CONFIDENCE AS AN EXPRESSION OF COMMITMENT: WHY MISPLACED EXPRESSIONS OF CONFIDENCE BACKFIRE

NON COPYEDITED VERSION, PLEASE DO NOT QUOTE

Vullioud, C.
Université de Neuchâtel
Centre de Sciences Cognitives
Espace Louis-Agassiz 1
Neuchâtel 2000
Switzerland

Clément, F.
Université de Neuchâtel
Centre de Sciences Cognitives
Espace Louis-Agassiz 1
Neuchâtel 2000
Switzerland

Scott-Phillips, T.
Evolutionary Anthropology Research Group
Department of Anthropology
Durham University
Dawson Building, South Road, Durham DH1 3LE, UK

Mercier, H.
Université de Neuchâtel
Centre de Sciences Cognitives
Espace Louis-Agassiz 1
Neuchâtel 2000
Switzerland
hugo.mercier@gmail.com

Abstract

Because communication can be abused by senders, it is not inherently stable. One way of stabilizing communication is for senders to commit to their messages. If a sender is committed to a message, she is willing to incur a cost (direct or reputational) if the message is found to be unreliable. This cost provides a reason for receivers to accept messages to which senders are committed. We suggest that expressions of confidence can be used as commitment signals: messages expressed more confidently commit their senders more. On this basis, we make three predictions: that confidently expressed messages are more persuasive (H1', already well established), that senders whose messages were accepted due to the senders' confidence but were then found to be unreliable should incur costs (H2'), and that if a message is accepted for reasons other than confidence, when it is found to be unreliable the sender should incur lower reputational costs than if the message had been accepted on the basis of the sender's confidence (H3'). A review of the literature revealed broadly supportive but still ambiguous evidence for H2' and no tests of H3'. In Experiments 1, 2, and 3 (testing H2') participants received the same advice from two senders, one being confident and the other unconfident. Participants were more likely to follow the advice of the confident sender, but once the advice was revealed to have been misguided, participants adjusted their trust so that they trusted the initially unconfident sender more than the confident sender. In Experiments 3 and 4 (testing H3') participants chose between either two senders differing in confidence or two senders differing in competence. Participants followed the advice of the confident sender and of the competent sender. When it was revealed that the advice was misguided, the confident sender suffered from a larger drop in trust than the competent sender. These results are relevant for communicative theories of overconfidence.

Keywords: Confidence, commitment, communication, overconfidence.

Word count: 9089

Confidence as an Expression of Commitment: Why Misplaced Expressions of Confidence Backfire

Communication between agents whose interests do not perfectly overlap is not inherently stable. Even if both could benefit from communication, the danger is always present that one would abuse communication for its own advantage. This observation holds at the proximal level and at the ultimate level. At the proximal level, economists and other social scientists have puzzled over the weight of ‘cheap talk’ (Farrell & Rabin, 1996): how can mere words influence others when lying is not inherently costly? At the ultimate level, evolutionary biologists have pointed out that communication can only be evolutionarily stable if it benefits both senders and receivers (Dawkins & Krebs, 1978; Krebs & Dawkins, 1984; Maynard Smith & Harper, 2003; Scott-Phillips, 2008). If senders do not benefit from communication, they stop sending; if receivers do not benefit from communication, they stop receiving. But what stops senders from sending signals that benefit only them, thereby threatening the stability of communication?

Several mechanisms can stabilize communication (Maynard Smith & Harper, 2003). For instance, some signals are inherently reliable because they cannot be faked—someone who says “I am not a mute” cannot be lying, a Red Deer stag can only emit some types of roars if it is large enough (see Maynard Smith & Harper, 2003). In humans, however, very few signals are of this sort, so that we need to resort to other mechanisms to ensure the stability of communication (Sperber et al., 2010). Here we focus on one of these mechanisms: commitment. We suggest that in human communication, senders commit to various degrees to their messages. A message to which the sender commits has, everything else equal, more influence on the receiver. One way to express commitment is confidence: an assertion uttered with more confidence commits its speaker more. We lay out and evaluate—through a literature review and four experiments—consequences of this view of expressions of confidence as commitment signals. In conclusion, we relate this view to theories that seek to explain overconfidence through its communicative effects.

Commitment and communication

Commitment can take many forms. Some consider that commitment can be purely internal. Such 'subjective commitment' (Fessler & Quintelier, 2013) consists in maintaining a course of action not because of its instrumental value, but because of its intrinsic qualities. Fessler and Quintelier (2013, p. 459) provide the example of a suicide bomber who follows through on his plans because this course of action reflects his moral outrage towards the targets of the bombing. In such a case, if the suicide bomber were to change his course of action, he would suffer no external costs, but psychic costs such as feeling he has betrayed a just cause. By contrast, objective commitment involves an actual cost attached to changing one's course of action (Fessler & Quintelier, 2013). Opening a retirement account which carries a heavy fee for withdrawals constitutes an objective commitment to saving for one's retirement. In this example, the costs are purely personal but many instances of objective commitment involve social costs. For instance, an individual who breaks a promise—which is a typical form of commitment—often only incurs reputational costs.

The risks an individual takes in committing—i.e. the chances of having to pay some costs if she fails to stay true to her commitment—should have a benefit, otherwise it is not clear why anybody would commit to anything. These benefits can take many forms—for instance, making sure that one isn't too poor upon retirement. In the context of communication, the benefit of commitment is typically increased credibility, and the ability to influence others credibility provides. When a receiver knows that a sender would incur some costs if her communication proved unreliable, this provides him with a reason to believe her. The role of commitment in communication can be more precisely laid out with the following hypothesizes:

> H1. Increased commitment should result in increased chances that a message is accepted, or increased weight granted to the message.[1]
>
> H2. If a message is found to have been unreliable (false, harmful), and the receiver had accepted the message on the basis of the sender's commitment, then the sender should suffer reputational costs.[2]

For commitment to play its hypothesized communicative role, it must be the case not only that a sender of unreliable signals suffers some costs (per H2), but also that these costs be higher than they would have been if she had not been committed. It is the cost added by commitment that allows commitment to play its role. We can thus add the following hypothesis:

> H3. If a message is accepted on another basis than commitment, and if the message is found to have been unreliable, then the sender's reputation should suffer less than if the message had mostly been accepted on the basis of commitment. This would happen for instance when a message is accepted because the receiver had deemed the sender competent.

