1, the statements were presented in random order and were surrounded by six primacy and six recency buffer items. Participants then completed a non-verbal filler task for 10 min. Finally, participants rated the validity of statements from two item sets, one of which had already been presented during the category-sorting task. The 40 statements were presented in random order. Importantly, for each participant, there was no overlap in stimulus sets between Sessions 1 and 2. At the end of the second session, participants again indicated whether they had participated seriously. Finally, they were thanked and debriefed.

Counterbalancing and computation of individual truth effect indices. We counterbalanced all stimulus sets across sessions and judgment tasks according to the plan shown in Table 1. Each participant was randomly assigned to one of the eight counterbalancing conditions. Participants in the first counterbalancing condition, for example, started Session 1 by performing the category-sorting task for the statements in Set A. Later, these participants rated the validity of all statements from Sets A and B, presented in a fully randomized order. Because the between-items criterion of the truth effect is usually calculated as the difference in validity ratings between a set of repeated statements and another set of new statements (Dechêne et al., 2010), individual truth effect indices were then calculated as the difference between the mean validity ratings for the repeated Set A and the non-repeated Set B. In Session 2, participants in the first counterbalancing condition completed the category-sorting task for statements in Set C, and later rated the validity of all statements from Sets C and D, presented in a fully randomized order. For these participants, individual truth effect indices in Session 2 were then calculated as the difference between the mean validity ratings for the repeated Set C and the non-repeated Set D. Truth effect indices in the other counterbalancing conditions were calculated analogously (see Table 1).

Calculating raw differences between mean validity ratings is but one way of operationalizing individual differences in the truth effect. To assess the generalizability of our findings, we also adopted a linear regression approach to calculate residual scores as an indicator of whether participants' validity ratings for repeated statements were higher or lower than would have been expected on the basis of their validity ratings for non-repeated statements.

Table 1
Counterbalancing of Stimulus Sets A, B, C, and D across sessions and judgment tasks in Experiment 1.

		Session 1		Session 2	
Counterbalancing condition	*N*	category sorting	validity rating	category sorting	validity rating
1	22	A	AB	C	CD
2	21	B	BA	D	DC
3	19	A	AB	D	DC
4	18	B	BA	C	CD
5	25	C	CD	A	AB
6	21	D	DC	B	BA
7	25	D	DC	A	AB
8	15	C	CD	B	BA

2.2. Results and discussion

We used R (Version 4.2.3; R Core Team, 2023) for all analyses reported below. Prior to conducting the main analysis, we compared validity ratings across the eight counterbalancing conditions with univariate analyses of variance. Validity ratings did not differ between counterbalancing conditions in either the first or the second session (all $ps > 0.05$). We therefore pooled across all counterbalancing conditions in the following analyses.

To check if a truth effect was present in Session 1, we conducted a repeated measures ANOVA using statements' repetition status (repeated vs. new) and their truth status (true vs. false) as independent variables, and participants' mean validity ratings as the dependent variable. There was only a significant main effect of repetition status, $F(1, 165) = 31.71$, $p < 0.001$, $\eta_p^2 = 0.16$, indicating that in Session 1, participants indeed assigned higher validity ratings to repeated statements ($M = 3.83$, $SD = 0.49$) than to new statements ($M = 3.64$, $SD = 0.42$). To test for the presence of a truth effect in Session 2, we conducted an analogous 2 (repetition status: repeated vs. new) x 2 (truth status: true vs. false) repeated measures ANOVA. Again, there was only a significant main effect of repetition status, $F(1, 165) = 28.30$, $p < 0.001$, $\eta_p^2 = 0.15$. As in Session 1, participants assigned higher validity ratings to repeated statements ($M = 3.78$, $SD = 0.47$) than to new statements ($M = 3.60$, $SD = 0.43$).

As our primary concern was the test–retest stability of the truth effect, we calculated an individual truth effect index for each participant in each session by subtracting the mean validity rating for new statements from the mean validity rating for repeated statements. This computation of raw mean differences is the most straightforward way to calculate a truth effect index based on the between-items criterion (Dechêne et al., 2010). In the first session, 64 % of all participants provided higher validity ratings for repeated statements compared to new statements, as indicated by a truth effect index that was numerically larger than zero. The same pattern was observed in the second session, where 63 % of all participants provided higher validity ratings for repeated statements compared to new statements.[1] However, the main question was whether participants were consistent in their tendency to exhibit a truth effect across both experimental sessions. As can be seen in Fig. 1, the truth effect was not stable at all at an individual level and truth effect indices were virtually uncorrelated across sessions, $r = 0.04$, $p = 0.579$, 95 % CI [-0.11, 0.19]. Thus, the magnitude of a participant's truth effect in Session 1 did not predict the magnitude of the same participant's truth effect in Session 2.

