Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cphp20

# Speaker trustworthiness: Shall confidence match evidence?

Mélinda Pozzi & Diana Mazzarella

**To cite this article:** Mélinda Pozzi & Diana Mazzarella (2024) Speaker trustworthiness: Shall confidence match evidence?, Philosophical Psychology, 37:1, 102-125, DOI: 10.1080/09515089.2023.2193220

**To link to this article:** https://doi.org/10.1080/09515089.2023.2193220

Published online: 24 Mar 2023.

PHILOSOPHICAL PSYCHOLOGY
2024, VOL. 37, NO. 1, 102–125
https://doi.org/10.1080/09515089.2023.2193220

ARTICLE

OPEN ACCESS Check for updates

# Speaker trustworthiness: Shall confidence match evidence?

Mélinda Pozzi and Diana Mazzarella

Cognitive Science Center, University of Neuchâtel, Neuchâtel, Switzerland

**ABSTRACT**

Overconfidence is typically damaging to one's reputation as a trustworthy source of information. Previous research shows that the reputational cost associated with conveying a piece of false information is higher for confident than unconfident speakers. When judging speaker trustworthiness, individuals do not exclusively rely on past accuracy but consider the extent to which speakers expressed a degree of confidence that matched the accuracy of their claims (their "confidence-accuracy calibration"). The present study experimentally examines the interplay between confidence, accuracy and a third factor, namely evidence, in the assessment of speaker trustworthiness. Experiment 1 probes the hypothesis that overconfidence does not backfire when a confident but inaccurate claim is justified: the trustworthiness of a confident speaker who turns out to be wrong is restored if the confidence expressed is based on strong evidence (good confidence-evidence calibration). Experiment 2 investigates the hypothesis that confidence can backfire if a confident and accurate claim is not justified: the trustworthiness of a confident speaker who turns out to be right is damaged if the confidence expressed is based on weak evidence (bad confidence-evidence calibration). Our results support both hypotheses and thus suggest that "confidence-evidence calibration" plays a crucial role in the assessment of speaker trustworthiness.

**ARTICLE HISTORY**
Received 1 November 2022
Accepted 14 March 2023

**KEYWORDS**
Commitment; reputation; epistemic trust; calibration

## 1. Introduction

Imagine that you are a member of a jury in court, and you are in charge of a car accident case. You read the deposition of the eyewitness, in which he identifies *with confidence* the passenger of one of the cars (suspected of shoplifting). Consider now the following two scenarios. The first scenario is one in which the witness identifies the wrong passenger, but is justified in

**CONTACT** Mélinda Pozzi pozzimelinda@gmail.com Cognitive Science Center, University of Neuchâtel, Pierre-à-Mazel 7, Neuchâtel CH-2000; Diana Mazzarella diana.mazzarella@unine.ch Cognitive Science Center, University of Neuchâtel, NeuchâtelSwitzerland

Supplemental data for this article can be accessed online at https://doi.org/10.1080/09515089.2023.2193220

doing so. It turns out that the passenger has an identical twin, who is also friends with the driver and whom the witness mistakenly took to be on the accident scene. In the second scenario, the witness identifies the right passenger. However, it turns out that the witness has just made a lucky guess as a CCTV camera footage shows that, from his perspective, the passenger had his back turned the whole time and his face was therefore not visible to the witness.

How would you judge the witness's trustworthiness in these two scenarios? Certainly, while in the first one, the witness was confident but wrong, in the second one, he was confident and right. But would your judgment be influenced by whether the witness was more or less justified in being confident? In the first scenario, the existence of the identical twin provided the witness with good evidence to identify the passenger with high confidence (although he turned out to be wrong). The same confidence, though, appears to be inadequately justified in the second scenario, in which the witness indeed identifies the right passenger, but has weak evidence to support his confidence.

The present study investigates the interplay between the confidence expressed, the accuracy of the message and its evidential basis in the assessment of speaker trustworthiness. We will proceed as follows. First, we draw on the literature in cognitive and evolutionary psychology to introduce the notion of "epistemic vigilance" and discuss which cognitive capacities allow humans to assess the trustworthiness of a source of information. Among these capacities, we focus on humans' ability to track and evaluate speaker commitments (section 1.1). We consider why the speaker's *expressed confidence* can function as a commitment signal, thus modulating the speaker's accountability for the truth/falsity of the message communicated (section 1.2). We then move to discuss the role of *evidence* in the assessment of speaker trustworthiness, and outline some possible implications for its interplay with confidence and accuracy (section 1.3). We present our experimental study on "confidence-evidence calibration" and its impact on speaker perceived trustworthiness (sections 2 and 3). We end by discussing the results in the context of the philosophical work on trust and commitment (section 4).

### *1.1. Epistemic vigilance and speaker trustworthiness*

Trust is essential for human interactions and communication. Trust allows humans to cooperate and achieve goals that they would not be able to achieve by themselves. Furthermore, it enables humans to share and acquire information that would be otherwise difficult or impossible to obtain. Because of the risk of defection and the risk of misinformation, humans have evolved a suite of cognitive mechanisms that allows them to assess others' trustworthiness and guide partner choice. Specifically, to benefit from communication while avoiding the risks of misinformation, humans are endowed with a capacity for "epistemic vigilance" (Sperber et al., 2010), which enables them to calibrate their trust. Which cognitive mechanisms underpin epistemic vigilance toward the source of the information?

Research in psychology has identified a variety of selective trust mechanisms, which develop early in ontogeny and operate upon different cues of speaker trustworthiness: speaker's past accuracy (Chow et al., 2008; Koenig et al., 2004), perceived competence and confidence (Sabbagh & Baldwin, 2001), expertise (Koenig & Jaswal, 2011), group membership (Elashi & Mills, 2014), perceived benevolence (Mascaro & Sperber, 2009), to mention just a few. Speakers that display such cues are typically considered more trustworthy than speakers that do not (who are perceived as inaccurate, incompetent, inexpert, out-group, malevolent, etc.). As a result, addressees are more likely to accept a message as true and more disposed to learn from the former than from the latter (for an overview, see Harris, 2012; Poulin Dubois & Brosseau-Liard, 2016; Robinson & Einav, 2014).

Another important component of human epistemic vigilance is related to the ability to track and evaluate speaker "epistemic commitments". When transmitting information, speakers typically undertake some commitment toward the truth of their claims and addressees expect them to respect this commitment. This expectation was captured by Grice (1975) in his conversational Maxim of Quality "Try to make your contribution one that is true", which expresses the assumption that rational and cooperative interlocutors provide truthful contributions to the conversation. However, communicators can explicitly endorse or distance themselves from the truth of their claims through different linguistic devices, thus modulating their epistemic commitments. For instance, Boulat and Maillat (2017) suggest that linguistic markers of epistemic modality and evidentiality can function as commitment signals. As an example, the epistemic modal "must" would signal a stronger epistemic commitment than "might" (compare "The key must be in the closet" with "The key might be in the closet", see also Pietrandrea, 2008). Furthermore, research in the philosophy of language and linguistic pragmatics has suggested that commitment can be pragmatically modulated. For instance, while asserting is a way of taking commitment toward the truth of the message, implicating or insinuating may reduce speaker commitment, and leave it open to the speaker the possibility to deny having had the intention to communicate a given content if this turns out to be false (see, for instance, Fricker, 2012; Mazzarella et al., 2018; Pinker et al., 2008).[1]