[1] Some caveats, which are not explored here, should be added to this hypothesis. The increased trust that results from increased commitment should be seen as multiplying the a priori trust in the sender rather than adding to it, so that completely mistrusted senders cannot rely on commitment to get their messages across. Moreover, expressed degrees of confidence that are implausibly high (e.g. "I am 100% sure my lottery number will come out") should also be dismissed.

[2] Reputational losses can affect either the perceived benevolence or the perceived competence of the sender (see Sperber *et al.*, 2010). In theory the loses due to failed commitments should mostly bear on the sender's perceived benevolence, but given that this prediction was not tested here, we do not elaborate further on this point.

Expressions of confidence as commitment signals

At least since Schelling's foundational work (Schelling, 1960), the communicative benefits of commitment have received much attention (in an evolutionary perspective, see, e.g. Fessler & Quintelier, 2013; Nesse, 2001). This attention has mostly focused on explicit commitments, such as promises (e.g. Schelling, 2001). However, other speech acts also commit their sender. In particular, assertions commit their sender to the truth of the proposition expressed (e.g. Searle, 1969). This suggests that a sender whose assertions are found to be false would suffer reputational costs. In practice, the distinction between speech acts is often blurred (e.g. Astington, 1988), and what matters is not simply whether one's speech act is, say, a promise or an assertion, but the degree of commitment that the sender expresses.

Human languages possess a variety of devices that enable senders to modulate their degree of commitment (Moeschler, 2013; Morency, Oswald, & de Saussure, 2008). For instance, a sender is more committed to the propositional content of her utterances than to their implicatures (Moeschler, 2013). Expressions of confidence also likely affect the degree to which the sender is understood by receivers to be committed to her statements. Expressions of confidence are ubiquitous in human communication, be they verbal ("I'm sure," "I guess," etc.) or non-verbal (gestures, tones, facial expressions). Indeed, the mechanisms which allow senders to gauge their level of confidence might have evolved for the purpose of communication (Shea et al., 2014). If expressions of confidence play the role of commitment signals, then the hypotheses formulated above about commitment in general should apply to expressions of confidence:

H1’. Increased confidence should result in increased chances that a message is accepted, or increased weight granted to the message (the same caveats as above apply).

H2’. If a message is found to have been unreliable (false, harmful), and the receiver had accepted the message on the basis of the sender’s confidence, then the sender should suffer reputational costs.

H3’. If a message is accepted on another basis than sender confidence, and if the message is found to have been unreliable, then the sender’s reputation should suffer less than if the message had mostly been accepted on the basis of confidence. This would happen for instance when a message is accepted because the receiver had deemed the sender competent.

H1’ and H2’ are similar to the hypotheses laid out about calibration in Tenney *et al*. (2008, p. 1369). In support of H1’, many experiments have revealed that confidence tends to increase the influence messages have on receivers (see, e.g., Price & Stone, 2004; Tenney, Small, Kondrad, Jaswal, & Spellman, 2011; Yaniv, 1997; and, for children, Brosseau-Liard, Cassels, & Birch, 2014). The goal of this article is to review the evidence relevant to H2’, to further test H2’, and to offer the first—to the best of our knowledge—tests of H3’.

Do receivers punish senders who were confident but wrong?

Experiments relevant to evaluating H2’ have yielded contradictory results. A first series of experiments unambiguously supports H2’. Tenney and her colleagues (Tenney, MacCoun, Spellman, & Hastie, 2007; Tenney et al., 2011, 2008) confronted participants with the testimony of two senders whose confidence calibration was manipulated. For instance, in

Experiment 1 of Tenney *et al.* (2008), the participants had to evaluate the testimony of two witnesses on the basis of the accuracy of two collateral statements (i.e. statements unrelated to the case used to evaluate the reliability of a witness' testimony). One witness was confident for both statements, while the other was confident for one statement and unconfident for the other. At first, the participants did not know whether the statements were accurate; they were then more likely to trust the confident witness. It was then revealed that each witness had been mistaken about one statement. As a result, the confident witness was poorly calibrated, having held confidently an inaccurate statement. In one condition, the less confident witness was well calibrated since she had been wrong on the uncertain statement. In this condition, after the accuracy feedback the participants found the less confident but better calibrated witness to be more credible than the more confident but less well calibrated witness, and they were more likely to believe her testimony. This experiment offers strong support for H2'. The participants initially accepted a piece of testimony because its sender was confident. When the confidence of the sender was revealed to have been unwarranted, the participants chose to trust a sender who had been less confident but who had not expressed unwarranted confidence.

Other experiments have found ambiguous support for H2'. Sah *et al.* (2013) asked participants to gauge the weight of individuals on the basis of a picture of these individuals and someone else's opinion (the senders' opinion). The senders had either high or low confidence, and they were either very accurate or very inaccurate. Inaccurate and confident senders were deemed, after the task had been completed, to be less credible than inaccurate and unconfident senders. However, the opinions of inaccurate but confident senders were not taken into account less than that of the inaccurate and unconfident senders (in spite of the absence of floor effects). In another type of experiment, participants had to evaluate two candidates: one who was very confident in his abilities, and one who was more cautious

(Tenney & Spellman, 2011). At first, confidence paid off, with better ratings for the confident candidate. Once it was revealed that both candidates had in fact the same qualities, they were both rated equally well. Thus, although confidence had no positive affect after it was revealed to have been mistaken, it had no negative effects either (at least in the short term, see below).

Another set of studies, using a very different methodology, reached similar conclusions (Kennedy, Anderson, & Moore, 2013). Participants first completed half of a task in a small group. They were then asked their perception of the status and the competence of each group member, including themselves. The participants' actual performance was then revealed to all. The groups reconvened and completed the second half of the task before answering the same status and competence questions. Participants were considered overconfident if they estimated their status to be higher than warranted by their actual performance. As in Anderson et al. (2012), before the participants had received performance feedback, those who were overconfident were seen as being more competent and as having a higher status (supporting H1'). After the performance feedback, the positive effects of overconfidence disappeared, but there were no negative effects (i.e. the participants who were initially overconfident were not perceived less well than those who had been initially well calibrated).