To ensure that our results regarding the test–retest stability of the truth effect were not dependent on a particular operationalization of the effect's magnitude, we also assessed the stability of individual differences in the truth effect based on residual scores calculated from linear regression. We argue that residual scores can be used as an indicator of individual differences in the truth effect because they allow for investigating whether validity ratings for repeated statements differ from what would be expected on the basis of validity ratings for non-repeated statements. To obtain residual scores, we predicted mean validity ratings for repeated statements from mean validity ratings for non-repeated statements in each of the two sessions by means of linear regression. It is important to note that this operationalization of individuals' susceptibility to the truth effect differs conceptually from the computation of raw difference scores: A positive residual indicates that validity ratings for repeated statements are higher than would have been expected on the basis of validity ratings for non-repeated statements. Likewise, a negative residual indicates that validity ratings for repeated statements are lower than would have been expected on the basis of validity ratings for non-repeated statements. To reassess the stability of individual differences in the truth effect, we correlated residual scores across sessions and found test–retest stability to be low, albeit significantly higher than zero, $r = 0.16$, $p = 0.034$, 95 % CI [0.01, 0.31]. Using the formulae provided by Raghunathan et al. (1996) as implemented in the R package cocor (Diedenhofen & Musch, 2015), this correlation was also found to be significantly larger than the corresponding correlation of $r = 0.04$ that was observed for raw difference scores, $z = -3.14$, $p = 0.002$.

In sum, the results of Experiment 1 indicate a surprisingly low test–retest stability of the truth effect. This finding held true irrespective of whether susceptibility to the truth effect was operationalized as a raw difference score between mean validity ratings for repeated and non-repeated statements, or whether it was computed using residual scores. Although the test–retest stability was found to be slightly higher in the latter case, it again remained well below test–retest stabilities reported for other phenomena in the area of judgment and decision making (e.g., Kantner & Lindsay, 2012; Kirby, 2009; Michalkiewicz & Erdfelder, 2016). Given these findings, it would, however, be premature to conclude that the truth effect is not stable at an individual level. While low stability coefficients might indeed imply that a phenomenon is unstable, they can also indicate that the test or the operationalization that was used provides an unreliable measure of the phenomenon (Crocker & Algina, 1986). To test this possibility, we additionally assessed the reliability of the truth effect's measurement using the split-half method.

Over the course of Experiment 1, the truth effect was measured using four mean differences, i.e., $\overline{A}$-$\overline{B}$, $\overline{B}$-$\overline{A}$, $\overline{C}$-$\overline{D}$, and $\overline{D}$-$\overline{C}$. For each of these four measures, we calculated an estimate of the split-half reliability of the truth effect. To this end, we pooled data across sessions and counterbalancing conditions as follows: The computation of the split-half reliability $r_{\overline{A}-\overline{B}}$ was based on validity ratings collected from participants in counterbalancing conditions 1 and 3 during Session 1 and from participants in counterbalancing conditions 5 and 7 during Session 2 (see Table 1). Likewise, $r_{\overline{B}-\overline{A}}$ was calculated using validity ratings from participants in counterbalancing conditions 2 and 4 during Session 1 and from participants in counterbalancing conditions 6 and 8 during Session 2. Validity ratings from counterbalancing conditions 5 and 8 during Session 1 and from counterbalancing conditions 1 and 4 during Session 2 were included in the

[1] That many participants demonstrated a negative truth effect may seem surprising at first glance. However, a study by Schnuerch et al. (2021) showed that the truth effect is inconsistent across people and that it is quite common to find participants showing a negative truth effect by assigning higher validity ratings to new than to repeated statements. In their analyses, Schnuerch et al. (2021) observed negative truth effects in five out of eight data sets. Negative truth effects for a subset of participants were also reported by Henderson et al. (2021) and Mattavelli et al. (2023).

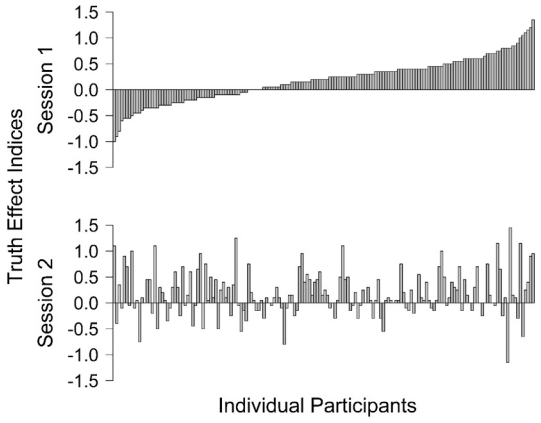


Fig. 1. Magnitude of the truth effect per participant in the first (top) and in the second session (bottom) of Experiment 1. Each bar represents the difference between a participant's mean validity rating for repeated statements and his or her mean validity rating for new statements. Positive values indicate that participants assigned higher validity ratings to repeated statements than to new statements. In both parts of the figure, participants are sorted according to their truth effect indices in Session 1 (in ascending order).

computation of the split-half reliability $r_{\overline{C-D}}$. Finally, $r_{\overline{D-C}}$ was based on validity ratings collected from participants in counterbalancing conditions 6 and 7 during Session 1 and from participants in counterbalancing conditions 2 and 3 during Session 2. To compute each of the four split-half reliabilities, we used the permutated splitting technique provided by the R package splithalfr (Pronk et al., 2022). For each participant, validity ratings were randomly distributed to two halves, with the constraint that both halves contained an equal number of validity ratings for repeated and true, repeated and false, new and true and new and false statements. For each half, a truth effect index was calculated using raw mean differences and the indices from both halves were then correlated across participants. The entire procedure was repeated 5,000 times and the resulting correlation coefficients were averaged across all 5,000 replications. Overall, four separate estimates of the truth effect's split-half reliability were computed. Reliabilities with 95 % confidence intervals (CI) were $r_{\overline{A-B}} = 0.04$, 95 % CI [-0.09, 0.17], $r_{\overline{B-A}} = -0.15$, 95 % CI [-0.34, 0.04], $r_{\overline{C-D}} = 0.11$, 95 % CI [-0.05, 0.27], and $r_{\overline{D-C}} = -0.08$, 95 % CI [-0.24, 0.09].[2]

Taken together, the results of Experiment 1 indicate that both estimates of the test–retest stability and the split-half reliability estimates of the truth effect were close to zero, a result that poses a serious obstacle to finding replicable relations between the truth effect and potentially related cognitive or personality traits. To assess the generalizability of our findings, we conducted a second experiment in a classroom setting which allowed us to use a different operationalization of the truth effect.

