Why Are We Convinced by the Ad Hominem Argument?: Bayesian Source Reliability and Pragma-Dialectical Discussion Rules

Mike Oaksford and Ulrike Hahn

Abstract There has been little empirical research on the ad hominem argument. What there is has been carried in the tradition of the pragma-dialectic approach by investigating the reasonableness of the ad hominem argument which is determined by the discussion stage in which it is deployed (van Eemeren et al., Fallacies and judgements of reasonableness. Springer, Dordrecht, 2009). The experiment reported in this chapter investigates how convincing people find the ad hominem argument from the emerging Bayesian epistemic perspective on argumentation (Hahn and Oaksford, Psychol Rev 114:704–732, 2007), in which people are argued to be sensitive to the reliability of the source of an argument. The experiment varied source reliability, initial degree of belief in the conclusion, and whether the ad hominem was a pro or a con argument. A Bayesian account of the effect of reliability on posterior degree of beliefs after hearing the argument provided excellent fits to the data. Moreover, the results were not consistent with the pragma-dialectic approach, as no differences were observed between conditions where a discussion rule was violated and a control condition where it was not violated. However, further experimentation is required to fully establish this conclusion.

Much of what we know of the world comes to us second-hand (e.g. Coady 1992). We actually pick up few of our beliefs via direct experience. Rather most of our beliefs derive from being told in the classroom, on the TV, in the newspapers, in books, by our parents, by our bosses, or by our friends. Whether or to what degree

M. Oaksford (✉)
Department of Psychological Science, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK
e-mail: mike.oaksford@bbk.ac.uk

U. Hahn
Department of Psychological Science, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

School of Psychology, Cardiff University, Cardiff, Wales, UK

F. Zenker (ed.), *Bayesian Argumentation: The practical side of probability*, Synthese Library 362, DOI 10.1007/978-94-007-5357-0_3,

we believe what we are told depends on many factors. For example, it depends on the coherence of what we are told with other beliefs and information (see, e.g. Olsson 2005; Shogenji 1999; Bovens and Hartmann 2003; Harris and Hahn 2009). However, it also depends critically on the source of the information. We are less likely to believe a proposition on being told that our informant was a sociopath, and we are more likely to believe experts than novices. It is perhaps not surprising therefore that our standards of argumentation embody some assessment of the credibility of an informant. For example, (1) seems a reasonable argument:

(1) Person A suggests that person B should invest in A's company.
Person C argues against A's suggestion by pointing out that A is a sociopath.

One might argue that whether A is a sociopath or not is irrelevant to whether investing in A's company will yield B a profit. However, human beings are in what has been called the "finitary predicament" (Cherniak 1986); they most often simply do not have the time or access to information to make more informed decisions. In financial terms, the man in the street is most often an "outsider" rather than an "insider". Without insider knowledge, C's interjection that A is a sociopath, and hence likely to act solely in her own interests rather than B's is highly relevant to B's decision whether or not to invest in A's company. Had more people identified Bernie Madoff as a sociopath, fewer may have lost their shirts investing with him in financial deals they never looked into further.

A problem for argument (1), which is seemingly against the person rather than the proposition they are advancing, is that it has traditionally been viewed as fallacious: it is an instance of the classical reasoning fallacy, the *argumentum ad hominem*. For example, what distinguishes (1) from (2):

(2) Person A (B's doctor) suggests that person B should give up smoking.
Person C argues against A's suggestion by pointing out that B smokes herself.

(2) seems clearly fallacious as whether B's doctor smokes or not cannot influence the beneficial health effects that B will experience by giving up smoking. Here, the reliability of the source of this advice is not only irrelevant but known to be irrelevant. That is, in this case, we do have insider information. It is now common knowledge that smoking is bad for your health in a way that it was not common knowledge that investing in Bernie Madoff's company would lead you to lose money. This epistemic difference seems to be important for how convinced people are by an ad hominem argument. That is, the extent to which B now believes that she should give up smoking or invest in A's company, given C's ad hominem argument, depends in part on what they already believe about these propositions.

An epistemic approach to reasoning fallacies, that may be able to capture these intuitions, has recently been adopted by a variety of researchers. For example, Ikuenobe (2004) argued that "a fallacy is fundamentally an epistemic error, involving the failure to provide in the form of a premise, adequate proof for a belief, or proposition in the form of a conclusion" (p. 193). However, he avoided articulating any explicit epistemic principles about what constitutes adequate proof or evidence. Hahn and Oaksford (2006, 2007) adopt a related approach arguing that

when the relevant arguments are reformulated with the conclusion as the hypothesis and the premise as the evidence, then Bayes' theorem provides a useful account of the epistemic adequacy of proof or what they call *argument strength*. There has been very little empirical research on argumentative fallacies. This is particularly true of the argumentum ad hominem, where other than the work summarised by van Eemeren et al. (2009), no other empirical studies appear to have been carried out on this fallacy. In this chapter, we report the results of an experiment investigating the effects on people's degree of belief in the conclusion of an ad hominem argument. We will conclude that a Bayesian source reliability model (Hahn et al. 2009) provides a good account of people's judgements of how persuaded they are by these arguments, which do not seem to be particularly sensitive to violations of *pragma-dialectical* discussion rules (see, *The Pragma-Dialectical Approach*, and van Eemeren et al. 1996, 2009).