These studies (Kennedy et al., 2013; Sah et al., 2013; Tenney & Spellman, 2011) seem to suggest that, contrary to H2', overconfident senders do not see their reputation suffer much. After their inaccuracy has been revealed, overconfident senders are not trusted less (or not much else in the case of Sah *et al.* 2013) than unconfident senders. However, these studies can be interpreted in a way that is compatible with H2'. They report a drop in trust or in status once someone is revealed to have been overconfident. That this drop does not compensate for the initial benefits of overconfidence might only reflect the scope of the experiments. If an individual had kept being overconfident, and this overconfidence had kept causing drops in trust and status, then that individual would have become less trustworthy, and would have been attributed lower status, than her better calibrated peers. Thus these studies do not flatly contradict H2'. Instead they suggest that for mistaken confidence to become costly, in some cases, it has to be large enough, or clear enough, or repeated enough times. That this is the case is suggested not only by the studies of Tenney *et al.* cited above—in which mistaken confidence might have been particularly salient—but also by the results of Paulhus (1998). In one of these experiments, participants met repeatedly over the course of several weeks. At first, self-enhancers—individuals who tend to be overconfident in their abilities—were perceived positively. After seven weeks, however, they were rated negatively on a variety of traits.

Overall, the evidence regarding H2' is thus ambiguous, although we surmise that if the experiments cited above that do not directly support H2' had been extended, the costs of being confident but wrong would have become clearer, and thus their support for H2' clearer as well.

The present experiments

The literature offers ambiguous support in favor of H2', and H3' has not been tested. With the overarching goal of testing the role of commitment in the expression of confidence, the present experiments further test H2' and offer the first tests of H3'. All the experiments follow a similar template. Two senders provide advice to the participants, with varying degrees of confidence (all experiments) or competence (Experiment 3 and 4). The participants take this information into account. It is then revealed that at least one of the senders was mistaken, and participants are asked to decide which of the senders they would rather punish and which sender they would trust in the future.

In Experiments 1 and 2, both senders are equally wrong in their advice, only varying in the degree of confidence with which the advice is expressed. H2’ predicts that the participants will inflict a higher cost on the more confident sender (through lower trust in particular). In Experiments 3 and 4 a sender, who is more confident (in one condition) or more competent (in the other condition) than the other sender, is proven wrong. H3’ predicts that the confident but wrong senders will see their reputation suffer more than that of the equally wrong but competent sender.

The experiments were designed to be engaging for the participants. Experiment 1, which was conducted in a classroom, used a simple, realistic situation and videos of the senders. Experiments 2, 3, and 4, which were conducted online, used textual advice, but they entailed a real, immediate cost for participants who accepted the wrong message. Finally, Experiment 4 tested the evolution of the participants’ trust in the senders by asking them to make another potentially costly decision between advice provided by the same two senders.

Experiment 1 (a and b)

Method

Participants

Ninety undergraduate students (59 females; M_{Age} 20.1; SD = 1.77) from a Swiss University took part in Experiment 1a and 42 undergraduate students (27 females; M_{Age} 21.6; SD = 3.36), also from a Swiss University, took part in Experiment 1b. Both groups were French speaking.

Materials and procedure

Experiments 1a and 1b took place in classrooms, before students attended a lecture. The students had been asked to bring headphones and either a laptop or a smartphone, and they answered the questions online on these devices. Experiments 1a and 1b were identical except for the wording of one question that was found to have been unclear in 1a.

In order to set up the situation in which the advice would be given, the participants were asked to imagine: “You have just started working as a middle manager in a big company. You have to meet the Swiss manager for international coordination to organize an important trip. As you don’t know how to reach this manager, you rely on the coffee break to ask two of your colleagues”[3]

A short movie then started in which two actors from the university theater group played the colleagues. The first shot went from the corridor to the break room where the two colleagues were standing. A screen appeared stating: “Hello, I’m trying to reach the Swiss manager for international coordination. Do you know who is he and where I can find him?” In the following shot, one of the colleagues answered: “Hi! International coordination, I know him! It’s Mr. Descloux, in building L, for Lausanne. You can believe me, I’m sure it’s him.” His tone was confident (the surnames were not counterbalanced, but we see no plausible reason to believe that this will have caused the effects observed).

In the final shot, which only showed the remaining colleague, he said: “Hi, hmm, I don’t know but I think that for the international coordination, it’s Mr. Grandjean, in building B, for Bern. But I’m really not sure”. His tone was unconfident. Actors and presentation order (i.e. confident first / unconfident first) were counterbalanced across participants, for a total of four different films.

After they had watched the short movie, half of the participants were asked to choose which of the two pieces of advice they wanted to follow. This was done to see whether participants who had not explicitly stated that they would follow the advice of the confident sender would still be more likely to punish him. All the participants were then told that both

[3] All texts are translated from French.

colleagues had been wrong, that the international coordination manager the participants were looking for was in fact someone else in a different building. Thus the only difference between the two colleagues was how confident they had been in their wrong answers.

The participants were then asked two questions (here for Experiment 1a). The first was aimed at testing the participants' choice of which colleague they would like to directly punish: "A few day later, your boss asked you to find somebody to put the 2000 invitations for the collaborators' dinner in their envelope and to stamp them, during lunch break. You are the team leader of the two colleagues seen during the coffee break. Whom do you give the task to?" (punishment question). The second question bore on the reputation of the colleagues as a sender in an unrelated area: "Since you are new in town, you are looking for a good kindergarten for your kids. Whom do you ask advice from?" (trust question). For both questions, participants had to pick one of the two senders.

As explained in the results section, the answers to the direct punishment question proved surprising in light of the other results. To test whether participants had understood the direct punishment question as intended, a few days after they had taken part in the experiment, they were asked if they had understood the question as asking about a punishment (forcing someone to do a boring task) or about a reward (trusting someone with a task). Fifty percent of the participants had interpreted the question as being about a reward, thereby invalidating the answers.

To fix this problem, in Experiment 1b we asked two questions instead of a single punishment question, and made the wording unambiguous: "You are the team leader of the two colleagues you saw during the coffee break. You are about to start two new projects. The first project is a project that has no interest or importance. Taking part in this project is demeaning and can be seen as a kind of punishment." Participants answered the first question, then read and answered the second: "The second project is an important and interesting project for a big client. Taking part in this project is gratifying and can be seen as a kind of reward.” The order of presentation of the questions was counterbalanced between participants in both Experiment 1a and 1b.

Results

Experiment 1a.