3. Experiment 2

In Experiment 1, the presence of a truth effect was tested by comparing validity ratings between a set of repeated statements and another set of new statements. This way of operationalizing the truth effect uses a *between-items criterion* (Dechêne et al., 2010). However, the truth effect can also be assessed by asking participants to evaluate the truth of the same set of statements twice. This alternative operationalization of the truth effect uses a *within-items criterion* (Dechêne et al., 2010). In Experiment 2, we tested whether the low stability and reliability estimates of the truth effect we observed in Experiment 1 would replicate when using the within-items criterion, which is arguably a purer measure of the truth effect because it does not confound the magnitude of the effect with potential a priori differences in the credibility of item sets.

For the truth effect to be assessed based on the within-items criterion, participants need to evaluate the validity of a given set of statements at two different points in time. This way of measuring the truth effect has been used successfully in a large number of previous studies (see Dechêne et al., 2010). However, Nadarevic and Erdfelder (2014) observed no truth effect when they asked participants to judge the validity of the same set of statements twice within a very short time interval of only 10 min. This was probably because participants were either still able to reproduce their original judgments in the second validity-rating phase, or because performing the same validity-rating task twice in close succession induced skepticism and caused participants to discount processing fluency as a cue for validity. In Experiment 2, we therefore ensured that there was a sufficiently large time interval of one week between the validity-rating tasks. Because determining the truth effect's test–retest stability requires not one, but two indices of the truth effect's magnitude, we asked participants to complete the standard validity-rating task on three occasions separated by one week each. Due to this experimental procedure, two individual truth effect indices could be calculated using the within-items criterion: The first index compared validity ratings collected for a set of statements that was first presented in Session 1 and then repeated in Session 2; the second index compared validity ratings collected for a different set of statements that was first presented in Session 2 and then

[2] As the Spearman-Brown adjustment tends to inflate negative correlations (Pronk et al., 2022), we chose to report unadjusted Pearson correlations as reliability estimates instead.

repeated in Session 3. As in Experiment 1, we correlated these two indices to determine the test–retest stability of the truth effect. In addition, we again computed estimates of the split-half reliability to check whether the operationalization of the truth effect that was used in Experiment 2 provides a reliable measurement of the phenomenon.

3.1. Method

Participants. Psychology students at the University of Düsseldorf participated in the experiment for course credit. On three successive weeks, paper-and-pencil test booklets were distributed at the beginning of a lecture on differential psychology. At the beginning of each session, participants were asked to mark their test booklet with a self-generated code, which allowed us to maintain participants' anonymity when matching their booklets across sessions. The number of participants who completed the test booklets in Sessions 1, 2, and 3 was 154, 163, and 149, respectively. One hundred and twenty-one participants attended all three sessions. We discarded the data from three participants who indicated that they had misunderstood the instructions. Data from another two participants also had to be discarded because they were mistakenly given the wrong test booklet in one of the sessions. Our final sample therefore consisted of 116 participants (93 female, 22 male, 1 missing response). To protect participants' anonymity, age was assessed in age categories only. Six percent of the participants were below 18 years of age, 78 % reported being between 18 and 24 years of age, and 12 % reported being between 25 and 30 years of age. Only 3 % of the participants reported being over 30 years old.

Materials. We used the same four stimulus sets (A, B, C, D) as in Experiment 1. Test booklets in each of the three sessions contained two of these sets. In addition to the 2 x 20 = 40 critical statements, we also presented eight unique and easy-to-solve filler statements per session, four of which were true and four of which were false. By including easy filler statements, we tried to maintain participants' motivation by fostering the impression that their knowledge was helpful in judging the factual truth of the presented statements. Thus, the test booklets in each of the three sessions consisted of a total of 40 critical and 8 filler statements, all of which had to be rated for validity. However, validity ratings for the filler items were not included in any of the later analyses.

Procedure. The experiment was announced as a quiz being conducted to select items for a tricky knowledge test. At the beginning of Session 1, all participants were informed that they would have to rate the validity of true and false statements on three successive occasions. We also told participants that they might be uncertain when evaluating the truth of some of the statements, but that they should nevertheless judge the validity of each statement to the best of their knowledge. At the beginning of Sessions 2 and 3, we also told participants that all statements were randomly selected from a large pool of statements, and that due to this random sampling they might encounter some statements they were already familiar with. In each session, participants then rated the validity of 40 critical and 8 filler statements on a 6-point scale ranging from 1 (*definitely false*) to 6 (*definitely true*). Half of the 40 critical statements presented in Session 1 were presented again in Session 2, and 20 new statements first presented in Session 2 were presented again in Session 3, along with 20 additional new statements. In Session 3, participants also provided some basic demographic data and indicated whether they had looked up the truth status of any of the statements they had been presented with in any of the three sessions. Finally, all participants were thanked and debriefed. Each of the three sessions lasted about 12 to 16 min. The time intervals between Sessions 1 and 2 and between Sessions 2 and 3 were 8 and 7 days, respectively.