We begin this chapter with a review of the types of the argument ad hominem that have been identified in the literature. We then introduce the pragma-dialectical approach to the argument ad hominem and the experimental research that has been done on how *reasonable* these arguments are perceived to be (van Eemeren, et al. 2009). We then introduce the Bayesian approach to argumentative fallacies and derive some predictions for an experiment. We then describe the experiment and the results. In the discussion, we outline the conclusion that we draw from this research.

1 Types of the Argumentum Ad Hominem

Three types of the argumentum ad hominem have been identified in the literature. These are the *abusive*, the *circumstantial*, and the *tu quoque* ("you, too") forms. According to Walton (2000, 2009), all forms of the argument ad hominem attempt to undermine a person's argument for a proposition by suggesting that the person is of "bad character". As Walton observes, this means that these arguments can have non-fallacious uses, "the real function of an *ad hominem* argument (when properly used) is to attack an arguer's credibility in order to criticise the argument she advocates", Walton (2000, p. 102).

The form of the abusive argument ad hominem is as follows (from Walton 2000):

(3) Person A is of bad character.
Therefore, A's argument X should not be accepted.

An example of this type of argument is given in (3′):

(3′) Person A knows nothing about cars.
Therefore, A's argument that you should buy a Ford should not be accepted.

This is the same form as (1) above. Accusing someone of lack of knowledge or of being a sociopath without further substantiation is potentially abusive. This is

also known as the *direct* form of the argument. As we will see, according to many theorists (van Eemeren, et al. 2009; Walton 2000, 2009), whether this use of the argument ad hominem is fallacious depends on the stage of argumentation and the type of argumentation in which our interlocutors are engaged.

The form of the circumstantial, or *indirect*, argument ad hominem is as follows (from Walton 2000):

(4) A advocates argument X.
A has carried out an action, or set of actions, that imply that A is personally committed to the opposite of X.
Therefore A is a bad person.
Therefore A's argument X should not be accepted.

An example of this type of argument is given in (4′):

(4′) A advocates using company C because it is the best.
A has married the daughter of the CEO of C and so may not be personally committed to C being the best.
Therefore, A is a bad person.
Therefore, A's argument for employing C should not be accepted.

In this type of argument ad hominen, a circumstantial inconsistency is pointed out, that is, there is circumstantial evidence that A would advocate this argument even if he didn't personally believe it.

The *tu quoque* type of the argument ad hominem is a special case of the circumstantial argument type in which the circumstantial inconsistency involves A (not) carrying out exactly the actions that she is advocating people should (not) carry out. This is identical to argument (2) above:

(5) A advocates not smoking.
A smokes herself and so cannot be personally committed to not smoking.
Therefore A is a bad person.
Therefore A's argument for not smoking should not be accepted.

We now look at how these fallacies have been dealt with in the pragma-dialectical approach to argumentation.

2 The Pragma-Dialectical Approach

In their *pragma-dialectical* approach, van Eemeren and Grootendorst (2004) developed a normative theory of discourse rules that define the illegitimate moves that the participants in an argument can make at different stages and in different types of arguments. Some moves may be fallacious in the context of one type of argument but not in another. For example, in a *quarrel*, "where each participant tries to hit out verbally at the other . . . [and which is] characterised by an almost total absence of logical reasoning and by heightened emotions" (Walton 1990, p. 414), arguing ad

hominem may be appropriate. However, in a *critical discussion*, in which the goal is to "resolve a difference of opinion by means of methodical exchange of discussion moves" (van Eemeren and Grootendorst 2004, p. 22), arguing ad hominem might not be appropriate. Following the pragma-dialectical approach in this regard, we concentrate here solely on the critical discussion in which some kind of rational standard seems to be required.

In the pragma-dialectical approach, fallacies arise because the participants in an argument make wrong moves, that is, moves not licensed by the rules of discourse. Whether the ad hominem argument is fallaciously used in a critical discussion depends on the stage to which the argument has progressed. This argument may be used perfectly legitimately at the *argumentative* stage of a critical discussion if one party has made an appeal to authority in order to support a standpoint that they have introduced earlier at the *confrontation* stage. So, for example, A may propose that smoking is bad for you, and B may enquire why A believes this. At this point, A may argue that his doctor informed him that smoking is bad for you, that is, A makes an appeal to authority to backup his claim introduced at the confrontation stage. A and B have now entered the argumentative stage of the critical discussion, at which point it is legitimate for B to question A's appeal to authority by pointing out that A's doctor himself smokes which seems to belie his ability to act as an authoritative source. However, in (5), where A is the doctor introducing this argument himself at the confrontation stage, it would not be legitimate to disallow his introduction of this claim by deploying the *tu quoque* ad hominem argument. For van Eemeren et al. (2009, p. 21), deploying the ad hominem argument at the confrontation stage constitutes a violation of the *freedom rule* that "discussants may not prevent each other from advancing standpoints or from calling standpoints into question". This is because the goal of a critical discussion is, as we stated above, to "resolve a difference of opinion by means of methodical exchange of discussion moves". If the doctor's introduction of the claim that smoking is bad for you is dismissed at the confrontation stage simply because he himself smokes then there is no possibility of moving forward in resolving any difference of opinion between the doctor and his interlocutor.