Why is it important for humans to track and evaluate epistemic commitment? Commitment is an invitation to trust: committed speakers invite the addressee to trust them and accept their claims as true (Vullioud et al., 2017). Crucially, though, by doing this, speakers put their reputation at risk, where "reputation" is to be understood as the "track record as an informant" (Vallinder & Olsson, 2014). Indeed, committed speakers suffer higher reputational costs than non-committed speakers if their message turns out to be inaccurate (Mazzarella et al., 2018; Vullioud et al., 2017). It is this reputational risk that, according to Vullioud et al. (2017), allows expressing and tracking commitments to be advantageous, on average, for both speakers and addressees. On the one hand, speakers who take higher responsibility by displaying commitment (being ready to justify their message if necessary and to pay the consequences if the information is found to be inaccurate) should have good reasons to incur this risk (e.g., they probably have strong evidence for what they say), and addressees should thus be more likely to accept the message, making commitment beneficial to speakers. On the other hand, given the risk of overcommitment, committed speakers are more likely to share true information, making commitment beneficial to addresses. There is thus a trade-off for speakers between the benefits that come with the expression of stronger commitments and the potential costs for reputation as a trustworthy source of information. This balance between risks and benefits is thought to have made commitment a crucial evolutionary aspect in the stabilization of human communication (Vullioud et al., 2017) since it benefits both speakers (higher acceptance of their messages) and the addressees (lower risk of misinformation).

Crucially, it is worth noticing that research in philosophy and epistemology has put a great emphasis on the role of commitment for speaker trustworthiness. For instance, Hawley (2014, 2019) defines trustworthiness as the avoidance of unfulfilled commitments. Trusting someone is to believe that this person has a commitment, and to rely on this person to fulfil that commitment. In her view, being trustworthy requires fulfilling one's commitments, but also taking commitments that one can fulfil. Applied to testimony, epistemic trustworthiness would involve fulfilling one's commitment to say the truth, and undertaking epistemic commitments that are warranted (for instance, by committing to the truth of a message for which one has a good evidential basis). Building on this perspective, commitment and evidence appear to be strongly intertwined. In what follows, we will focus on one way of expressing commitment, namely confidence (section 1.2), and discuss its interplay with evidence (section 1.3).

### *1.2. Confidence as a commitment signal*

A common linguistic device to adjust commitment is the verbal expression of confidence and doubt used to indicate the speaker's strength of feeling or belief in the truth of the information communicated (Clark, 1990). Expressing confidence is a means for taking commitment while expressing unconfidence is a way of avoiding commitment. One of the most frequent ways for speakers to express confidence is the use of certainty expressions, which fall on a continuum of cases from high-confidence phrases to low-confidence phrases (with words such as *positive*, *certain*, *sure*, *think*, and *suppose* being consistently rated in this ordering; see Wesson & Pulford, 2009). Crucially, the use of these phrases may have a direct impact on the force of the statement: a speaker who employs phrases such as "I suppose that" or "I guess that" is not making an assertion and therefore avoids taking responsibility for the truth of the message (Marsili, 2018). Furthermore, speakers can also display confidence with other linguistic cues such as speech rate, intonation, volume, and non-linguistic cues such as gestures, facial expressions, and posture. Non-linguistic cues are typically vaguer than linguistic ones and are thus perceived as less committal (Tenney et al., 2019).

There is plenty of evidence in the psychological literature supporting the claim that confidence functions as a commitment signal. First, as committing is beneficial for speakers, the same benefits should be found when speakers express confidence. Indeed, confident speakers are more likely to be believed (Tenney et al., 2007, 2008, 2019; Vullioud et al., 2017), and speakers' confidence increases addressees' learning (Birch et al., 2020). People favor speakers who make assertions to those marking their claims with hedging expressions such as "as far as I know" (C. Moore et al., 1989; Sabbagh & Baldwin, 2001; Tenney et al., 2007, 2008). This preferential bias for confident speakers is so pervasive that it has received the name of "confidence heuristic" in the psychological literature (e.g., Birch et al., 2020; Kominsky et al., 2016). Second, as committing puts the speaker's reputation at stake, overconfidence (i.e., being confident when the information transmitted is inaccurate) should lead to reputational loss. In line with this, several studies have shown that overconfidence is damaging to the informant's reputation as a trustworthy source of information: confident speakers who transmit inaccurate information suffer higher costs (punishment and loss of credibility) than unconfident speakers (Mazzarella et al., 2018; Tenney et al., 2007, 2008, 2019). Crucially, these costs are higher than the costs incurred by inaccurate speakers whom one has trusted based on factors other than the expression of commitment (for instance, based on their previous accuracy; see Vullioud et al., 2017).

Given that confidence functions as a commitment signal, vigilant interlocutors should track and evaluate speaker's past calibration between their expression of confidence and the accuracy of their claims ("confidence-accuracy calibration", see Tenney et al., 2007, 2008, 2019). Humans are well-calibrated when their expression of confidence corresponds to the accuracy of the information (i.e., confident when the information is accurate, unconfident when the information is inaccurate), and they may pay for a bad confidence-accuracy calibration (i.e., confident when the information is inaccurate, unconfident when the information is accurate). Addresses expect speakers to show good confidence-accuracy calibration ("presumption of calibration"), and they tend to distrust bad-calibrated communicators when they have access to this calibration cue (Tenney et al., 2007, 2008, 2019; Vullioud et al., 2017).

To conclude this discussion on confidence as a commitment signal, let us go back to Hawley's notion of trustworthiness. As suggested earlier, according to Hawley (2019), being trustworthy requires fulfilling one's commitments, as well as taking commitments that one can fulfil. In light of this, we suggest that confidence-accuracy calibration matters for this first requirement: expressing confidence is a way of acquiring commitment toward the truth of one's claim, and being accurate is the fulfillment of that commitment. In the next section, we investigate the second requirement: taking commitments that one can fulfil. Specifically, we raise the question of whether the perceived trustworthiness of confident speakers should be affected by the quality of the evidence available to them. What if a bad-calibrated overconfident speaker is justified by strong evidence? And what about a well-calibrated confident speaker who is right but not justified by enough evidence? We explore the importance of evidence when evaluating speaker trustworthiness, and outline why the confidence expressed (which signals the speaker's epistemic commitment) should be justified by the speaker's evidence.

### *1.3. Adjusting commitment to evidence: confidence-evidence calibration*

People typically expect informants to provide information that is not only true but also supported by adequate evidence. This expectation is captured by the second Gricean sub-maxim of Quality: "Do not say that for which you lack adequate evidence" (Grice, 1975). Cooperative speakers generally assert information for which they have a good evidential basis, and addresses expect them to have evidence for what they assert. To probe this intuition, Kneer (2018) experimentally investigated under which conditions an assertion is judged as acceptable. He tested four different norms of assertions that have been proposed in the literature: the "truth" norm saying that one should assert that *p* only if *p* is true, the "knowledge" norm saying that one should assert that *p* only if one knows that *p* (i.e., *p* is true and justified), the "belief" norm saying that one should assert that *p* only if one believes that *p*, and the "justified belief" norm saying that one should assert that *p* only if one believes that *p* and one has justification for it. According to the two last norms, it is acceptable to make an assertion even if it is false. In a series of experiments, Kneer (2018) found that in a situation in which a speaker asserts an accidentally true belief that is justified, participants considered *p* as not known but still assertable (against the knowledge account), while in a situation in which a speaker asserts a false belief that is justified, participants considered *p* as not known and not true but still assertable (against both the knowledge account and the truth account). However, in both situations, participants considered *p* as justified and assertable, supporting the justified belief account (*p* is assertable when it is believed and justified). To control that the findings were indeed the results of a justified belief norm and not simply a belief norm, Kneer (2018) compared a situation in which the speaker asserts that *p* and has good evidence for it with another situation in which the speaker has poor evidence. The results show that in the first situation, participants considered *p* as believed, justified and assertable (which supports both the justified belief account and the belief account), but in the second situation participants considered *p* as believed, not justified and not assertable (contradicting the belief account and supporting the justified belief account). Kneer (2018) concluded that the justified belief norm is the only norm of assertion that could explain all these findings. Justification is not only sufficient (*p* is assertable when it is justified, even if it is not true), but it is also a necessary condition for an assertion (*p* is not assertable when it is not justified, even if it is true).[2] As a result, it may be argued that, because speakers are expected to assert what they have adequate evidence for, speaker trustworthiness may be affected by whether speakers satisfy this expectation, over and beyond the actual truth or accidental falsity of their claim.