Counterbalancing and computation of individual truth effect indices. To control for stimulus-specific effects, we randomly assigned participants to one of two counterbalancing conditions (see Table 2), using their month of birth as a randomization device. Because participants were assigned to one of two counterbalancing conditions and took part in three consecutive sessions, six different test booklets were needed. We prepared three versions of each test booklet to control for stimulus-specific order effects. In each version, the filler items were placed at the same fixed positions, whereas the critical statements were arranged in a different random order.

In the first counterbalancing condition, participants started the experiment in Session 1 by rating the validity of statements in Sets A and C. In Session 2, the same participants rated the truth of statements in Sets C and D. Finally, participants ended the experiment in Session 3 by rating the validity of statements in Sets D and B. Thus, a first truth effect index based on the within-items criterion could be computed using the validity ratings for statements in Set C, while a second index could be computed using the validity ratings for statements in Set D. In the second counterbalancing condition, the order of the assignment of Sets C and D was reversed (see Table 2). In this condition, a first truth effect index was computed using the validity ratings for statements in Set D, and a second index was computed using the validity ratings for statements in Set C. Validity ratings for statements in Sets A and B were not used to compute a within-items index, as statements in these sets were presented only once throughout the experiment.

3.2. Results and Discussion

Seven participants admitted that they had read up on the truth status of one or two of the critical statements in between sessions. We therefore discarded the corresponding validity judgments for these participants, resulting in a loss of 17 validity ratings. In a next step, we compared the validity ratings across the two counterbalancing conditions for each of the three sessions using *t*-tests for independent samples. Mean validity ratings did not differ between the counterbalancing conditions in any session (all $ps > 0.05$). Therefore, we pooled across the two counterbalancing conditions in the following analyses.

The truth effect was assessed using the within-items criterion, i.e., by comparing validity ratings for participants' first and second encounter with a given set of statements (Dechêne et al., 2010). Because statements that were presented only once throughout the experiment are not needed for assessing the within-items criterion of the truth effect, we did not include the corresponding validity ratings in the subsequent analyses. Validity ratings for repeated statements were also excluded whenever a given participant failed to judge a particular statement twice, that is, whenever there were missing values on at least one occasion. Across participants, this

Table 2
Counterbalancing of Stimulus Sets A, B, C, and D across sessions in Experiment 2.

Counterbalancing condition	*N*	Session 1	Session 2	Session 3
1	56	AC	CD	DB
2	60	AD	DC	CB

applied to 44 out of the 4.640 (0.95 %) statements that were shown twice over the course of the experiment.

To test whether repetition led to an increase in the perceived validity of statements that had been shown in both Sessions 1 and 2, we conducted a repeated measures ANOVA using experimental session (Session 1 vs. Session 2) and statements' truth status (true vs. false) as independent variables, and participants' mean validity ratings as the dependent variable. As expected, there was a significant main effect of session, $F(1, 115) = 12.50$, $p < 0.001$, $\eta_p^2 = 0.10$, confirming that participants assigned higher validity ratings to statements during their second occurrence in Session 2 ($M = 3.79$, $SD = 0.32$) than during their first occurrence in Session 1 ($M = 3.69$, $SD = 0.35$). The results also showed that participants found it difficult to judge the statements' truth status; they even assigned slightly higher validity ratings to false statements ($M = 3.79$, $SD = 0.35$) than to true statements ($M = 3.69$, $SD = 0.38$), $F(1, 115) = 5.55$, $p = 0.020$, $\eta_p^2 = 0.05$. Importantly, however, the truth effect was not moderated by the factual correctness of statements, since there was no interaction between session and truth status, $F < 1$.

To test whether repetition led to an increase in the perceived validity of statements that had been shown in both Sessions 2 and 3, we conducted another repeated measures ANOVA using experimental session (Session 2 vs. Session 3) and statements' truth status (true vs. false) as independent variables, and participants' mean validity ratings as the dependent variable. Again, there was a significant main effect of session, $F(1, 115) = 51.48$, $p < 0.001$, $\eta_p^2 = 0.31$, confirming that participants assigned higher validity ratings to statements during their second occurrence in Session 3 ($M = 3.79$, $SD = 0.32$) than during their first occurrence in Session 2 ($M = 3.60$, $SD = 0.28$). False statements ($M = 3.73$, $SD = 0.34$) were again rated slightly more credible than true statements ($M = 3.66$, $SD = 0.35$). However, this difference did not reach significance, $F(1, 115) = 3.75$, $p = 0.055$, $\eta_p^2 = 0.03$. More importantly, the truth effect was again not moderated by the factual correctness of statements, since there was no interaction between session and truth status, $F < 1$.[3]

To assess the test–retest stability of the truth effect, we calculated two within-items indices of the effect for each participant based on raw difference scores. The first index represented the mean shift in validity ratings for repeatedly presented statements from Session 1 to Session 2. This index was numerically larger than zero for 57 % of participants, indicating that the majority of participants provided higher validity ratings to statements when they were shown for the second time. The second index represented the mean shift in validity ratings for repeatedly presented statements from Session 2 to Session 3. This index was numerically larger than zero for 69 % of participants. While these findings confirmed the occurrence of a truth effect, our primary concern was whether the effect was stable at the individual level. As can be seen in Fig. 2, this was not the case. The magnitude of the truth effect participants exhibited for statements presented in Sessions 1 and 2 did not predict the magnitude of the truth effect they exhibited for statements presented in Sessions 2 and 3, $r = 0.12$, $p = 0.191$, 95 % CI [-0.06, 0.30].