Van Eemeren et al. (2009) have also investigated people's attitudes to the argument ad hominem in a series of experiments. These experiments used all three types of the argument ad hominem which were always introduced in the confrontation stage of a critical discussion. Consequently, the arguments were always deployed such that their use was fallacious and constituted a violation of the freedom rule. A control condition was also used that did not violate the freedom rule. General conversational pragmatics also suggests that within the different types of argument ad hominem there will be differences in how reasonable each is viewed. Van Eemeren et al. (2009) hypothesise that general issues of politeness indicate that the abusive form will be less reasonable than the circumstantial and the circumstantial less reasonable than the tu quoque. In order to rule out any general politeness explanation of the results, van Eemeren et al. (2009) also used three different discussion contexts, a domestic discussion, a political debate, and scientific discussion. The important distinction is between the scientific context and the

remaining two. One would expect people to be particularly sensitive to violations of discussion moves in scientific discussion as this is one of the paradigmatic examples of a critical discussion where the goal is to get things right. Thus, if people are sensitive to the reasonableness or soundness of an argument then one would expect them to consider fallacies in a scientific discussion to be less reasonable than fallacies in the other two discussion contexts.

As the discussion in the last paragraph indicates, the dependent variable used in van Eemeren et al.'s (2009) experiments was a rating of *reasonableness*. The idea behind this variable was to—as far as possible—address people's understanding of the discourse rules governing argumentation in a critical discussion. That is, by analogy with logic, van Eemeren et al. were attempting to assess the soundness of the discussion rules regardless of specific content used. They suggest that differences in reasonableness between discussion contexts do not mean that different norms apply in each context but rather that the criteria for the application of the same norms may vary from context to context.

Briefly summarising their results, they found that all the arguments ad hominem implying a violation of the freedom rule were judged less reasonable than the control condition where there was no violation of this rule. They also found that the individual arguments differed in reasonableness in the order predicted, that is, the abusive form was judged less reasonable than the circumstantial and the circumstantial was judged less reasonable than the *tu quoque*. Moreover, overall, the use of any of these argument forms was judged less reasonable in the scientific discussion context than the other two contexts, and their use was judged equally reasonable in the domestic and political contexts. In sum, van Eemeren et al. (2009) argue that people are sensitive to violations of the freedom rule and to the pragmatics of politeness but that the observed differences between discussion contexts rule out a general politeness explanation of their results.

3 The Bayesian Approach

In this section, we briefly introduce the Bayesian approach to argumentation and where it contrasts with the pragma-dialectical approach. In the latter approach, fallacies arise because the participants in an argument make wrong moves, that is, moves not licensed by the rules of discourse. For example, here are two arguments ad hominem cited in van Eemeren et al. (2009, p. 56) but deriving from Brinton (1995):

(6) Candidate Jones has no right to moralise about the family; he was once seen arguing with his wife.
(7) Candidate Jones has no right to moralise about the family; since he cheats on his wife.

(6) seems much less acceptable than (7). This difference cannot reside in the discourse context which is the same in both cases, a political discussion, nor in the type of ad hominem argument as they are both the direct or *abusive* form of this

argument. Consequently, for the pragma-dialectical approach, the difference between these arguments seems to have to reside in the type of the argumentative discourse in which people are engaged and in the stage of the argument (van Eemeren and Grootendorst 2004). So, in (7) a rule must have been violated, but in (6) no rule has been violated.

Our Bayesian account begins from the observation that pairs of arguments like (6) and (7) would seem to be differentially acceptable even in the same argumentative context. So assuming a critical discussion, they could both be used fallaciously in the confrontation stage violating the freedom rule, but nonetheless it could be argued that (7) is more convincing than (6). Moreover, both would be acceptable as refutations of Candidate Jones' authority to moralise about the family in the argumentative stage. But again, (7) would be more convincing than (6). Thus, it seems perfectly feasible for both (6) and (7) to occur in the same argumentative context, for example, a critical discussion, and at the same stage but (7) would still be more convincing than would (6). According to the Bayesian theory, the difference must be due to the difference in the content of the argument, which is analysed using Bayesian probability theory to provide an account of *argument strength*. Thus, the approach attempts to capture the uncertain nature of argumentation, emphasised by previous researchers (e.g. Perelman and Olbrechts-Tyteca 1969), while also providing a normative standard, as emphasised in the pragma-dialectical approach.