Eighty-five percent of the participants, who had been asked whose advice they wanted to follow, decided to follow the advice of the confident sender (39/46; binomial p = .001).[4] For the remaining two questions, there were no differences between these participants and those who had not been asked to specify which of the pieces of advice they wanted to follow (Mann-Whitney; punishment question: Z= -.65, p = .514; reputation question: Z= -.46, p = .644), and their results were aggregated. For the question intended to bear on punishment, 63% of these participants chose the unconfident sender (57/90; binomial p = .015). However, as mentioned above, the punishment question was problematic. For the trust question, 71% of the participants trusted the unconfident sender (64/90; binomial p < .001).

Experiment 1b.

Seventy-six percent of the participants, who had been asked whose advice they wanted to follow, decided to follow the advice of the confident sender (16/21; binomial p = .027). Overall, 83% of the participants (35/42; binomial p < .001) answered that they would punish the highly confident sender while 69% (29/42; binomial p = 0.02) answered that they would reward the unconfident sender.

In Experiment 1, participants behaved in line with H1’—they were more likely to believe a confident than an unconfident sender—and with H2’—they were more likely to punish, and less likely to trust, on an unrelated matter, a sender who was confident but was

[4] All data are available in the ESM.

then proven to have been wrong, than a sender who had been equally wrong but had been less confident. Experiment 2 seeks to replicate the result regarding H2’ using a different context and different tools (online experiment).

Experiment 2 (a and b)

Method

Participants

Forty participants (17 females; M_{Age} 33.30; SD = 11.01) took part in experiment 2a, and 42 participants (16 females; M_{Age} 34.05; SD = 10.94) in experiment 2b. The participants were recruited through Amazon Mechanical Turk (MTurk).

Materials and procedure

Experiments 2a and 2b were conducted online. Experiment 2a is a conceptual replication of Experiment 1 in which the two senders provide the same advice, only varying in their degree of confidence. Experiment 2b is a control experiment in which the senders give correct pieces of advice, designed to insure that participants do not have a general bias against confident senders. We predict that in Experiment 2b, by contrast to Experiment 2a, participants will punish less and trust more, after feedback, the more confident sender.

Each experiment comprised two tasks. In the first task, participants took on the role of an adviser. They were told that another participant would have to type a text, and that they had to advise them about which text they should type in order to make the typing easier and faster. The participants were presented, for a short time (5s) with two texts, one of which contained many difficult words which made it longer to type than the other (see ESM). The participants then had to say which text they would advise another participant to choose, and to write a short statement qualifying their answer. This first task had two goals. First, to make it more believable that the advice the participants received in the second task could have been given by another participant. Second, to make participants believe that the task of the advisor was not trivially easy, so that bad advice could be attributed to an honest error.

In the second task, the participants took on the role of the advisee. They had to choose one text among two to type, and were thus motivated to choose the text that would take less time to type. The only indication they had as to which text would take less time to type took the form of two pieces of advice provided by senders described as previous participants in the experiment (in fact we created the pieces of advice ourselves). Both senders advised to select the same text, but they offered different statements in support of their advice. The confident sender's advice was accompanied by this statement "I'm 100% sure this text is the easiest of the two," while the unconfident sender's advice was accompanied by "It was very quick, I couldn't see well, so I'm not so sure." Both senders were either wrong (Experiment 2a) or right (Experiment 2b), only differing in their degree of confidence. After participants had chosen a text to type, they were told that they had chosen either the longer text (Experiment 2a) or the shorter text (Experiment 2b). The participants then typed the texts, which took approximately one minute for the long text—in Experiment 2a—and 30 seconds for the short text—Experiment 2b). The texts were provided to them in a picture format, so that they could not cut and paste their content, and the participants could not move on to the next screen until they had typed the exact text provided.

After they had typed the texts, participants were asked two forced-choice questions similar to the questions asked in the first experiment. The first was aimed at testing the participants' choice of which senders they would like to directly punish: "If you could stop one of the two participants whose advice you saw from receiving a bonus, who would you pick?" (punishment question). The second question bore on the reputation of the senders: "If you had to do the experiment again, and you could only receive advice from one of these two participants, who would you pick?" (trust question). Question order as well as order of presentation of the senders (confident vs. unconfident) were counterbalanced between participants. The detail of all the information provided to the participants, screen by screen, is available in the ESM.

Results

Experiment 2a.

Eighty-eight percent of the participants (35/40; binomial $p < .001$) picked the text advised by the two senders. Participants who had not followed the recommendation of the senders were excluded of further analysis. All of the participants who had followed the advice preferred to punish the confident sender (35/35, binomial $p < .001$) and 91% (32/35; binomial $p < .001$) indicated that they would trust the unconfident sender.

Experiment 2b and comparison with 2a.

Seventy-six percent of the participants (32/42; binomial $p = .001$) followed the advice given by the two senders. Participants who had not followed the recommendation of the senders were excluded of further analysis. Twenty-two percent of the participants who had followed the advice (7/32; binomial $p = .002$) preferred to punish the confident sender and 9% (3/32; binomial $p < .001$) indicated that they would trust the unconfident sender to complete the task again. Compared to Experiment 2a, in Experiment 2b participants were more likely to punish the unconfident sender (Mann-Whitney $Z= -6.6$, $p < .001$), and more likely to trust the confident sender as a better sender (Mann-Whitney $Z= -6.7$, $p < .001$).

Experiment 2a supports H2': between two senders who were equally wrong, participants tend to punish more, and to trust less, the more confident sender. Experiment 2b shows that these results do not stem from a general bias against confident senders. When both senders are equally right, participants tend to punish more, and to trust less, the less confident sender. Experiment 3 uses the same procedure as Experiment 2 in order to test H3’ (and incidentally to replicate Experiment 2a).

Experiment 3

Method

Participants

Ninety-nine participants (34 females; M_{Age} 32.58; SD = 8.84) were recruited through Mturk to participate in experiment 3.

Materials and procedure

Experiment 3 was designed to test H3’ through a between-participant design with two conditions: a Competence Condition and a Confidence Condition. The Competence Condition was broadly similar to Experiment 2a with three crucial differences. As in Experiment 2a, the participants were confronted with two senders. The first departure from Experiment 2a is that the difference in confidence between the sender was removed: both used neutral expressions to accompany their advice (“Text one looked like the shorter one to type” and “Seemed to be overall the easier of the two” respectively). Second, a difference of competence between the two senders was introduced. Participants were told that the two senders had different track records of success at the task of picking the easier: “in previous experiments, he or she [i.e. the sender] correctly chose the easiest text 12 out of 13 times [respectively 3 out of 13 times].” Third, the two senders advised to select different texts—while they advised to select the same text in Experiment 2a. We introduced this change in order to test whether competent senders were initially believed more or less than confident senders (relative to incompetent and unconfident senders respectively).