As in Experiment 1, we supplemented our main analysis by calculating residual scores as an alternative index for individual differences in the truth effect. In a first linear regression, we predicted mean validity ratings for repeated statements in Session 2 based on mean validity ratings for the same statements previously shown in Session 1. Similarly, in a second linear regression, we predicted mean validity ratings for repeated statements in Session 3 based on mean validity ratings for the same statements previously shown in Session 2. We thus obtained two residuals per participant, which we then correlated to estimate the stability of individual differences in the magnitude of the truth effect. Although this correlation was also rather low, $r = 0.27$, $p = 0.003$, 95 % CI [0.09, 0.43], it again turned out to be significantly larger than the corresponding correlation of $r = 0.12$ that was observed for raw difference scores, $z = -2.70$, $p = 0.007$. Thus, this result parallels the findings from Experiment 1.

So far, the results of Experiment 2 indicate that although a truth effect was present at two different points in time at an aggregate level, the test–retest stability of the effect was again rather low. This finding held true for both operationalizations of the truth effect that were used, although the effect's stability was found to be somewhat higher when calculations were based on residual scores. Like in Experiment 1, we complemented our results regarding the test–retest stability of the truth effect by further analyses testing the reliability of the effect's measurement.

Over the course of Experiment 2, the truth effect was measured using two mean differences, i.e., ΔC and ΔD, allowing us to generate two separate estimates of its split-half reliability. For this purpose, we pooled data across sessions and counterbalancing conditions as follows: The computation of the split-half reliability $r_{\Delta C}$ was based on validity ratings for statements in Set C that participants in counterbalancing condition 1 encountered in Sessions 1 and 2 and that participants in counterbalancing condition 2 encountered in Sessions 2 and 3 (see Table 2). Likewise, the split-half reliability $r_{\Delta D}$ was calculated using validity ratings for statements in Set D that participants in counterbalancing condition 1 encountered in Sessions 2 and 3 and that participants in counterbalancing condition 2 encountered in Sessions 1 and 2. Again, we used the permutated splitting technique provided by the R package splithalfr (Pronk et al., 2022) to compute each of the two split-half reliabilities. Before the splitting procedure, the change in validity ratings between a

[3] Apart from assessing the truth effect via the within-items criterion, the design of Experiment 2 also allowed us to assess the effect via the between-items criterion. The results of additional analyses based on the between-items criterion confirmed the findings of our main analyses: Repeated statements were judged to be significantly more credible than new statements both in Session 2 and in Session 3 (all *ps* < 0.001).

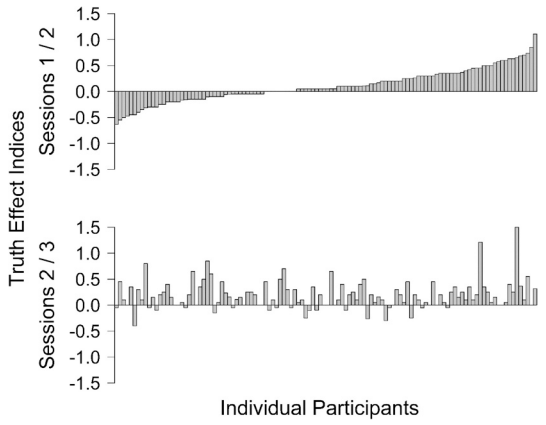


Fig. 2. Participants' individual truth effect indices based on raw mean differences in Experiment 2. Bars represent either the shift in validity ratings for statements that were presented in the first and the second session (top) or the shift in validity ratings for statements that were presented in the second and the third session (bottom). Positive values indicate that participants assigned higher validity ratings to statements during their second presentation. In both parts of the figure, participants are sorted according to the size of their truth effect calculated on the basis of validity ratings from Sessions 1 and 2 (in ascending order).

statement's first and second presentation was computed on an item level for each participant. These difference scores were then randomly distributed to two halves per participant, with the constraint that both halves contained an equal number of difference scores for true and false statements. For each half, a mean difference score was calculated and the indices from both halves were then correlated across participants. The entire procedure was repeated 5,000 times and the resulting correlation coefficients were averaged across all 5,000 replications. The two resulting split-half reliabilities were $r_{\Delta C} = 0.20$, 95 % CI [0.02, 0.38], and $r_{\Delta D} = 0.04$, 95 % CI [-0.09, 0.16].

In sum, the results of Experiment 2 closely replicate the general findings of Experiment 1. Even though the test–retest stability as well as the reliability of the truth effect's measurement appear to be somewhat higher when the truth effect is assessed using the within-items criterion rather than using the between-items criterion, the truth effect's stability and reliability are almost negligible in both cases.