What are the implications for the relationship between confidence, accuracy and evidence? One may argue that a confident speaker who conveys a piece of information which turns out to be wrong may be less likely to suffer a reputation cost if the confidence expressed was justified by strong evidence. If the falsity of the claim could not have been foreseen and appears highly unlikely based on the available evidence, confidence would be justified and should not be punished. Preliminary evidence from experimental psychology suggests that this is indeed the case. Tenney et al. (2008) study shows that the reputation of a confident speaker that is found to be inaccurate (and thus suffers a reputational cost) is restored when the mistake turns out to be justified. This indicates that speaker trustworthiness is affected by confidence-evidence calibration, more than confidence-accuracy calibration. The first objective of our experimental study is to corroborate the hypothesis that overconfidence does not backfire when a confident but inaccurate claim is justified and thus to replicate Tenney et al. (2008) preliminary findings in this direction (see Experiment 1).

Crucially, though, to examine whether confidence-evidence calibration trumps confidence-accuracy calibration in the assessment of speaker trustworthiness, one also needs to investigate cases in which confident speakers turn out to be accurate, although their confidence was not supported by adequate evidence. In these cases, speakers are accurate only *by chance*, and the invitation to trust carried by their expression of confidence is thus fundamentally unwarranted. Previous literature in epistemology has long discussed cases of “epistemic luck”, that is, cases in which a belief turns out to be true because of mere luck, and vigorously debated the extent to which a proper account of knowledge should exclude them (starting from the seminal work of Gettier, 1963; see Engel, 2011 for a review). Similar concerns can arise for cases of *testimonial luck*, that is, situations in which a testimony turns out to be accurate because of mere luck. The second objective of our experimental study is thus to investigate the implications of testimonial luck for the assessment of speaker trustworthiness (Experiment 2). More specifically, we aim at testing the hypothesis that confidence can backfire if a confident and accurate claim is not justified. To our knowledge, no previous experimental work has addressed this issue.

There are at least two different reasons why a confident speaker who is accurate only by chance should suffer a reputational loss. The first one involves counterfactual considerations: while the claim is accurate in the actual world, given the available evidence, the same claim could have easily turned out to be wrong, if things went slightly differently (that is, in nearly all nearby possible worlds in which the confident speaker had the same evidence). For this reason, confidence appears to be misplaced, and speaker trustworthiness should be revised downwards (see Pritchard, 2003 for an analogous modal characterization of epistemic luck). The second reason concerns the predictability of future interaction. If the confident claim turned out to be accurate only by chance, there is no guarantee that the speaker will provide valuable (that is, epistemically warranted) information in the future. As a result, although the speaker is accurate today, there is no reason to trust them in the future.

Overall, we suggest that speakers may be held accountable for expressing a level of confidence (and therefore commitment) that matches the quality of the evidence available to them. When transmitting information, speakers may be more or less confident about its accuracy depending on how they acquired this information. As a result, they would be well calibrated when their expression of confidence corresponded to the strength of the evidence (i.e., confident when they have strong evidence, unconfident when they have weak evidence), and they may pay a reputational cost for a bad confidence-evidence calibration (i.e., confident when they have weak evidence, unconfident when they have strong evidence), independently of the accuracy of the information. The present study investigates whether confidence-evidence calibration matters more than confidence-accuracy calibration when judging speaker trustworthiness.

## 2. The present study

The present study has two aims. First, to replicate the findings of Tenney et al. (2008) in support of the hypothesis that overconfidence does not backfire when a confident but inaccurate claim is justified: the trustworthiness of a confident speaker who turns out to be wrong is restored if the confidence expressed is based on strong evidence (Hypothesis 1 – Experiment 1). Crucially, this replication would allow to establish the same findings across different experimental settings (from Tenney et al.'s pen and pencil experiment with participants in the lab to our web-based experiment with online participants) and thus support the viability of our methodological approach. Second, the study aimed at investigating whether confidence can backfire if a confident and accurate claim is not justified: the trustworthiness of a confident speaker who turns out to be right is damaged if the confidence expressed is based on weak evidence (Hypothesis 2 – Experiment 2). The pre-registration of the study can be found on OSF with the following link: https://osf.io/fbv8g/?view_only=d3e7a5c5c7ef49439d5dbaabea5db912. The study was approved by the Ethics Committee of the University of Neuchâtel.

To test our two hypotheses, we conducted two online experiments in which participants were presented with two testimonies concerning a car accident, one from a confident witness and the other from an unconfident witness (both identified as males). In Experiment 1, both witnesses were inaccurate but were justified by strong evidence. In Experiment 2, both witnesses were accurate but had weak evidence. So, in both experiments, the accuracy and the strength of evidence were kept stable between the two witnesses, and the level of confidence differed between the two witnesses. The material was adapted from Tenney et al. (2008, Experiment 2), and we used the same type of evidence employed in their study, that is, perceptual evidence. Participants judged the credibility of the two witnesses (on a scale from 1 to 6) and were asked to choose which of the two depositions they believed at different times during the experiment, as information about accuracy and evidence unfolds. Participants were asked to justify their choices (as an attention check). We, therefore, had two measures of speaker trustworthiness: credibility (continuous measure) and believability (binary measure). Both trustworthiness measures were also taken from Tenney et al. (2008). We measured trustworthiness at three distinct times: (1) participants have no information about the accuracy and strength of evidence, (2) participants get feedback about accuracy, and (3) participants get feedback about evidence. This allowed us to investigate whether the confident witness would regain their trustworthiness when their inaccuracy was found to be justified by strong evidence (Experiment 1) and whether they would lose their trustworthiness when their accuracy was discovered to be not sufficiently evidenced (Experiment 2). Based on the expectation that we would be able to replicate Tenney et al.'s findings, the two experiments were run at the same time with participants being randomly assigned to one of the two experiments. For the sake of exposition, we present the two experiments one after the other.

## 2.1. Experiment 1

Experiment 1 tested Hypothesis 1: the trustworthiness of a confident speaker who turns out to be wrong is restored if the confidence expressed is based on strong evidence.

### 2.1.1. Method

*Participants.* The sample size was determined based on a power analysis. Since there are no agreed-upon sample size calculations for mixed models, we made our sample size calculation based on the statistical tests used by Tenney et al. (2008). The largest resulting sample size was selected, i.e., 107 participants. The sample size was based on having 80% power to detect an effect size $w = 0.3$ (medium effect size) with $df = 2$ for a chi-square. This sample was approximately the same as the original study (i.e., 105 participants in Tenney et al., 2008).