Before discussing the technicalities of the Bayesian approach, it is worth contrasting it with van Eemeren et al. (2009) by quoting them in detail here. They are commenting on the contrast between their approach and the large body of empirical work on *persuasion* in social psychology (for a recent review, see, e.g. Johnson et al. 2005; for a detailed treatment, see, e.g. Eagly and Chaiken 1993):

> Although it may by and large be expected that there will be some connection between the persuasiveness of argumentation and the reasonableness of argumentation (who would let themselves be convinced by unreasonable argumentation?), it should nevertheless be clear that the contents of the two terms, *persuasiveness* and *reasonableness*, do not coincide. Sound, or in other words reasonable, argumentation does not have to be perceived as convincing *per se*. Other more psychologically tinted factors such as someone's original attitude regarding the defended standpoint, the credibility of the source, the involvement of whoever must be convinced of the defended standpoint, and so on, play a part in convincing that is not to be underestimated. (Van Eemeren et al. 2009, pp. 32–33)

Our Bayesian approach attempts to make sense of how persuaded or convinced somebody is by an argument by providing a *rational analysis* of argumentation in Anderson's (1990) sense.

According to Anderson (1990), rational analysis requires six steps:

Step 1. Specify precisely the goals of the cognitive system.
Step 2. Develop a formal model of the environment to which the system is adapted.
Step 3. Make minimal assumptions about computational limitations.
Step 4. Derive the optimal behaviour function given 1–3 above. (This requires formal analysis using rational norms, such as probability theory and decision theory.)
Step 5. Examine the empirical evidence to see whether the predictions of the behaviour function are confirmed.
Step 6. Repeat, iteratively refining the theory.

Paraphrasing van Eemeren et al. (1996, p. 5), the goal of argumentation is to increase (or decrease) "the acceptability of a controversial standpoint for a listener or reader, by putting forward a constellation of propositions intended to justify (or refute) the standpoint before a 'rational judge'" (step 1). The environment is given in part by the audience, its prior beliefs about a subject and the *rational judge* constraint (step 2). For the moment, we leave step 3 to one side (perhaps to be incorporated on iterating the process at stage 6). To provide step 4, we employ Bayesian probability theory to capture how far someone's prior degree of belief in a controversial standpoint should be modified by propositions advanced for or against it. The experiment we report later in this chapter initiates step 5.

Our approach can be seen as pursuing complimentary goals to both the empirical research on persuasion and the pragma-dialectical approach. On the one hand, we provide a much needed evaluative account of argument persuasiveness in social psychology (Voss and Van Dyke 2001); on the other hand, we supplement a normative account of the reasonableness of argument with a normative account of how convincing people should find these arguments.

The Bayesian approach uses Bayes' theorem and the subjective or epistemic interpretation of probability (Gillies 2000) to capture, "someone's original attitude regarding the defended standpoint" and "the credibility of the source". Bayes' theorem is as follows:

$$P(C|a) = \frac{P(a|C)P(C)}{P(a|C)P(C) + P(a|\neg C)P(\neg C)} \tag{8}$$

Bayes' theorem states that one's posterior degree of belief in a conclusion, C, in light of an argument, a, $P(C|a)$, is a function of one's initial, prior degree of belief, $P(C)$ (i.e. "someone's. . .original attitude"), and how likely it is that the argument one is presented with is true if one's initial conclusion was true, $P(a|C)$, as opposed to if it was false, $P(a|\neg C)$. The ratio of these latter two quantities, the likelihood ratio, provides a natural measure of the diagnosticity of the argument—that is, its informativeness regarding the conclusion in question. The most basic aspect of diagnosticity is that if $P(a|C) > P(a|\neg C)$, then the argument will result in an increase in belief in C, whereas if $P(a|C) < P(a|\neg C)$, then the argument will result in a decrease. This has immediate implications for arguments from different sources. We can imagine an encounter with someone who we believe to be truthful as opposed to someone who we believe to be a liar: information from the truthful source will increase our belief in the conclusion, whereas we will consider the opposite of what the liar says to be more likely to be true.

We adopt the same approach to source reliability as Hahn et al. (2009; see also Bovens and Hartmann 2003). The argument, a, above is always a report by one of the interlocuters of the evidence that they take to support or refute C. Factoring in the reliability, R, of the source can be achieved by decomposing the likelihoods to incorporate R in the marginals, for example,

$$P(a|C) = P(a|C,R)P(R) + P(a|C,\neg R)P(\neg R) \tag{9}$$

(and similarly for $P(a|\neg C)$). An unreliable informant is assumed to be as likely to emit argument a as $\neg a$ whether the conclusion is true or false, that is. $P(a|C,\neg R) = P(a|\neg C,\neg R) = .5$. This decomposition allows the contributions of the argument to altering degrees of belief to be separated out from source reliability. Hahn et al. (2009) showed that because of the multiplicative relation between the probability that the source is reliable, $P(R)$, and the likelihoods, source reliability places limits on the posterior degree of belief. They did this by varying the likelihood ratio due solely to the argument, that is, $P(a|C,R)/P(a|\neg C,R)$, and the probability that the source is reliable, $P(R)$. For example, for a completely reliable source, $P(R) = 1$, assuming an uninformative prior, $P(C) = .5$, as the argument likelihood ratio increases the posterior, $P(C|a)$, approaches 1. However, with $P(R) = .6$, the posterior asymptotes at $P(C|a) = .8$ as the argument likelihood ratio increases.[1]