4. General discussion

In the past, several authors have noted a lack of research on individual differences in the magnitude of the truth effect and its correlates (Arkes et al., 1991; Dechêne et al., 2010). This research gap has only recently begun to be addressed and the first studies that have tried to thoroughly investigate the relationship between the truth effect and cognitive as well as personality variables have yielded rather mixed results. In fact, partially inconsistent and non-replicable findings have been observed both across (e.g., De keersmaecker et al., 2020; Stump et al., 2022) as well as within studies (De keersmaecker et al., 2020). In the present study, we therefore investigated the essential preconditions for detecting replicable associations between the magnitude of the truth effect and other cognitive variables or personality traits. Replicable relations with other variables can only be expected if the truth effect can be measured reliably and is temporally stable at an individual level. If these requirements are not met, however, any search for correlates of the truth effect is bound to fail.

In both of our experiments, the reliability of measures of the truth effect was found to be low and test–retest stabilities were found to be much lower than stability coefficients reported for other phenomena in the domain of judgment and decision making (e.g., Kantner & Lindsay, 2012; Kirby, 2009; Michalkiewicz & Erdfelder, 2016). In particular, test–retest stability was found to be very low regardless of whether the truth effect was measured using the between-items criterion (Experiment 1) or the within-items criterion (Experiment 2). Furthermore, although stability was shown to be slightly higher when based on residual rather than difference scores, the general finding of a low test–retest stability was obtained for both operationalizations of the truth effect's magnitude.

Although our results seem to suggest that the truth effect is not an individually stable phenomenon, it is important to note that low stability coefficients can also result from the use of unreliable measures (Crocker & Algina, 1986). The question of how the truth effect can be measured reliably at the individual level is of major importance when investigating the effect from an individual difference perspective. The present study contributes to answering this question by estimating the split-half reliability of raw mean differences between validity ratings for new and repeated statements, a measure that is routinely used to assess the truth effect (e.g., De keersmaecker et al., 2020; Moritz et al., 2012). In both of our experiments, the split-half reliability of these difference scores turned out to be very low. Although these results may seem surprising at first glance, they fit in quite well with other research findings in the domain of judgment and decision making. Buchner and Brandt (2003), for example, showed that the split-half reliability of "illusion-oriented implicit memory measures" (p. 194), such as those derived from fame and preference judgment tasks, can be unsatisfactory. Stated more broadly, this means that measures typically used in the investigation of memory-based illusions, such as the false fame effect (e.g., Jacoby et al., 1989) or the mere exposure effect (e.g., Zajonc, 1968), may suffer from reliability problems. In a related study, Pohl

(1999) assessed the reliability of hindsight bias, a phenomenon that can be defined as "an overestimation of foresight knowledge following the receipt of outcome knowledge" (Harley et al., 2004, p. 962). Reanalyzing several hindsight bias studies, Pohl (1999) found that the reliability of the effect was very low. Taken together, these findings highlight the necessity to empirically establish the reliability of measures in the area of judgment and decision making before embarking on a quest to investigate potential correlates of the effects in question.

When investigating an effect from an individual difference perspective, another important aspect should be kept in mind: Hedge et al. (2018) noted that researchers in experimental psychology often implicitly assume that cognitive tasks producing robust and easily replicable effects will also serve well as objective measures of individual variation. However, this is not necessarily true, as, according to Hedge et al. (2018), cognitive paradigms in experimental research are usually designed and selected for providing robust effects, which is why they are typically accompanied by low between-subjects variability. A low between-subjects variability, however, necessarily reduces reliability and therefore poses a major impediment to finding replicable relations between the magnitude of an effect and potentially related cognitive or personality traits (Hedge et al., 2018). Following this line of reasoning, it is entirely possible for an effect to be robust and unreliable at the same time.

With regard to the influence of repetition on judged truth, our findings suggest that current operationalizations of the truth effect have to be reconsidered if the effect is to be measured reliably at an individual level. Studies that aim to investigate the truth effect from an individual difference perspective will have to address these reliability concerns. It might be possible—albeit costly—to improve the reliability of the truth effect's measurement by using a much larger number of items since—all other things being equal—using a larger number of items increases reliability (e.g., Buchner & Wippich, 2000; see also Nunnally & Bernstein, 1994). Moreover, as difference scores have at times been criticized for being unreliable (for a review, see Zumbo, 1999), future studies should use and evaluate alternative and more advanced operationalizations of the truth effect. Multinomial processing tree (MPT) models, for example, allow to estimate the probability with which specific cognitive processes occur and contribute to observed behavior. MPT models of the truth effect can be used to estimate the probability with which participants rely on processing fluency when asked to judge the truth of a (repeated) statement (Fazio et al., 2015; Unkelbach & Stahl, 2009). While traditional MPT models provide probability estimates for the occurrence of cognitive processes at the group level, recently developed hierarchical models explicitly account for participant heterogeneity by providing parameter estimates for individual participants, making these models particularly suitable for research on individual differences (Heck et al., 2018; Smith & Batchelder, 2010). It might therefore be worthwhile to use hierarchical MPT models if they can help to more reliably assess the truth effect's magnitude at an individual level. However, these models require categorical data rather than the rating-scale data collected in the present experiments.