The experiment was implemented on Qualtrics, and participants were recruited via Prolific. Only native adult English speakers could take part in the experiment (pre-screening on Prolific), and they were paid 0.88£/1.12$ (the experiment took about 7 minutes). The final sample size was 108 participants (30 men, 78 women, $M_{\text{age}} = 37.19$, $SD = 12.73$). We did not exclude any participants (all participants completed the entire experiment and all the participants gave relevant justifications for their answers).

*Materials and design.* All the vignettes are given in the Supplementary material.

At Time 1, participants read two witnesses' depositions describing a car accident. The witnesses disagreed on which car was at fault in the accident and identified the passenger in one of the vehicles. One witness was confident about all aspects of his testimony, i.e., the accident, the weather that day, and the identification of the passenger. The other witness was confident about the accident and the weather, but not about the identification of the passenger. The confident witness was therefore committed to the truth of the passenger identification, while the unconfident witness was not. Which informant was confident/unconfident (i.e., Witness 1 or Witness 2) was counterbalanced. Both witnesses identified the same person as the passenger (a man sitting in the third row of the courtroom). After reading the two testimonies, participants rated the credibility of each witness on a scale from 1 (not credible) to 6 (credible). They also made a binary decision as to which witness's deposition they believed and justified this choice.

At Time 2, participants got feedback about the accuracy of the weather report and the passenger identification (but not about the rest of the deposition). Participants learned that both witnesses were accurate about the weather on the day of the accident but inaccurate about the identification of the passenger. Participants rated again the credibility of each witness as well as made a binary decision as to which witness's deposition they believed, and justified this choice.

At Time 3, participants got feedback about the strength of the evidence that both witnesses had to identify the passenger. Participants read that the two witnesses had strong evidence (they were justified in being inaccurate): they were told that the passenger had an identical twin who was also friends with the driver. The evidence on the basis on which the two witnesses had formulated their claims was therefore perceptual and strong. Participants rated again the credibility of each witness as well as made a binary decision as to which witness's deposition they believed, and justified this choice.

We expected that, at Time 1, in the absence of any information about accuracy and strength of evidence, the confident witness (i.e., confident about the accident, the weather, and the identification of the passenger) would be judged more trustworthy (i.e., he would be rated as more credible and would be more likely to be believed) than the unconfident witness (i.e., confident about the accident and the weather, but not about the identification of the passenger). At Time 2, when both witnesses turn out to be inaccurate (i.e., they identified the wrong passenger), the confident witness would lose his trustworthiness (both for credibility and believability) to the benefit of the unconfident witness. At Time 3, when the inaccuracy is found to be justified by strong evidence (i.e., the passenger had an identical twin who was also friends with the driver), the confident informant's trustworthiness would be restored (both for credibility and believability).

### 2.1.2. Results

All the analyses were performed in R (v. 4.2.1; R Core Team, 2022) using R Studio (v. 2.4.1; RStudio Team, 2022). We fitted a cumulative linear mixed model (CLMM) for credibility ratings which included confidence and time as fixed effects, and participant as a random effect. Furthermore, we fitted a generalized linear mixed model (GLMM) for believability which included time as a fixed effect, and participant as a random effect. Data and R scripts are available on OSF (https://osf.io/z92tg/?view_only=7402859320fd43dea4b652e1ef4bfbb5). Results show a significant interaction effect of confidence and time on credibility ($\chi^2$ (2) = 106.896, $p < .001$, see Figure 1). Believability changed over time as predicted ($\chi^2(2) = 59.229$, $p < .001$, see Figure 2). For the post-hoc comparisons,[3] we used paired Wilcoxon signed-rank tests for credibility ratings, and binomial tests for believability. In the absence of any information about accuracy and strength of evidence (Time 1), the confident witness was judged more trustworthy: he was rated as more credible ($M = 4.38$, $M_d = 4$, $SD = 1.05$) than the unconfident witness ($M = 3.79$, $M_d = 4$, $SD = 1.10$, $Z = 1851$, $p < .001$, $r = .53$), and was more likely to be believed

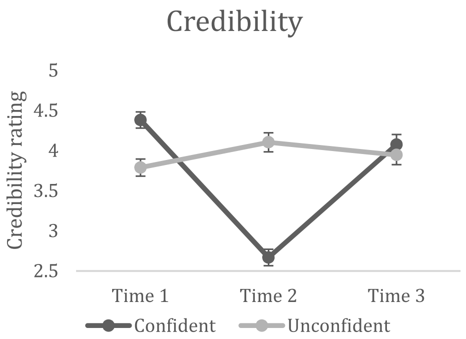

Figure 1. Credibility scores of the confident (black) and unconfident (gray) witnesses, on a scale from 1 "not credible" to 6 "credible".

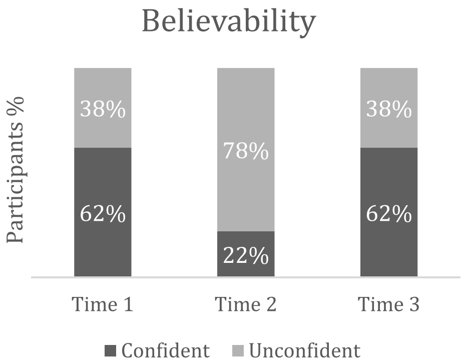

Figure 2. Percentage of participants who believed the deposition of the confident (black) or unconfident (gray) witness.

than the unconfident witness (63% confident witness, 68/108, binomial $p = .009$, Cohen's $g = .13$). At Time 2, when both witnesses turned out to be inaccurate, the confident witness lost his trustworthiness to the benefit of the unconfident witness: he was rated as less credible ($M = 2.67$, $M_d = 3$, $SD = 1.07$) than the unconfident witness ($M = 4.10$, $M_d = 4$, $SD = 1.24$, $Z = 273$, $p < .001$, $r = -.85$), and was less likely to be believed than the unconfident witness (22% confident witness, 24/108, binomial $p < .001$, Cohen's $g = -.28$). At Time 3, when the inaccuracy was found to be justified by strong evidence, the confident witness' trustworthiness was (partially) restored: the confident witness was rated as credible ($M = 4.07$, $M_d = 4$, $SD = 1.30$) as the unconfident witness ($M = 3.94$, $M_d = 4$, $SD = 1.26$, $Z = 1342.5$, $p = .552$, $r = .08$), and was more likely to be believed than the unconfident witness (63% confident witness, 68/108, binomial $p = .009$, Cohen's $g = .13$).

### 2.1.3. Discussion

The results of Experiment 1 supported our first hypothesis: Trustworthiness is (at least partially) recovered when a confident and inaccurate claim is justified by strong evidence. At Time 1, when confidence was the only factor participants could rely on, the confident witness was rated as more credible and was more likely to be believed. This result is in line with the idea that, in the absence of other cues of trustworthiness, individuals may adopt a confidence heuristic, as well as with previous findings showing that confidence acts as a commitment signal, and thus increases the likelihood that the message will be accepted as true. At Time 2, when the confident witness turned out to be overconfident (confident and inaccurate), he lost his trustworthiness to the benefit of the unconfident witness. This finding supports the role of confidence-accuracy calibration in the assessment of speaker trustworthiness: participants judged a bad-calibrated (confident and inaccurate) speaker less trustworthy than a well-calibrated (unconfident and inaccurate) speaker.