The Confidence Condition was identical to Experiment 2a with one exception: as in the Competence Condition, the two senders provided different advice about which texts to select. Since the test of H3' consists in a comparison across conditions, what matters is not that the senders within each condition are equally wrong, but that the confident sender and the competent sender are equally wrong, as is the case (since they are both wrong while the other sender is right). Order of presentation of the senders, question order, and, in the Competence Condition, matching between the neutral statements and the senders, were counterbalanced. The questions were the same as in Experiment 2.

To summarize, participants in the Competence Condition were exposed to two senders of different initial competence, while participants in the Confidence Condition were exposed to two senders of different confidence. In both conditions, the sender who we expected to be initially believed—the competent sender in the Competence Condition and the confident sender in the Confidence Condition—was proven wrong. We then asked participants questions about which sender they would like to inflict costs on—either directly or by trusting them less.

Results

Participants were as likely to select the text advised by the confident speaker in the Confidence Condition (76%, 37/49; binomial $p < .001$) than they were to select the text advised by the competent speaker in the Competence Condition (78%, 39/50; binomial $p < .001$) (Mann-Whitney Z= -.29, $p = .770$). Participants who had not followed the advice of the competent or the confident senders were excluded from further analyses.

In the Confidence Condition, 86% of the participants (32/37; binomial $p < .001$) preferred to punish the confident sender and 89% (33/37; binomial $p < .001$) indicated that they would trust the unconfident sender to complete the task again. In the Competence

Condition, 77% of the participants preferred to punish the competent sender (30/39; binomial *p* = .001) but only 56% (22/39; binomial *p* = .522) indicated that they would trust the incompetent sender to complete the task again. There was no difference between the conditions in the answers to the punishment question (Mann-Whitney *Z*= -1.1, *p* = .286), but the participants who had accepted the advice of the competent senders were more likely to trust them in the future than the participants who had accepted the advice of the confident senders were to trust the confident senders in the future (Mann-Whitney *Z*= -3.7, *p* = .002).

By contrast with the other experiments, in Experiment 3 one sender was right while the other was wrong. We could hardly have expected participants to punish the sender who actually gave them sound advice over the one who gave them unreliable advice, even if the latter was more competent. As a result, the answers to the punishment question are not as relevant here as in the other experiments: they cannot properly test H3'. By contrast, the trust question can adequately test H3', since a participant can trust someone who has been wrong once over someone who has been right once, if other factors make up for this difference. The results of the trust question support H3'. Even though initial trust was equally high in the confident sender and the competent sender, and that both senders proved to be equally wrong, final trust was higher in the competent sender than in the confident sender. This shows that, when other factors are controlled for, the reputation of the confident sender suffered more than that of the competent sender. Experiment 4 replicates Experiment 3 while increasing its validity by introducing stakes in the final trust question.

Experiment 4

Method

Participants

Seventy-nine participants (37 females; M_{Age} 32.49; SD = 9.99) were recruited through MTurk to take part in the experiment.

Materials and procedure

Experiment 4 was similar to Experiment 3 with two differences. First, instead of asking participants who they would pick if they had to complete the task again, participants had to actually perform the same task, receiving advice from the same two senders. In this second task, both senders advised to select different texts so that we could measure which sender was trusted more. In the Competence Condition, the success rates of the two senders were provided again, having been updated to account for their failure in the first task. In the Confidence Condition, two new statements were adapted from those previously written by participants to express confidence ("I am absolutely sure in my decision") and lack of confidence ("looked like easier to type but I'm not really sure"). Second, given that we had established in Experiment 3 that confidence and competence had the same influence on participants' choices in the first task, in this first task both senders in each condition gave the same advice. This makes the results of the punishment question more interesting.

Results

In the Confidence Condition, 93% of the participants (37/40; binomial $p < .001$) selected the text advised by both senders in the first task; in the Competence Condition, 95% of the participants did so (37/39; binomial $p < .001$). Participants not following the advice of the two participants were excluded from further analyses.

In the Confidence Condition, 85% of the participants (31/37; binomial $p < .001$) preferred to punish the confident sender and 65% (24/37; binomial $p = .099$) trusted the advice of the unconfident sender in the second task. In the Competence Condition, 49% of the participants (18/37; binomial p = 1.00) preferred to punish the competent sender and 30% (9/37; binomial p = .020) trusted advice of the incompetent sender in the second task. Participants were more likely to punish the confident sender than the competent sender (Mann-Whitney Z= -3.2, p = .002) and they were more likely to trust the competent sender than the confident sender (Mann-Whitney Z= -3.0, p = .003).

These results offer strong support for H3'. Even though the confident sender's message and the competent sender's message were revealed to have been equally wrong, the confident sender was subsequently punished more, and trusted less than the competent sender.

Replications

To ensure the reliability of our findings, we replicated the results from all online experiments (Experiments 2, 3, and 4). A total of 413 participants were recruited through MTurk in three sessions: Experiment 2a and 2b (8 excluded, final Ns: 2a=37, 2b=38, 27 females; M_{Age} 34.61; SD = 9.57), experiment 3 (11 excluded, final N=189; 83 females; M_{Age} 35.92; SD = 11.94), experiment 4 (12 excluded, final N=149; 78 females; M_{Age} 33.42; SD = 9.20). For Experiments 2a and 2b, which were simple conceptual replications of previous experiments, we used the same Ns as in the first version of the present experiments. For Experiments 3 and 4, which were more novel—therefore potentially more contentious—and which included a comparison across conditions, we doubled the number of participants recruited in the first version of these experiments.

The 31 participants who were excluded had said they had already taken part in the same experiment previously or were not sure that they had not.

Experiment 2a

Eighty-four percent of the participants (31/37; binomial p < .001) selected the text advised by the two senders. Participants who had not followed the recommendation of the senders were excluded of further analysis. Seventy-one percent of the participants who had followed the advice preferred to punish the confident sender (22/31, binomial *p* = .029) and 74% (23/31; binomial *p* < .011) indicated that they would trust the unconfident sender. These results thus closely replicate those obtained previously.