Regarding alternative operationalizations of the truth effect, it is noteworthy that De keersmaecker et al. (2025) recently published a preprint[4] focusing on the reliability and stability of individual differences in the repetition-based truth effect. Using an experimental setup consisting of five sessions and collecting 108,600 truth judgments, De keersmaecker et al. (2025) operationalized the truth effect as the person-level random effects variance component for the truth effect slope within an ordinal (cumulative probit) mixed effects regression model. The results of their study are seemingly at odds with our own findings, as De keersmaecker et al. (2025) report reliable and stable differences in the repetition-based truth effect over time. While there is no obvious explanation for this discrepancy, it is important to note that their experimental setup differed markedly from other studies that attempted to find correlates of individual differences in the truth effect. Most of these studies have assessed the effect in only one or two experimental sessions, using truth judgments for a substantially smaller pool of repeated and non-repeated statements. In future studies, it may be necessary to use multi-session designs, considerably larger item pools, or more sophisticated ways of assessing the truth effect in order to measure individual differences in the effect and its correlates more reliably (cf. De keersmaecker et al., 2025). Such improved designs might also enable differentiation between individuals with a positive, negative or absent truth effect, which is difficult when the effect itself cannot be reliably measured. Furthermore, when studying the truth effect from an individual difference perspective, it might also be advisable to rely more heavily on the assessment of the effect via the within-items criterion, as this operationalization of the truth effect is based on the comparison of the same set of statements (instead of different sets of statements) and therefore provides a purer measure of the effect by reducing construct-irrelevant variance.

It may then be worthwhile, as a next step, to explore whether other procedural differences might also moderate the reliability of truth effect indices. For example, De keersmaecker et al. (2025) used a relatively short time interval between initial exposure and truth judgments, with only a very brief distraction phase separating the two. Future research might also investigate whether the stability of the truth effect depends on participants holding particular naïve theories relating repetition or fluency to truth.

As a final point, the results of the present study may be subject to criticism in that the sample sizes of 166 and 116 participants in Experiments 1 and 2, respectively, are too small to draw meaningful conclusions about correlations and about the truth effect's test–retest stability in particular. Although our sample sizes are similar to the mean sample size in previous truth effect studies (n = 153, according to a review of 181 studies; Henderson et al., 2022), Schönbrodt and Perugini (2013) have argued that sample sizes should typically approach 250 to achieve stable estimates for correlations. However, while larger sample sizes would have further increased the precision of the reliability estimates reported in the current study, it is important to note that even the upper limits of the confidence intervals for all test–retest stabilities and split-half reliabilities reported in the present work are far below a satisfactory level of reliability or stability. The upper bounds of the 95 % confidence intervals for test–retest stabilities based on difference scores and residual scores were 0.19 and 0.31 in Experiment 1, and 0.30 and 0.43 in Experiment 2, respectively; the upper bounds of the 95 %

[4] We would like to thank Reviewer 1 for bringing this study to our attention.

confidence intervals for split-half reliabilities were 0.17, 0.04, 0.27 and 0.09 in Experiment 1 and 0.38 and 0.16 in Experiment 2. Much higher levels of reliability and stability would be needed to meet reasonable standards of measurement.

While the present results, considered as a whole, cannot rule out the possibility that there are meaningful individual differences in the truth effect and that the effect's test–retest stability might be somewhat higher than the low stabilities that we observed, the findings presented here should be a cause for concern for studies that aim to investigate the effect from an individual difference perspective, because they suggest that the measures typically used to assess the truth effect are unreliable. However, any search for replicable relations between the truth effect and other cognitive or personality variables is bound to fail if the truth effect cannot be measured reliably at an individual level in the first place.

CRediT authorship contribution statement

Frank Calio: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lena Nadarevic:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Jochen Musch:** Writing – review & editing, Methodology, Conceptualization.

Acknowledgments

Funding: This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) to Lena Nadarevic – 451722987.

Data availability

A link to the data is included in the manuscript.

References

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

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*, 527–535. https://doi.org/10.3758/s13428-012-0265-2

Beck, R. C., & Triplett, M. F. (2009). Test-retest reliability of a group-administered paper-pencil measure of delay discounting. *Experimental and Clinical Psychopharmacology, 17*, 345–355. https://doi.org/10.1037/a0017078

Betsch, C. (2004). Präferenz für Intuition und Deliberation (PID): Inventar zur Erfassung von affekt- und kognitionsbasiertem Entscheiden [Preference for intuition and deliberation (PID): An inventory for assessing affect- and cognition-based decision-making]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 25*(4), 179–197. https://doi.org/10.1024/0170-1789.25.4.179

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

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

Buchner, A., & Brandt, M. (2003). Further evidence for systematic reliability differences between explicit and implicit memory tests. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 56*, 193–209. https://doi.org/10.1080/02724980244000260

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology, 40*, 227–259. https://doi.org/10.1006/cogp.1999.0731

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston.