This analysis captures our initial intuition that the reliability of an informant and prior knowledge interact such that the less knowledge you have the greater effect the ad hominem argument should have. The argument ad hominem is a counter argument which should dissuade someone from believing a conclusion given a prior argument. Prior to the ad hominem attack, a general *principle of charity* constrains an audience to view an interlocuter's utterances as making true statements (Davidson 1974; Wilson 1959). The ad hominem argument is an attack on this assumption to undermine the conclusion. In (1), for example, before hearing C's ad hominem argument, B *assumes* A is reliable, that is, $P(R) \approx 1$, and even though he is initially uncertain, $P(C) = .5$, he should now be highly confident that he should invest in A's company, $P_1(C|a) = .995$. However, C then informs him that the reliability assumption is wrong and that A is unreliable, for example, $P(R) = .5$, and so, re-computing, this reduces B's confidence so that $P_2(C|a) = .75$ a reduction of .245. Of course, if the source is assumed to be completely unreliable, $P(R) = 0$, there will be no change in degree of belief, that is, $P(C) = P_1(C|a)$. However, if B has some prior knowledge that inclines him to believe the conclusion already, as in (2), for example, $P(C) = .8$, the assumption that his doctor is reliable, that is, $P(R) \approx 1$, should lead to the conclusion that she should give up smoking with even higher confidence, $P_1(C|a) = .999$. Moreover, learning that her doctor smokes and so may be unreliable will not reduce B's degree of belief by much, so with $P(R) = .5$, $P_2(C|a) = .92$, that is, a reduction of only .076. Thus, by incorporating prior beliefs a Bayesian model seems to explain the epistemic differences between examples (1) and (2). Figure 1 illustrates this behaviour and also shows what happens if the argument likelihoods are not deterministic, that is, $P(a|C,R) = 1$ and $P(a|\neg C,R) = 0$. Figure 1 shows the models behaviour when $P(a|C,R) = .8$ and $P(a|\neg C,R) = .2$, that is, the argument likelihood ratio is 4.

[1] In this example and the next, a deterministic relationship is assumed between the conclusion and the argument, such that the conclusion guarantees the argument with probability 1 if true and probability zero if false.

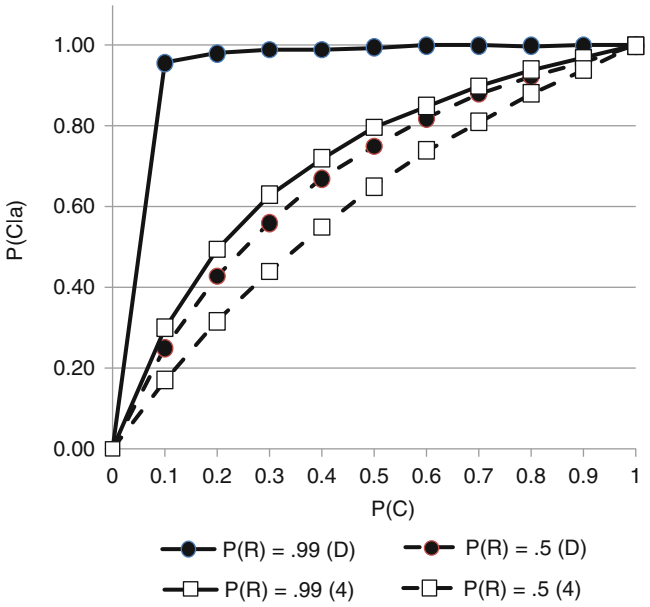

Fig. 1 The behaviour of the model showing changes in the posterior, $P(C|a)$, varying the prior, $P(C)$, with two different argument likelihoods (D = deterministic, 4), and two reliabilities, $P(R)$ (.99, .50)

4 An Experiment on the Argument Ad Hominem

We conducted an experiment using the argument ad hominem, but in contrast to van Eemeren et al. (2009), we asked participants to rate how convinced one of the interlocutors should be by these arguments in a critical discussion. The materials we used were derived from van Eemeren et al.'s (2009) study. However, we also derived *pro* and *con* versions of these arguments. The con version is the standard usage of the ad hominen argument which we illustrate using the *abusive* form used in the experiment:

(10) A: After listening to him, I think it might be possible that Ford cars simply drive better.
B: Actually, you should be certain that they don't drive better.
A: Why do you think that?
B: Because how would he know? He doesn't know the first thing about cars.