Experiment 2b and comparison with 2a

Ninety-two percent of the participants (35/38; binomial *p* < .001) followed the advice given by the two senders. Participants who had not followed the recommendation of the senders were excluded of further analysis. Thirty-one percent of the participants who had followed the advice (11/35; binomial *p* = .041) preferred to punish the confident sender and 6% (2/35; binomial *p* < .001) indicated that they would trust the unconfident sender to complete the task again. Compared to Experiment 2a, in Experiment 2b participants were more likely to punish the unconfident sender (Mann-Whitney *Z*= -3.2, *p* = .001), and more likely to trust the confident sender as a better sender (Mann-Whitney *Z*= -5.7, *p* < .001). These results thus closely replicate those obtained previously.

Experiment 3.

Participants were as likely to select the text advised by the confident speaker in the Confidence Condition (66%, 61/92; binomial *p* = .002) than they were to select the text advised by the competent speaker in the Competence Condition (71%, 71/97; binomial *p* < .001) (Mann-Whitney *Z*= -1.03, *p* = .303). Participants who had not followed the advice of the competent or the confident senders were excluded from further analyses.

In the Confidence Condition, 80% of the participants (49/61; binomial *p* < .001) preferred to punish the confident sender and 72% (44/61; binomial *p* = .001) indicated that they would trust the unconfident sender to complete the task again. In the Competence Condition, 54% of the participants preferred to punish the competent sender (38/71; binomial *p* = .635) but only 45% (32/71; binomial *p* = .477) indicated that they would trust the incompetent sender to complete the task again. Participants in the Confidence Condition were more likely to punish the confident sender than the participants in the Competence Condition were to punish the competent sender (Mann-Whitney *Z*= -3.2, *p* = .001), and the participants who had accepted the advice of the competent senders were more likely to trust them in the future than the participants who had accepted the advice of the confident senders were to trust the confident senders in the future (Mann-Whitney *Z*= -3.1, *p* = .002). These results thus closely replicate those obtained previously. The only potential difference was that participants were significantly more likely to punish the confident sender than the competent sender, whereas this difference was not significant in the original experiment. The results of the replications are thus, if anything, even more in line with H3'.

Experiment 4

In the Confidence Condition, 88% of the participants (68/77; binomial *p* < .001) selected the text advised by both senders in the first task; in the Competence Condition, 83% of the participants did so (60/72; binomial *p* < .001). Participants not following the advice of the two participants were excluded from further analyses.

In the Confidence Condition, 66% of the participants (45/68; binomial *p* = .010) preferred to punish the confident sender and 50% (34/68; binomial *p* = 1.000) trusted the advice of the unconfident sender in the second task. In the Competence Condition, 37% of the participants (22/60; binomial *p* = .052) preferred to punish the competent sender and 28% (17/60; binomial *p* = .001) trusted advice of the incompetent sender in the second task. Participants were more likely to punish the confident sender than the competent sender (Mann-Whitney *Z*= -3.2, *p* = .001) and they were more likely to trust the competent sender than the confident sender (Mann-Whitney *Z*= -2.5, *p* = .013). These results thus closely replicate those obtained previously.

General discussion

The goal of this series of experiments was to test two hypotheses: H2’, that senders whose messages are accepted because they are confident suffer a reputation loss when their messages are found to have been misleading; and H3’, that this reputation loss is greater than that incurred by senders whose messages were accepted for other reasons (here, competence). Incidentally, all experiments also found support for H1’ (confidence increases message acceptance). Experiments 1, 2, and 4 support H2’. In these experiment, participants receive the same advice from two senders, one being confident and the other unconfident. At first, participants are more likely to follow the advice of the confident sender. However, once the advice is revealed to have been misguided, participants adjust their trust so that they trust the initially unconfident sender more.

Experiments 3 and 4 support H3’. Participants choose between either two senders differing in confidence or two senders differing in competence. At first, participants follow the advice of the confident sender and of the competent sender—and they do so equally strongly. When it is revealed that the advice is misguided, participants are more likely to trust the initially unconfident sender. By comparison, the drop in trust incurred by the competent but wrong sender is less severe, since after the feedback, the participants either do not trust the competent sender less than the incompetent sender (Experiment 3), or they keep trusting the competent sender more (Experiment 4). Experiments 2, 3, and 4 were successfully replicated using the same population, demonstrating the robustness of their results. The results from all the experiments are summarized in Table 1.

Experiment	1a	1b	2a	3			4		
N	90	42	40 _37_	49 _92_		50 _97_	40 _77_		39 _72_
Which sender incorrect?	Both senders incorrect		Both senders incorrect	Only confident or competent sender incorrect			Both senders incorrect		
Trait of the favored sender	Confident		Confident	Confident		Competent	Confident		Competent
Initially trust the favored sender	85%	76%	88% _84%_	76% _66%_		78% _71%_	93% _88%_		95% _83%_
After feedback: Punish the favored sender	Question unclear	83%***	100% *** _71%*_	86% _80%_	 **	77% _54%_	85% _66%_	** **	49% _37%_
After feedback: Trust the favored sender	29%***	31%**	9% *** _26%*_	11% _28%_	** **	44% _55%_	35% _50%_	** *	70% _72%_

Table 1: Summary of results for Experiments 1 to 4. In all the experiments the favored sender (confident or competent) was wrong. The results of the replications of experiments 2 to 4 are in italics. For the two 'After feedback' lines, the percentages are computed on the basis of the participants who trusted the favored sender (confident or competent). Stars denote the level of significance (* < 0.05, ** < 0.01, *** < 0.001). For Experiments 1 and 2, stars represent differences from chance performance. For Experiments 3 and 4, stars represent differences between conditions.

Besides offering support for H3', which had not been previously tested, the present experiments extend the literature related to H2' in different ways. Given the ambiguity in the existing literature regarding H2', the simple adjunction of more evidence in support of this hypothesis is pertinent. Moreover, the experiments extend previous results supporting H2' (in particular the experiments of Tenney *et al.* 2007, 2008, 2011) in at least two ways. First, the experiments reveal that the drop in reputation incurred by overconfident senders extends beyond the domain in which they have been found to be overconfident: In Experiment 1a, the senders were proven to be wrong on a work matter, and they were then less trusted on a family matter. This suggests that experiments which only test for a drop of trust following confident but unreliable messages in the same domain as that of the message might be underestimating the costs of mistaken confidence. Second, three of the experiments introduced costs for following the misguided advice (having to type a much longer text, Experiments 2, 3, and 4) and one incentivized the choice of which sender to trust after the feedback (following the best advice might lead participants to type a shorter text, Experiment 4).