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

De Keersmaecker, J., Wiernik, B. M., Roets, A., & Unkelbach, C. (2025). Truth by repetition reliably differs between people over time. *PsyArXiv*. https://doi.org/10.31234/osf.io/aeukr_v1

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

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10*(4), Article e0121945. https://doi.org/10.1371/journal.pone.0121945

DiFonzo, N., Beckstead, J. W., Stupak, N., & Walders, K. (2016). Validity judgments of rumors heard multiple times: The shape of the truth effect. *Social Influence, 11*, 22–39. https://doi.org/10.1080/15534510.2015.1137224

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General, 144*, 993–1002. https://doi.org/10.1037/xge0000098

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 75–90. https://doi.org/10.1037/0033-295X.109.1.75

Harley, E. M., Carlsen, K. A., & Loftus, G. R. (2004). The "saw-it-all-along" effect: Demonstrations of visual hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 960–968. https://doi.org/10.1037/0278-7393.30.5.960

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

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An r package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50*, 264–284. https://doi.org/10.3758/s13428-017-0869-7

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henderson, E. L., Simons, D. J., & Barr, D. J. (2021). The trajectory of truth: A longitudinal study of the illusory truth effect. *Journal of Cognition, 4*(1), 29. https://doi.org/10.5334/joc.161

Henderson, E. L., Westwood, S. J., & Simons, D. J. (2022). A reproducible systematic map of research on the illusory truth effect. *Psychonomic Bulletin & Review, 29*, 1065–1088. https://doi.org/10.3758/s13423-021-01995-w
Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56*, 326–338. https://doi.org/10.1037/0022-3514.56.3.326
Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40*, 1163–1177. https://doi.org/10.3758/s13421-012-0226-0
Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review, 21*, 1272–1280. https://doi.org/10.3758/s13423-014-0608-3
Kirby, K. N. (2009). One-year temporal stability of delay-discount rates. *Psychonomic Bulletin & Review, 16*, 457–462. https://doi.org/10.3758/PBR.16.3.457
Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review, 103*, 263–283. https://doi.org/10.1037/0033-295X.103.2.263
Ladowsky-Brooks, R. L. (2010). The truth effect in relation to neuropsychological functioning in traumatic brain injury. *Brain Injury, 24*, 1343–1349. https://doi.org/10.3109/02699052.2010.506856
Law, S., Hawkins, S. A., & Craik, F. I. M. (1998). Repetition-induced belief in the elderly: Rehabilitating age-related memory deficits. *Journal of Consumer Research, 25* (2), 91–107. https://doi.org/10.1086/209529
Maio, G. R., & Esses, V. M. (2001). The need for affect: Individual differences in the motivation to approach or avoid emotions. *Journal of Personality, 69*, 583–615. https://doi.org/10.1111/1467-6494.694156
Mattavelli, S., Corneille, O., & Unkelbach, C. (2023). Truth by repetition without repetition: Testing the effect of instructed repetition on truth judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(8), 1264–1279. https://doi.org/10.1037/xlm0001170
Michalkiewicz, M., & Erdfelder, E. (2016). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition, 44*, 454–468. https://doi.org/10.3758/s13421-015-0567-6
Moritz, S., Köther, U., Woodward, T. S., Veckenstedt, R., Dechêne, A., & Stahl, C. (2012). Repetition is good? An internet trial on the illusory truth effect in schizophrenia and nonclinical participants. *Journal of Behavior Therapy and Experimental Psychiatry, 43*, 1058–1063. https://doi.org/10.1016/j.jbtep.2012.04.004
Mutter, S. A., Lindsey, S. E., & Pliske, R. M. (1995). Aging and credibility judgment. *Aging and Cognition, 2*(2), 89–107. https://doi.org/10.1080/13825589508256590
Nadarevic, L. (2022). Illusory truth effect. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (3rd ed, pp. 225–240). London, England: Routledge.
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
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*, Article 102866. https://doi.org/10.1016/j.concog.2019.102866
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.
Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*, 972–987. https://doi.org/10.1037/0022-3514.76.6.972
Pohl, R. F. (1999). *Hindsight bias: Robust, but not reliable*. Giessen, Germany: Justus-Liebig-University. Unpublished manuscript.
Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin and Review, 29*, 44–54. https://doi.org/10.3758/s13423-021-01948-3
R Core Team (2023). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.
Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1*, 178–183. https://doi.org/10.1037/1082-989X.1.2.178
Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*, 338–342. https://doi.org/10.1006/ccog.1999.0386
Schnuerch, M., Nadarevic, L., & Rouder, J. N. (2021). The truth revisited: Bayesian analysis of individual differences in the truth effect. *Psychonomic Bulletin and Review, 28*, 750–765. https://doi.org/10.3758/s13423-020-01814-8
Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609–612. https://doi.org/10.1016/j.jrp.2013.05.009
Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*, 332–348. https://doi.org/10.1207/s15327663jcp1404_2
Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology, 54*, 167–183. https://doi.org/10.1016/j.jmp.2009.06.007
Stump, A., Rummel, J., & Voss, A. (2022). Is it all about the feeling? Affective and (meta-) cognitive mechanisms underlying the truth effect. *Psychological Research Psychologische Forschung, 86*, 12–36. https://doi.org/10.1007/s00426-020-01459-1
Sundar, A., Kardes, F. R., & Wright, S. A. (2015). The influence of repetitive health messages and sensitivity to fluency on the truth effect in advertising. *Journal of Advertising, 44*, 375–387. https://doi.org/10.1080/00913367.2015.1045154
Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning, 20*, 147–168. https://doi.org/10.1080/13546783.2013.844729
Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 219–230. https://doi.org/10.1037/0278-7393.33.1.219
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
Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition, 18*, 22–38. https://doi.org/10.1016/j.concog.2008.09.006
Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2, Part 2), 1–27. https://doi.org/10.1037/h0025848
Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in Social Science Methodology* (pp. 269–304). Greenwich, CT: JAI Press.