At Time 3, when the overconfident witness (confident and inaccurate) turned out to be justified (strong evidence), his trustworthiness was restored (although not completely). This result supports the claim that a good confidence-evidence calibration can counteract the negative effects of a bad confidence-accuracy calibration: participants partially restored the reputational loss of the speaker displaying good confidence-evidence calibration (confident and strong evidence), even if the speaker was overconfident (confident and inaccurate).

Overall, these results are in line with the findings of Tenney et al. (2008). Interestingly, though, our results reveal that the overconfident witness still incurred some cost (although he was justified), as his credibility rating was slightly (but significantly) lower at Time 3 compared to Time 1. There was no evidence of such a cost in Tenney et al. (2008), where the confident witness's credibility ratings at Time 3 were comparable to those at Time 1, and they were higher than the credibility ratings for the unconfident witness at Time 3. In contrast to this, our data suggest that failing to fulfil a commitment (i.e., failing to convey a true message when the claim is made with confidence) can still have an impact on speaker trustworthiness, even if the speaker is perceived to be justified to take this commitment (i.e., by strong evidence).

## 2.2. Experiment 2

Experiment 2 tested Hypothesis 2: the trustworthiness of a confident speaker who turns out to be right is damaged if the confidence expressed is based on weak evidence.

### 2.2.1. Method

*Participants.* The rationale for the sample size, the recruitment procedure and the exclusion criterion were the same as for Experiment 1. The final sample size comprised 109 native adult English speakers (30 men, 78 women, 1 non-binary/third gender, $M_{\text{age}} = 37.73$, $SD = 13.28$) recruited via Prolific and paid 0.88£/1.12$ for participating in the online experiment on Qualtrics. No participant was excluded from the data.

*Materials and design.* All the vignettes are given in the Supplementary material.

Time 1 of Experiment 2 was identical to Time 1 of Experiment 1. At Time 2, participants got feedback about the accuracy of the weather report and the passenger identification (but not about the rest of the deposition). Participants learned that both witnesses were accurate about the weather on the day of the accident and (contrary to Experiment 1) about the identification of the passenger. Participants rated again the credibility of each witness and made a binary decision as to which witness's deposition they believed, and justified this choice.

At Time 3, participants got feedback about the strength of the evidence that both witnesses had to identify the passenger. Participants read that the two witnesses had weak evidence to support their claims (contrary to Experiment 1). They were told that the CCTV camera footage of a restaurant revealed that the passenger had his back turned the whole time and his face was not visible from the perspective of the witnesses. The evidence was therefore perceptual and weak. Participants rated again the credibility of each witness as well as made a binary decision as to which witness's deposition they believed, and justified this choice.

As for Experiment 1, we expected that, in the absence of any information about accuracy and evidence (Time 1), the confident witness (i.e., confident about the accident, the weather, and the identification of the passenger) would be judged more trustworthy (i.e., he would be rated as more credible and would be more likely to be believed) than the unconfident witness (i.e., confident about the accident and the weather, but not about the identification of the passenger). At Time 2, when both informants turn out to be accurate (i.e., they identified the right passenger), the confident informant would keep his trustworthiness (both for credibility and believability). At Time 3, when the testimony of the informants is found to be warranted by weak evidence (i.e., the CCTV camera footage of a restaurant revealed that the passenger had his back turned the whole time and his face was not visible), the confident informant would suffer a reputational loss (both for credibility and believability).

### 2.2.2. Results

All the analyses were performed in R (v. 4.2.1; R Core Team, 2022) using R Studio (v. 2.4.1; RStudio Team, 2022). We fitted a cumulative linear mixed model (CLMM) for credibility ratings which included confidence and time as fixed effects, and participant as a random effect. Furthermore, we fitted a generalized linear mixed model (GLMM) for believability which included time as a fixed effect, and participant as a random effect. Data and R scripts are available on OSF (https://osf.io/z92tg/?view_only=7402859320fd43dea4b652e1ef4bfbb5). There was a significant interaction effect of confidence and time on credibility ($\chi^2(2) = 39.016$, $p < .001$, see Figure 3). Believability changed over time as predicted ($\chi^2(2) = 54.792$, $p < .001$, see Figure 4). For the post-hoc comparisons,[4] we used paired Wilcoxon signed-rank tests for credibility ratings, and binomial tests for believability. In the absence of any information about accuracy and strength of evidence (Time 1), the confident witness was judged more trustworthy: he was rated as more credible ($M = 4.42$, $M_d = 5$, $SD = 1.16$) than the unconfident witness ($M = 4.03$, $M_d = 4$, $SD = 0.99$, $Z = 1762.5$, $p = .002$, $r = .42$), and was more likely to be believed than the unconfident witness (65% confident witness, 71/109, binomial $p = .002$, Cohen's $g = .15$). At Time 2, when both witnesses turned out to be accurate, the confident witness kept his trustworthiness, and even increased it. The confident witness was still judged more trustworthy than the unconfident witness: he was rated as more credible ($M = 4.79$, $M_d = 5$, $SD = 1.24$) than the unconfident witness ($M = 4.10$, $M_d = 4$, $SD = 1.24$, $Z = 1697.5$, $p < .001$, $r = .58$), and was more likely to be believed than the unconfident witness (74% confident witness, 81/109, binomial $p < .001$, Cohen's $g = .24$). At Time 3, when the testimony of the

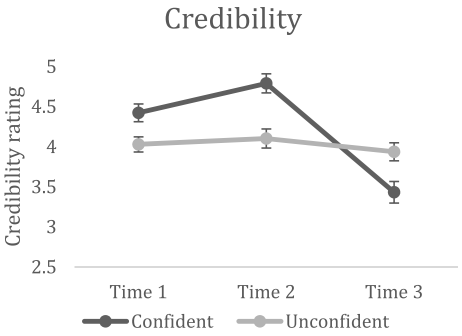

Figure 3. Credibility scores of the confident (black) and unconfident (gray) witnesses, on a scale from 1 "not credible" to 6 "credible".

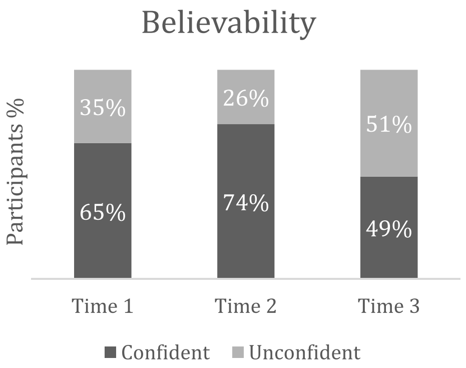

Figure 4. Percentage of participants who believed the deposition of the confident (black) or unconfident (gray) witness.

witnesses was found to be warranted by weak evidence, the confident witness suffered a reputational loss: he was rated as less credible ($M = 3.43$, $M_d = 4$, $SD = 1.41$) than the unconfident witness ($M = 3.94$, $M_d = 4$, $SD = 1.17$, $Z = 770.5$, $p = .001$, $r = -.43$), and was as likely to be believed as the unconfident witness (49% confident witness, 53/109, binomial $p = .848$, Cohen's $g = -.01$).