Conclusion

One of the mechanisms senders rely on to get receivers to accept their messages is commitment. By committing to their messages, they accept to incur a cost if the messages are found to be unreliable (H2), a cost that has to be greater than the cost they would have incurred if their unreliable messages had been accepted for reasons other than commitment (H3). Knowing of this cost, receivers have a reason to accept the messages senders commit to (H1). We suggested that expressions of confidence could play the role of commitment signals, leading to the formulation of the equivalent hypotheses for confidence instead of commitment more generally (H1’, H2’, and H3’, see General Discussion above). Four experiments provided incidental support for H1’ (which was already solidly supported), new support for H2’ (which was supported, but only ambiguously), and some initial support for H3’ (which had never been tested).

Considered with the existing evidence reviewed above, we now believe there is strong support for H2’: it seems that mistaken confidence backfires and hurts senders. Even though what we have developed here is a (partial) theory of expressed confidence, and not a theory of overconfidence more generally, the hypotheses we examined (H2' in particular) are relevant for some theories of overconfidence. H2' provides another reason why overconfidence should be costly: not only can overconfidence lead to personally damaging decisions (e.g., Barber & Odean, 2001), but, if expressed, it can hurt one's reputation. This extra cost makes it even more puzzling that overconfidence seems to be such a common phenomenon (e.g., Kahneman, 2011).

Some theories of overconfidence posit that overconfidence yields benefits that outweigh its costs. For instance, a model suggests that overconfidence allows agents to compete more effectively over resources (Johnson & Fowler, 2011), and another that overconfidence leads to a better mental health (Taylor & Brown, 1988). By contrast, other theories have suggested that some forms of overconfidence exist because the expression of overconfidence yields benefits that are conferred by others (social benefits). In particular, according to the status-enhancement theory of overconfidence, overconfidence confers social benefits because "overly positive self-views help individuals convince others that they are more capable than they actually are" (Anderson, Brion, Moore, & Kennedy, 2012, p. 718; see also, e.g. Trivers, 2011).

The status-enhancement theory of overconfidence predicts that individuals who express overconfidence should get social benefits, and that these benefits should be higher than the costs they might incur if their overconfidence were revealed (Kennedy et al., 2013). In this theory, overconfidence is not necessarily attached to specific statements (as in the present experiments), but rather with one's general abilities. However, we believe that such confidence might still constitute a form of commitment: people would commit not to a specific statement, but to the strength of their general abilities. If this were the case, then overconfidence should be punished, in that individuals who are consistently confident beyond their abilities should be seen as less reliable than individuals whose confidence matches their abilities.

As noted above, some experimental results suggested that overconfident individuals were not punished in this way (Kennedy et al., 2013). However, in these experiments participants still decreased the trust they granted overconfident individuals when their overconfidence was revealed. If we extrapolate from this trend, then an individual who would remain overconfident, or who would be too overconfident from the start, would end up being trusted less than a better calibrated individual. Indeed, as suggested in the introduction, this is what the rest of the literature (to which we can now add the current results) suggests.

This does not mean that the status-enhancement theory of overconfidence cannot apply in some cases. Individuals could be mistaken about the risks of overconfidence. Individuals could also find themselves in situations in which overconfidence has low costs, either because the senders' relative lack of competence is unlikely to be revealed (e.g. experts who make vague predictions), or because the senders are mostly engaged in one shot interactions (e.g. car dealers).

In spite of these potential exceptions, the idea that expressing overconfidence is not generally a successful strategy fits well with many results suggesting that some forms of overconfidence are not as robust as was once thought. Overconfidence can take at least the three following forms (Moore & Healy, 2008). Overplacement is saying that we are better than others when we are not (e.g. when most people believe they are smarter than the median individual). Overestimation is saying we are better than we are (e.g. when people say they can solve problems they can't solve). Overprecision is making statements that are more precise than warranted (e.g. when people say they are 95% sure the value of a stock will increase when in fact it has only 75% chances of increasing).

Overplacement and overestimation are not robust. Many studies that were supposed to demonstrate overplacement and overestimation have been contested on statistical grounds (Benoît & Dubra, 2011; Harris & Hahn, 2011). The amount of overplacement and overestimation varies widely as a function of different factors: the relative difficulty of the questions on which participants have to estimate their performance (Lichtenstein & Fischhoff, 1977), the participants' culture (Heine & Lehman, 1995), the ease with which overplacement can be justified (Dunning, Meyerowitz, & Holzberg, 1989), the amount of feedback provided to the participants (Rose & Windschitl, 2008), and so forth (e.g. Galesic, Olsson, & Rieskamp, 2012). The amount of variation in overplacement and overestimation is such that reversals are common. For instance, participants tend to underestimate their performance on easy problems (Lichtenstein & Fischhoff, 1977), and they believe they are below average when it comes to uncommon abilities (Moore, 2007). Note that in these experiments confidence is usually not measured behaviorally (e.g. by testing which tasks the participants are willing to engage in), but by asking participants to express their degree of confidence. As a result, the current hypotheses should apply. Considerations of the potential social costs caused by unwarranted expressions of confidence might help explain the pattern of data.

By contrast with overplacement and overestimation, overprecision is much more robust (Moore, Tenney, & Haran, in press). Moreover, overprecision is the form of overconfidence which is closest to the overconfidence displayed by the senders in our experiments. We thus seem to face the following puzzle: being overprecise is costly yet common. We suggest that the conversational norm theory of overprecision can solve this puzzle (Yaniv & Foster, 1995). According to this theory, if people are overprecise, it is because they favor informativeness in the tradeoff between informativeness and accuracy. Since more precise statements are less likely to be accurate, overprecision tends to decrease accuracy. However, more precise statements are more informative. To take an extreme example, if you ask a realtor to estimate the value of your house and she says “between $10 and $100,000,000,” she is bound to be right, but her statement is also so vague as to be irrelevant (see, Sperber & Wilson, 1995).

The conversational norm theory of overconfidence is supported by data both on the sender’s side and on the receiver’s side. On the sender’s side, individuals appropriately tailor the preciseness of their statements to the context—for instance by providing more precise time when the individual who is asking is going to catch a train (Van der Henst, Carles, & Sperber, 2002). On the receiver’s side, participants prefer a precise estimate (e.g. between 140 and 150 for the number of countries belonging to the U.N.) to a vague one (50 to 300), even after the second is revealed to be more accurate (the correct answer was 159) (Yaniv & Foster, 1995). It thus seems that by making statements more relevant, overprecision yields some benefits for receivers.