Initial degree of belief ($P_1(C|a)$), that is, "after listening to him", was partially manipulated by replacing the phrase "I think it might be possible" (weak) with "I strongly believe" (strong) in the opening comment. Person A always expresses an initial opinion based on hearing an argument from a third party in support of the claim that, "Ford cars simply drive better". Creating a pro argument involved a simple change in the opening comment:

(11) A: Even after listening to him, I think it might be possible that Ford cars don't drive any better.
B: Well, you can be certain that they don't drive better.
A: Why do you think that?
B: Because how would he know? He doesn't know the first thing about cars.

Here, B's ad hominem argument against the third party is in the same direction as A's opening claim because A still doesn't believe it despite the argument. We took two measures in this experiment. After B's final ad hominem attack, participants were asked how convinced A should now be in the claim that "Ford cars *simply* (don't) drive (any) better" on a 0–100 scale. That is, we measured $P_2(C|a)$ for A after B argues that $P(R) < 1$. In a second phase of the experiment, we also measured participants' assessment of A's degree of belief in their opening comment ($P_1(C|a)$). This comment is of course conditioned on having heard an argument by a third party supporting the claim that "Ford cars simply drive better". That is, we measured $P_1(C|a)$, that is, before A had any reason to believe that $P(R) < 1$, that is, when A, by the principle of charity, should assume that $P(R) \approx 1$. The measurement was taken on the same 0–100 scale.

In these dialogues, the argument put forward by the third party remains fixed. Moreover, it is assumed that that reliability remains fixed across types of ad hominem argument at least when the principle of charity holds. That is, when participants judge A's opening claim for $P_1(C|a)$ *before* B argues that $P(R) < 1$. With all elements of the likelihood ratio fixed the only way the posterior, $P_1(C|a)$, can vary is if the prior, $P(C)$, varies. Let us consider Fig. 1 for the $P(R) = .99$, argument likelihood ratio (LR) $= 4$ case. When $P_1(C|a) = .8$, which might be the case when the conclusion is strongly believed based on the con argument materials, $P(C) = .5$. If, when presented with the con argument where the conclusion might only be possible, $P_1(C|a)$ falls to .63, then, assuming everything else remains the same, $P(C)$ must fall to .3. So, because in the materials everything else remains fixed, differences in $P_1(C|a)$, that is, *before* B argues that $P(R) < 1$, must equate to changes in A's prior degree of belief, $P(C)$. In fitting the model to our data, we will use the prior ratings of $P(C|a)$, when $P(R) \approx 1$, to calculate $P(C)$, which we will assume is updated, on learning that $P(R) < 1$.

This analysis is also consistent with intuition. Compare the following intra-individual case:

(12) A After listening to him, I think it might be possible that Ford cars simply drive better.
A After listening to him, I strongly believe that BMW cars simply drive better.

Assuming A has heard exactly the same argument in both cases for why Fords/BMW drive better than, say Saabs, and assuming a principle of charity, it seems intuitively clear that A must have believed that BMWs drive better than Fords before hearing the argument put forward by this third party. Moreover, the following interindividual case seems intuitively to prompt the same explanation:

(13) A After listening to him, I think it might be possible that Ford cars simply drive better.
A′ After listening to him, I strongly believe that Ford cars simply drive better.

If A and A′ have heard exactly the same argument in both cases and assuming a principle of charity, it is intuitively clear that A′ must have believed that Fords drive better more than A before they heard the argument put forward by this third party. Example (12) is more convincing than (13) because A and A′ might differ in terms of gullibility, which is better modelled by differences in the argument likelihoods they assign to the same argument.

The principle quantitative prediction for this experiment is that the results can be shown to conform to the pattern in Fig. 1, making reasonable assignments of values to the parameters of the model. This hypothesis creates further predictions. So, in all cases, the judgement of convincingness taken in the second phase of the experiment, corresponding to $P_1(C|a)$ when $P(R) \approx 1$ (which we refer to as the 1st posterior), should be higher than when taken in the first phase, corresponding to $P_2(C|a)$ when $P(R) < 1$ (which we refer to as the 2nd posterior). Con arguments, where the initial claim is affirmative (e.g. "Ford cars drive better"), should have higher convincingness than pro arguments where the initial claim is negative (e.g. "Ford cars don't drive better"). Furthermore, when A's initial claim is strong ("I strongly believe"), convincingness should be higher than when this claim is weak ("I think it might be possible"). However, differences for weak or strong initial claims are only likely to come out in interactions. For example, a weak pro argument should come out just below .5, a weak con argument just above .5; but a strong pro argument should come out well below .5 and strong con well above. But this means that the mean of the weak and strong arguments would be predicted to be about the same, that is, .5 approximately. So while an interaction with pro vs. con arguments is predicted, there should be no main effect of weak vs. strong initial claims. There are other hypotheses about which we are neutral that can be tested. For example, if judgements of how convinced people are by an argument show some role for soundness, as investigated by van Eemeren et al. (2009) using the reasonableness measure, then differences between the types of the argument ad hominem would be predicted. That is, the *tu quoque* would be judged more convincing than the *circumstantial*, and the *circumstantial* would be judged more convincing than the *abusive*. Moreover, if people are sensitive to violations of the freedom rule in judgements of how convincing they find an argument, then convincingness should be lower for all the ad hominem argument types than for the control condition.