### 2.2.3. Discussion

The results of Experiment 2 supported our second hypothesis: Trustworthiness is damaged if an accurate and confident claim is not sufficiently evidenced. As in Experiment 1, at Time 1, when participants had information only about the witness's confidence, the confident witness was rated as more credible and was more likely to be believed than the unconfident witness. At Time 2, when both witnesses turned out to be accurate, the confident witness increased his perceived trustworthiness. This result confirms the importance of confidence-accuracy calibration: participants judged a well-calibrated (confident and accurate) speaker more trustworthy than a poorly calibrated (unconfident and accurate) speaker. Interestingly, though, this result also shows that underconfidence is not damaging (the unconfident witness did not lose their trustworthiness when he was found to be poorly calibrated), thus suggesting that underconfidence was not perceived as insincere or uncooperative.

Finally, at Time 3, when the well-calibrated (confident and accurate) witness turned out to be not justified (he was found to be accurate "by chance"), he lost his trustworthiness. This result supports the hypothesis that a speaker showing bad confidence-evidence calibration pays a reputational price even if this speaker displays a good confidence-accuracy calibration. This shows that confidence-evidence calibration can trump confidence-accuracy calibration when it comes to assessing speaker trustworthiness. Making a commitment that one *can* fulfil (in the epistemic sense of *can*) may therefore be more important than actually fulfilling the commitment.

## 3. General discussion

Epistemic vigilance allows humans to face the risk of misinformation and calibrate their trust toward the source of information. This calibration is sensitive to a variety of cues of speaker trustworthiness. In this paper, we were interested in two of these cues (and their interplay): the speakers' display of their epistemic commitments via the expression of confidence, and the quality of the evidence available to them, i.e., the confidence-evidence calibration. We tested two hypotheses. First, that overconfidence does not backfire when a confident but inaccurate claim is justified (Hypothesis 1). Second, that confidence can backfire if a confident and accurate claim is not justified (Hypothesis 2). By experimentally manipulating the time at which participants received information about the confidence, accuracy and evidence of the two witnesses, we were able to investigate their respective (and cumulative) contributions to participants' judgments of speaker trustworthiness and examine how they evolved. In both experiments, trustworthiness judgments appeared to be sensitive to good/bad confidence-accuracy calibration (Time 2). The trustworthiness of the confident witness who displayed a bad confidence-accuracy calibration was revised downwards (Experiment 1), while the trustworthiness of the confident speaker was reinforced when the confidence-accuracy calibration turned out to be good (Experiment 2). Interestingly, though, while confidence-accuracy calibration played a role in reassessing the trustworthiness of the confident speaker, it did not have a comparable impact on the perceived trustworthiness of the unconfident speaker. More specifically, the trustworthiness of the well-calibrated unconfident speaker was reinforced (Experiment 1), but bad confidence-accuracy calibration did not lead to any reputational loss for the unconfident speaker (Experiment 2).

Finally, the results show that, when assessing the trustworthiness of a confident speaker, confidence-evidence calibration plays a bigger role than confidence-accuracy calibration. Indeed, the trustworthiness of a confident and inaccurate speaker is adjusted upwards when participants receive information about the speaker's good confidence-evidence calibration (Experiment 1). Furthermore, speaker trustworthiness is revised downwards when a confident and accurate speaker is found to be poorly justified, thus displaying bad confidence-evidence calibration (Experiment 2). Interestingly, the impact of confidence-evidence calibration was stronger for the confident witness than for the unconfident witness, whose credibility ratings appear to be stable from Time 2 to Time 3 in both experiments.

Overall, our results supported our hypotheses and indicate that confidence-evidence calibration is crucial for speaker trustworthiness. On the one hand, a good confidence-evidence calibration can counteract (at least partially) the reputational costs related to a bad confidence-accuracy calibration. On the other hand, a bad confidence-evidence calibration can override a good confidence-accuracy calibration, and damage the perceived speaker trustworthiness.

Having good evidence in support of a confident claim thus matters more than actually saying the truth, and asserting an accurate information does not make us trustworthy if one lacks evidence for it (i.e., one is right by chance). One should thus avoid expressing confidence when one is not justified by evidence. Indeed, as Kominsky et al. (2016) suggest, in contexts in which it is not possible to have evidence (or high certainty), it is, therefore, more appropriate not to appear too confident. In these cases, showing confidence is a sign of incompetence (or miscalibration), while showing ignorance is a sign of expertise (knowing the limits), or what Kominsky and colleagues call "virtuous ignorance". Interestingly, Kominsky et al. (2016) found that people are more likely to believe a confident speaker when the information is knowable, and a cautious (or ignorant) speaker when the information is not knowable. Moreover, it has been claimed in the philosophy of lies that asserting something that one is not sure is true ("partial truths") can be as epistemically damaging for the addressee as asserting something one is not sure it is false ("partial lies"), and that speakers should thus only assert what they are confident about (Trpin et al., 2020).

The importance of the evidential cue for evaluating the speaker trustworthiness is reflected in the fact that the capacity to detect and use this cue develops quite early in ontogeny. A few studies have provided some insights into the role of evidence in children's assessment of speaker trustworthiness. For instance, Pillow (1989) showed that children preferred to learn from individuals whose testimony could be warranted by direct perceptual evidence, thus displaying the ability to infer another person's knowledge or ignorance on the basis of recent perceptual experience, and orient their trust choices accordingly. Interestingly for our study, Einav and Robinson (2011) showed that the assessment of speaker trustworthiness is sensitive to *how* speakers achieved their prior accuracy. Starting from the age of 4, when confronted with two equally accurate partners, children were more likely to trust the testimony of an informant who has demonstrated epistemic autonomy in the past (an unaided informant) than the testimony of an informant who relied systematically on a third party to provide accurate information (an assisted informant). This indicates that even children are sensitive to the fact that mere accuracy is not enough for true knowledge, and it may not be a predictor of speaker trustworthiness in the long term.

These results are echoed in the adult literature by studies that show that claims of first-hand evidence increase the probability to be trusted, and humans are therefore likely to use them when this is advantageous (Castelain et al., 2019). In contrast with this, claims of secondhand evidence (i.e., the evidence comes from a third party) decrease the likelihood to be trusted (Mahr & Csibra, 2021), although the identity of the source of information appears to play a crucial role: a message based on secondhand evidence is more likely to be accepted as true if the original source is credible and close to the speaker (Altay et al., 2020). Overall, this set of studies indicates that epistemic vigilance toward the source of information comprises mechanisms dedicated to the assessment of the evidential basis of a speaker's claims and that the output of this evaluation is used to calibrate trust.

Finally, it is worth discussing the relevance of our study to the distinction between overconfidence and underconfidence and its social implications. Researchers have mainly focused on one type of bad confidence-accuracy calibration, i.e., overconfidence (expressing confidence when the information is inaccurate). But speakers can also be poorly calibrated when they express uncertainty while the information is accurate, i.e., by being underconfident. This bad calibration may lead the addressee to disregard true information. Being too cautious may prevent accurate and potentially important information from being transmitted, and underconfident speakers should therefore also be exposed to costs. However, while underconfidence (or "understating") is also a sign of bad calibration, it is often perceived as sincerer (or less misleading) or more cooperative than overconfidence (or "overstating") (Marsili, 2018). This is confirmed by our results showing that an unconfident speaker is not judged as less trustworthy when the claim that he could have expressed with more confidence turns out to be true (poor confidence-accuracy calibration). Moreover, the trustworthiness of the unconfident speaker does not seem to be impacted by the confidence-evidence calibration.