Crucially, it is also possible that overprecision doesn’t entail any costs for receivers. For overprecise statements to be costly, they have to be taken at face value. For instance, when a participant discovers that another participant thought the number of countries belonging to the U.N. was between 140 and 150, she might not take that to mean that the participant is certain of this estimate, only that this is her best guess that would still be relevant enough to be useful. That this is how receivers understand messages is suggested by the fact that, everything else equal, receivers tend to heavily discount senders’ opinions when these opinions contradict their own views (see, e.g., Yaniv, 2004). Receivers would often be better off taking the senders’ opinion into account more, not less, so that even if the senders have been overprecise, this overprecision is likely to have played a positive role (i.e. by making receivers take the message into account more, even if not quite enough).

Moreover, overprecision does not seem to get in the way of the effective communication of degrees of confidence. When participants have to complete a perceptual task in dyads, they are able to determine which dyad member is more confident, so that the dyad is able to select the answer favored by the more confident participant, which is generally the correct answer (Bahrami et al., 2010). On balance, it thus seems that overprecision is beneficial rather than costly for receivers. It is thus not surprising that it should not be punished and that it should be so common and robust.

These considerations suggest that the prevalence of overprecision can be reconciled with the current hypotheses, in particular H2' which posits that expressed mistaken confidence should be costly. Our hypotheses about commitment and confidence do not bear on the literal meaning of the statements, but on the meaning attributed to the sender. For instance, a sender would obviously not commit to the literal meaning of an ironic or a metaphorical statement. Similarly, some contexts call for modesty, others for bluster. To the extent that senders manage to get their thoughts across effectively, then they should be understood to be committed to the appropriate degree. If overprecision does not, on average, lead receivers to attribute to senders a misleadingly high degree of confidence, then it should not be routinely punished.

An important challenge for future research is to integrate these insights from the experimental literature with general theories of pragmatics, and human communicative behavior more generally. For instance, some approaches emphasize the importance of 'face' (see Brown & Levinson, 1987). From that perspective expressions of different degrees of confidence—lower confidence in particular—allow the speaker to preserve face even when what they communicate might be false. In any case, there is much scope for further experimental research: existing experiments are rudimentary in contrast to the complexity of the expressions of commitment found in everyday dialogue. The current framework can create a useful bridge between linguistic, psychological, and evolutionary theories pertaining to the expression of confidence and commitment more generally.

Acknowledgements. H.M. acknowledges a grant from the Swiss National Science Foundation (grant PZ00P1_142388).

References

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718.

Astington, J. W. (1988). Children's understanding of the speech act of promising. *Journal of Child Language*, *15*(1), 157–173.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, 261–292.

Benoît, J.-P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, *79*(5), 1591–1625.

Brosseau-Liard, P., Cassels, T., & Birch, S. (2014). You Seem Certain but You Were Wrong Before: Developmental Change in Preschoolers' Relative Trust in Accurate versus Confident Speakers. *PloS One*, *9*(9), e108308.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge university press.

Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (pp. 282–309). Oxford: Basil Blackwell Scientific Publications.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: the role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, *57*(6), 1082–1090.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, *10*(110–118).

Fessler, D. M., & Quintelier, K. (2013). Suicide bombers, weddings, and prison tattoos: an evolutionary perspective on subjective commitment and objective commitment. In K. Sterelny, R. Joyce, B. Calcott, & B. Fraser (Eds.), *Cooperation and its Evolution* (pp. 459–484). Cambridge: MIT Press.

Galesic, M., Olsson, H., & Rieskamp, J. (2012). Social sampling explains apparent biases in judgments of social environments. *Psychological Science*, 956797612445313.

Harris, A. J., & Hahn, U. (2011). Unrealistic optimism about future life events: a cautionary note. *Psychological Review*, *118*(1), 135.

Heine, S. J., & Lehman, D. R. (1995). Cultural variation in unrealistic optimism: Does the West feel more vulnerable than the East? *Journal of Personality and Social Psychology*, *68*(4), 595.

Johnson, D. D., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317–320.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar Straus & Giroux.

Kennedy, J. A., Anderson, C., & Moore, D. A. (2013). When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence. *Organizational Behavior and Human Decision Processes*, *122*(2), 266–279.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind-reading and manipulation? In J. R. Krebs & N. B. Davies (Eds.), *Behavioural Ecology: An Evolutionary Approach* (Vol. 2ème, pp. 390–402). Oxford: Basil Blackwell Scientific Publications.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183.

Maynard Smith, J., & Harper, D. (2003). *Animal signals*. Oxford: Oxford University Press.

Moeschler, J. (2013). Is a speaker-based pragmatics possible? Or how can a hearer infer a speaker's commitment? *Journal of Pragmatics*, *48*(1), 84–97.

Moore, D. A. (2007). Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes*, *102*(1), 42–58.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502.

Moore, D. A., Tenney, E. R., & Haran, U. (in press). Overprecision in judgment. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making*. New York: Wiley.

Morency, P., Oswald, S., & de Saussure, L. (2008). Explicitness, implicitness and commitment attribution: A cognitive pragmatic approach. *Belgian Journal of Linguistics*, *22*(1), 197–219.

Nesse, R. (Ed.). (2001). *Evolution and the capacity for commitment*. New York: Russell Sage Foundation.

Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology*, *74*(5), 1197.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*, 39–57.

Rose, J. P., & Windschitl, P. D. (2008). How egocentrism and optimism change in response to feedback in repeated competitions. *Organizational Behavior and Human Decision Processes*, *105*(2), 201–220.

Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, *121*(2), 246–255.

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA.: Harvard University Press.

Schelling, T. C. (2001). Commitment: Deliberate Versus Involuntary. In R. Nesse (Ed.), *Evolution and the capacity for commitment* (pp. 48–56). New York: Russell Sage Foundation.

Scott-Phillips, T. C. (2008). Defining biological communication. *Journal of Evolutionary Biology*, *21*(2), 387–395.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, *25*(4), 359–393.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. New York: Wiley-Blackwell.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193.

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.

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: how young children and adults assess credibility. *Developmental Psychology*, *47*(4), 1065.

Tenney, E. R., & Spellman, B. A. (2011). Complex social consequences of self-knowledge. *Social Psychological and Personality Science*, *2*(4), 343–350.

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.

Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. New York: Basic Books.

Van der Henst, J.-B., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind & Language*, *17*(5), 457–466.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, *69*(3), 237–249.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*, 1–13.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424.