5 Method

Participants. Thirty two participants were recruited from undergraduate psychology students at Birkbeck College, University of London. Twenty one of the participants had English as their first language. The remainder were also English speakers and

readers. The participants were unpaid and not familiar with the materials used in the study.

Design. The experiment was a $2 \times 4 \times 2 \times 2$ completely within subjects design with posterior judgement (1st vs. 2nd posterior), Ad hominen fallacy type (abusive, circumstantial, tu quoque, and a non-fallacious control), initial claim strength (weak vs. strong), and argument direction (pro vs. con) as independent variables and convincingness ratings (0–100) as the dependent variable.

Materials. The materials were contained in a two-part booklet. Part 1 established estimates of people's 2nd Posterior degrees of belief in the conclusion. Part 2 established estimates of people's 1st posterior degrees of belief in the conclusion. Part 1 had 17 pages with instructions and two practice dialogues on the first page. The 16 experimental dialogues were arranged in random orders and displayed one per page on the subsequent 16 pages. The 16 dialogues comprised a set of 4 dialogues for each of the 4 fallacy levels. Each dialogue was composed of a four-line fictional exchange between two interlocutors, A and B. Within each fallacy dialogue set, interlocutor B presents an identical argument for each of the four dialogues, for example, "Because how would he know? He doesn't know the first thing about cars". However, the position of interlocutor A changes across the four dialogues as the argument direction, pro vs. con., and the strength of A's initial were manipulated, for example, "After listening to him, I *think it's possible* (strongly believe) that Ford cars *simply* (don't) drive better". See the Appendix for the materials for all four dialogues including the control.

The reason for placing the directional manipulations in A's opening statement rather than in B's arguments is that *ad hominem* arguments by definition always "damn the source". It was therefore simpler and required less manipulation of the remaining dialogue content to hold B's *ad hominem* argument as unidirectional. After each dialogue, participants were instructed to rate how convinced they felt that A should now be in their opening statement having listened to the argument present by B. The rating was on a scale of 0% (not convinced at all) to 100% (completely convinced).

Part 2 contained a prior belief questionnaire. Two practice examples were followed by the 16 opening statements made by interlocutor A in each of the part 1 dialogues. Participants were asked to provide ratings on the same scale as above of how strong they thought A's belief was in each of their opening statements. The remainder of part 2 contained questions about the participants followed by a short-written debrief explaining the purpose of the experiment.

Procedure. Participants were tested individually under laboratory conditions and were not timed. They were initially given a consent form to be signed and an information sheet containing details of the purpose of the project. Participants were then given the two-part booklet and instructed to work through the dialogues in part 1 in order, providing an argument strength estimate for each dialogue before moving onto the next one. Once the dialogues in part 1 were completed, the participants were instructed to complete the initial questionnaire and the final questions in part 2 before being debriefed.

6 Results and Discussion

The results of the experiment are shown in Table 1. The raw data using the 0–100 ratings were first transformed into the 0–1 probability scale by dividing by 100. For the pro arguments, this gave an estimate of $P(\neg C|a)$ and so these were further transformed by subtracting from one to give an estimate of $P(C|a)$ which acted as the dependent variable. Before assessing the overall fit of the Bayesian model to these data, we first statistically assessed the other hypotheses outlined in the introduction to this experiment. We conducted a $2 \times 4 \times 2 \times 2$ ANOVA with posterior judgement (1st vs. 2nd posterior), ad hominem fallacy (abusive, circumstantial, tu quoque, and a non-fallacious control), initial claim strength (weak vs. strong), and argument direction (pro vs. con) as within subjects factors and with $P(C/\text{a})$ as the dependent variable.