This indicates that overconfidence is more damaging than underconfidence. Why is this the case? This may be explained in terms of commitment: while an underconfident speaker avoids any epistemic commitment, an overconfident speaker commits to the truth of the message communicated, making the violation of this commitment more costly (Marsili, 2018). In line with our results, this would imply that only committed speakers are impacted by poor confidence-accuracy as well as poor confidence-evidence calibration. This is arguably due to the relationship between speaker communicative benefits and reputational costs: great benefits (higher message acceptance, higher perceived credibility) imply high costs (reputation loss), while small benefits (lower message acceptance) imply low costs (no or smaller reputation loss). An unconfident speaker does not commit to the truth of the message and thus does not put their reputation at stake. In fact, when the risks for reputation are low, it is not advantageous for a speaker to express with unconfidence a claim for which they have good support (and for which they are thus more likely to be right), as they miss the opportunity of getting greater benefits. This may explain why underconfidence is rare compared to overconfidence (D. A. Moore et al., 2015).

These observations have implications for the way speakers should manage their reputation while using communication to achieve their goals. On one hand, while expressing confidence may increase the chance to get one's message accepted, it may damage one's epistemic reputation if this confidence is not justified, making overconfidence detrimental in the long term. On the other hand, while unconfidence may make us less convincing when confidence is the only cue that the addressee can use to evaluate our trustworthiness, it pays off in situations in which information cannot be known (Kominsky et al., 2016), or when one lacks adequate evidence (as suggested by our results).

To conclude, it is worth noticing that, in the present study, the role of confidence-accuracy calibration and confidence-evidence calibration in the assessment of speaker trustworthiness was investigated in the context of a courtroom decision in which the need for accuracy and evidence is particularly salient (see Borg & Connolly, 2022). Of course, the costs of bad calibrations on speaker trustworthiness may be more or less pronounced depending on the context, the relevance of the information, and the possible consequences for the receiver. Although we expect this overall pattern to generalize, future experimental studies should extend these findings to different conversational contexts.

## 4. Conclusions

Philosophical analyses of trustworthiness consider this notion as tightly linked with commitment. Specifically, Hawley (2014, 2019) defined trustworthiness as the avoidance of unfilled commitments. Being trustworthy would thus require fulfilling the commitments one has already acquired, and taking commitments one can fulfil. In this paper, we focused on speaker trustworthiness, that is, trustworthiness as a source of information via communication. The results of our experiments showed that failing to fulfil a commitment (i.e., failing to convey a true message when this is expressed with confidence) has a negative impact on speaker trustworthiness, but that taking a commitment that is not warranted (i.e., by expressing confidence based on weak evidence) appears to be even more costly for speakers' reputation as trustworthy sources of information. Furthermore, failing to fulfil a commitment is less costly if this commitment is warranted (i.e., the confident speaker has strong evidence). We thus suggest that Hawley's requirements to establish trustworthiness are not on a par: taking commitments that one can fulfil is more important than actually fulfilling them; one can still be considered as trustworthy even when failing to fulfil a commitment that one was justified in taking. This has important implications for the psychological literature on overconfidence: confidence-evidence calibration can override confidence-accuracy calibration, such that a speaker displaying good confidence-evidence calibration can still be considered as trustworthy even if they manifest poor confidence-accuracy calibration.

To conclude, this study shows that speaker trustworthiness as a source of information depends on how confidence expression is calibrated to the speaker's evidential basis, and that to keep their own reputation, a speaker should first and foremost commit to what they have evidence for. Committing to the truth of a message only when one is justified is a better sign of epistemic responsibility than actually sharing true information, which may be accidentally transmitted by people who are not deeply concerned about truth.

## Notes

1. While deniability is always possible, it is not always plausible, and implausible denials may have a detrimental effect on speaker trustworthiness (Hawley, 2019). See Mazzarella (2021) for an account of what makes a denial be perceived as plausible.
2. Kneer (2021) replicated these results in the U.S.A, Germany and Japan, showing that this justified belief norm is shared by different cultures and languages.
3. The statistical significance of the post-hoc comparisons is assessed based on a Bonferroni correction (i.e., $\alpha = 0.05/9 = 0.006$). Concerning the evolution of the confident witness credibility, he was rated as less credible at Time 2 ($M = 2.67$, $M_d = 3$, $SD = 1.07$) than at Time 1 ($M = 4.38$, $M_d = 4$, $SD = 1.05$, $Z = 4497.5$, $p < .001$, $r = .93$), more credible at Time 3 ($M = 4.07$, $M_d = 4$, $SD = 1.30$) than at Time 2 ($M = 2.67$, $M_d = 3$, $SD = 1.07$, $Z = 342.5$, $p < .001$, $r = -.84$), but less credible at Time 3 ($M = 4.07$, $M_d = 4$, $SD = 1.30$) than at Time 1 ($M = 4.38$, $M_d = 4$, $SD = 1.05$, $Z = 918.5$, $p = .005$, $r = .44$). Concerning the evolution of the unconfident witness credibility, he was rated as more credible at Time 2 ($M = 4.10$, $M_d = 4$, $SD = 1.24$) than at Time 1 ($M = 3.79$, $M_d = 4$, $SD = 1.10$, $Z = 513$, $p = .004$, $r = -.42$), as credible at Time 3 ($M = 3.94$, $M_d = 4$, $SD = 1.26$) as at Time 2 ($M = 4.10$, $M_d = 4$, $SD = 1.24$, $Z = 865$, $p = .097$, $r = .26$), and as credible at Time 3 ($M = 3.94$, $M_d = 4$, $SD = 1.26$) as at Time 1 ($M = 3.79$, $M_d = 4$, $SD = 1.10$, $Z = 257$, $p = .033$, $r = -.37$).
4. The statistical significance of the post-hoc comparisons is assessed based on a Bonferroni correction (i.e., $\alpha = 0.05/9 = 0.006$). Concerning the evolution of the confident witness credibility, he was rated as more credible at Time 2 ($M = 4.79$, $M_d = 5$, $SD = 1.24$) than at Time 1 ($M = 4.42$, $M_d = 5$, $SD = 1.16$, $Z = 230.5$, $p < .001$, $r = -.64$), less credible at Time 3 ($M = 3.43$, $M_d = 4$, $SD = 1.41$) than at Time 2 ($M = 4.79$, $M_d = 5$, $SD = 1.24$, $Z = 3637$, $p < .001$, $r = .90$), and less credible at Time 3 ($M = 3.43$, $M_d = 4$, $SD = 1.41$) than at Time 1 ($M = 4.42$, $M_d = 5$, $SD = 1.16$, $Z = 2662$, $p < .001$, $r = .82$). Concerning the evolution of the unconfident witness credibility, he was rated as credible at Time 2 ($M = 4.10$, $M_d = 4$, $SD = 1.24$) as at Time 1 ($M = 4.03$, $M_d = 4$, $SD = 0.99$, $Z = 649$, $p = .545$, $r = -.09$), as credible at Time 3 ($M = 3.94$, $M_d = 4$, $SD = 1.17$) as at Time 2 ($M = 4.10$, $M_d = 4$, $SD = 1.24$, $Z = 1052$,

$p = .195$, $r = .19$), and as credible at Time 3 ($M = 3.94$, $M_d = 4$, $SD = 1.17$) as at Time 1 ($M = 4.03$, $M_d = 4$, $SD = 0.99$, $Z = 1091$, $p = .399$, $r = .12$).

## Acknowledgements

We are very grateful to the editors and the two anonymous reviewers for their insightful comments and suggestions. We warmly thank Elizabeth Tenney for sharing the material of her study and Giorgio Arcara, Gökhan Gönül and Marie Aguirre for their advice concerning the statistical analyses. Finally, we thank Bianca Cepollaro for her valuable comments on the theoretical background of this work.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Mélinda Pozzi http://orcid.org/0000-0001-7702-2804
Diana Mazzarella http://orcid.org/0000-0002-7650-7196

## Data availability statement

Data and R scripts are available on OSF: https://osf.io/z92tg/?view_only=7402859320fd43dea4b652e1ef4bfbb5

## References

Altay, S., Claidière, N., & Mercier, H. (2020). It happened to a friend of a friend: Inaccurate source reporting in rumour diffusion. *Evolutionary Human Sciences*, *2*, E49. https://doi.org/10.1017/ehs.2020.53

Birch, S. A., Severson, R. L., & Baimel, A. (2020). Children's understanding of when a person's confidence and hesitancy is a cue to their credibility. *Plos One*, *15*(1), e0227026. https://doi.org/10.1371/journal.pone.0227026

Borg, E., & Connolly, P. J. (2022). Exploring Linguistic Liability. *Oxford Studies in Philosophy of Language*, *2*, 1. https://doi.org/10.1093/oso/9780192844613.003.0001

Boulat, K., & Maillat, D. (2017). She said you said I saw it with my own eyes: A pragmatic account of commitment. In *Formal models in the study of language* (pp. 261–279). Springer. https://doi.org/10.1007/978-3-319-48832-514

Castelain, T., Floyd, S., & Mercier, H. (2019). *Evidentiality and flexibility of source reporting.* https://doi.org/10.31234/osf.io/qpb82

Chow, V., Poulin-dubois, D., & Lewis, J. (2008). To see or not to see: Infants prefer to follow the gaze of a reliable looker. *Developmental Science*, *11*(5), 761–770. https://doi.org/10.1111/j.1467-7687.2008.00726.x

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology*, *9*(3), 203–235. https://doi.org/10.1007/bf02686861

Einav, S., & Robinson, E. J. (2011). When being right is not enough: Four-year-olds distinguish knowledgeable informants from merely accurate informants. *Psychological Science*, *22*(10), 1250–1253. https://doi.org/10.1177/0956797611416998

Elashi, F. B., & Mills, C. M. (2014). Do children trust based on group membership or prior accuracy? The role of novel group membership in children's trust decisions. *Journal of Experimental Child Psychology*, *128*, 88–104. https://doi.org/10.1016/j.jecp.2014.07.003

Engel, M., Jr. (2011). Epistemic Luck. In J. Fieser & B. Dowden (Eds.), *Internet Encyclopedia of Philosophy*. https://www.Iep.utm.edu/epi-luck/ .

Fricker, E. (2012). I—elizabeth Fricker: Stating and Insinuating. *Aristotelian Society Supplementary Volume*, *86*(1), 61–94. https://doi.org/10.1111/j.1467-8349.2012.00208.x Blackwell Publishing Ltd.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123. https://doi.org/10.1093/analys/23.6.121

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, vol. 3, speech acts* (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811003

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press. https://doi.org/10.4159/harvard.9780674065192

Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, *48*(1), 1–20. https://doi.org/10.1111/nous.12000

Hawley, K. (2019). *How to be trustworthy*. Oxford University Press. https://doi.org/10.1093/oso/9780198843900.001.0001

Kneer, M. (2018). The norm of assertion: Empirical data. *Cognition*, *177*, 165–171. https://doi.org/10.1016/j.cognition.2018.03.020

Kneer, M. (2021). Norms of assertion in the United States, Germany, and Japan. *Proceedings of the National Academy of Sciences*, *118*(37), e2105365118. https://doi.org/10.1073/pnas.2105365118

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694–698. https://doi.org/10.1111/j.0956-7976.2004.00742.x

Koenig, M. A., & Jaswal, V. K. (2011). Characterizing children's expectations about expertise and incompetence: Halo or pitchfork effects? *Child Development*, *82*(5), 1634–1647. https://doi.org/10.1111/j.1467-8624.2011.01618.x

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology*, *52*(1), 31. https://doi.org/10.1037/dev0000065

Mahr, J. B., & Csibra, G. (2021). The effect of source claims on statement believability and speaker accountability. *Memory & Cognition*, *49*(8), 1505–1525. https://doi.org/10.3758/s13421-021-01186-x

Marsili, N. (2018). Lying and certainty. In *The oxford handbook of lying* (pp. 169–182). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198736578.013.12

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380. https://doi.org/10.1016/j.cognition.2009.05.012

Mazzarella, D. (2021). "I didn't mean to suggest anything like that!": Deniability and context reconstruction. *Mind & Language*, *38*(1), 218–236. https://doi.org/10.1111/mila.12377

Mazzarella, D., Reinecke, R., Noveck, I., & Mercier, H. (2018). Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics*, *133*, 15–27. https://doi.org/10.1016/j.pragma.2018.05.009

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development*, *60*(1), 167–171. https://doi.org/10.2307/1131082

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. *The Wiley Blackwell Handbook of Judgment and Decision Making*, *2*, 182–209. https://doi.org/10.1002/9781118468333.ch6

Pietrandrea, P. (2008). Certamente and sicuramente: Encoding dynamic and discursive aspects of commitment in Italian. *Belgian Journal of Linguistics*, *22*(1), 221–246. https://doi.org/10.1075/bjl.22.11pie

Pillow, B. H. (1989). Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology*, *47*(1), 116–129. https://doi.org/10.1016/0022-0965(89)90066-0

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. https://doi.org/10.1073/pnas.0707192105

Poulin Dubois, D., & Brosseau-Liard, P. (2016). The developmental origins of selective social learning. *Current Directions in Psychological Science*, *25*(1), 60–64. https://doi.org/10.1177/0963721415613962

Pritchard, D. (2003). Virtue epistemology and epistemic luck. *Metaphilosophy*, *34*(1–2), 106–130. https://doi.org/10.1111/1467-9973.00263

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. (v. 4.2.1). R Foundation for Statistical Computing. https://www.R-project.org/

Robinson, E. J., & Einav, S. (Eds.). (2014). *Trust and skepticism: Children's selective learning from testimony*. Psychology Press. https://doi.org/10.4324/9781315849362

RStudio Team. (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC. https://www.rstudio.com/

Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, *72*(4), 1054–1070. https://doi.org/10.1111/1467-8624.00334

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46–50. https://doi.org/10.1111/j.1467-9280.2007.01847.x

Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, *44*(5), 1368–1375. https://doi.org/10.1016/j.jesp.2008.04.006

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, *116*(3), 396–415. https://doi.org/10.1037/pspi0000150

Trpin, B., Dobrosovestnova, A., & Götzendorfer, S. J. (2020). Lying, more or less: A computer simulation study of graded lies and trust dynamics. *Synthese*, *199*(1–2), 1–28. https://doi.org/10.1007/s11229-020-02746-5

Vallinder, A., & Olsson, E. J. (2014). Trust and the value of overconfidence: A Bayesian perspective on social network communication. *Synthese*, *191*(9), 1991–2007. https://doi.org/10.1007/s11229-013-0375-0

Vullioud, C., Clément, F., Scott-Phillips, T., & Mercier, H. (2017). Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior*, *38*(1), 9–17. https://doi.org/10.1016/j.evolhumbehav.2016.06.002

Wesson, C. J., & Pulford, B. D. (2009). Verbal expressions of confidence and doubt. *Psychological Reports*, *105*(1), 151–160. https://doi.org/10.2466/pr0.105.1.151-160