As predicted, there was a main effect of posterior judgement such that 2nd posterior, $P_2(C|a)$ (mean = .37, SE = .016), corresponding to $P(C|a)$ when $P(R) < 1$, was significantly lower than the 1st posterior, $P_1(C|a)$ (mean = .50, SE = .009), corresponding to $P(C|a)$ when $P(R) \approx 1$, $F(1, 31) = 81.88$, $MSe = .06$, $\eta^2 = .73$, $p < .0001$). Again as predicted, con arguments (mean = .53, SE = .025), where the initial claim is affirmative (e.g. "Ford cars drive better"), had higher $P(C|a)$ values than pro arguments (mean = .34, SE = .013), where the initial claim is negative (e.g. "Ford cars don't drive better"), $F(1, 31) = 33.62$, $MSe = .30$, $\eta^2 = .52$, $p < .0001$). There was no significant main effect of initial claim strength, $F(1, 31) = 1.40$, $MSe = .03$, $\eta^2 = .04$, $p = .25$). However, this variable, as predicted, interacted strongly with other factors. As discussed in the introduction to this experiment, it interacted strongly with pro vs. con arguments. The difference between pro and con arguments was far greater when the initial claim was strong (pro: mean = .27, SE = .003; con: mean = .61, SE = .017) than when it was weak (pro: mean = .41, SE = .025; con: mean = .45, SE = .015), $F(1, 31) = 95.61$, $MSe = .06$, $\eta^2 = .76$, $p < .0001$). It interacted much more weakly with posterior judgement. The difference between 1st and 2nd posterior was greater when the initial claim was weak (1st posterior: mean = .50, SE = .009; 2nd posterior: mean = .35, SE = .016) than when it was strong (1st posterior: mean = .50, SE = .011; 2nd posterior: mean = .38, SE = .019), $F(1, 31) = 4.54$, $MSe = .01$, $\eta^2 = .13$, $p < .05$. There were a variety of other weak three-way interactions, which mainly involved the type of ad hominem argument.

There was no main effect of ad hominem argument type, $F(3, 93) = 1.10$, neither were any of the Helmert contrasts significant, that is, control vs. *tu quoque*, the control vs. *tu quoque* and circumstantial collapsed, and the control vs. *tu quoque*, circumstantial and the abusive collapsed . In particular, the comparison between the control and the fallacies collapsed was not significant, $F(1, 31) < 1$. Ad hominem argument type did show a weak two way interaction with strength of initial claim, but it was further modified by a three-way interaction with pro vs. con arguments, $F(3, 93) = 5.42$, $MSe = .01$, $\eta^2 = .15$, $p < .005$. To explore these possible effects further, we reasoned that it is the comparative effects of the

Table 1 Results of the experiment showing the mean $P(C|a)$ values (SDs) for initial claim strength (weak or strong), pro vs. con argument, 1st vs. 2nd posterior (post) judgement, for each fallacy type and the control

	Weak claim				Strong claim			
	Pro argument		Con argument		Pro argument		Con argument	
Fallacy	1st post	2nd post	1st post	2nd post	1st post	2nd post	1st post	2nd post
Abusive	.48(.28)	.38(.22)	.52(.14)	.39(.21)	.34(.22)	.32(.26)	.66(.17)	.51(.21)
Circumstantial	.47(.21)	.30(.24)	.54(.19)	.35(.16)	.32(.20)	.22(.22)	.68(.20)	.51(.18)
Tu quoque	.51(.15)	.35(.17)	.51(.15)	.40(.20)	.23(.19)	.21(.21)	.74(.19)	.51(.16)
Control	.44(.20)	.31(.23)	.56(.17)	.35(.19)	.26(.23)	.23(.29)	.77(.16)	.54(.23)
Totals (SE)	.48(.03)	.34(.03)	.53(.02)	.37(.02)	.29(.03)	.25(.04)	.71(.03)	.52(.02)

different argument forms on changing degree of belief that are important to determining whether differences in the soundness of these forms affects how convincing they are as well as how reasonable. Interactions with pro vs. con arguments or the strength of the initial claim, while important for testing the Bayesian model, are tangential to whether these different argument forms have differential effects. We therefore investigated the difference between argument types by looking at 2nd posterior ratings alone summed across these other variables and the 1st posterior–2nd posterior difference scores aggregated in the same way.

Looking solely at the 2nd posterior ratings, there was a significant effect of ad hominem argument type, $F(3, 93) = 2.54$, $MSe = .01$, $\eta^2 = .08$, $p < .05$ (one-tailed). Reverse Helmerts contrasts showed that the abusive argument form had higher mean $P(C|a)$ than the other argument types collapsed, $F(1, 31) = 5.74$, $MSe = .01$, $\eta^2 = .16$, $p < .025$, and that there were no significant difference between circumstantial and tu quoque collapsed with the control, nor between tu quoque and the control. Thus, the abusive form does not lead to as low values of $P(C|a)$ as the other ad hominem argument types, and there are no differences between these other types and the control. We also analysed the 1st posterior–2nd posterior difference scores. There was a significant effect of ad hominem argument type, $F(3, 93) = 2.37$, $MSe = .01$, $\eta^2 = .07$, $p < .05$ (one-tailed). Replicating the analysis for the 2nd posterior rating alone, reverse Helmerts contrasts showed that the abusive argument form lead to a lower difference in $P(C|a)$ than the other argument types collapsed, $F(1, 31) = 6.34$, $MSe = .01$, $\eta^2 = .17$, $p < .01$, and that there were no significant differences between circumstantial and tu quoque collapsed with the control, nor between tu quoque and the control. We also looked at all pairwise comparisons using paired *t*-tests. Only the comparisons with the abusive form approached significance but even these were not significant once an appropriate Bonferroni correction for multiple tests was applied.

These results would appear to indicate that, in contradistinction to van Eemeren et al's. (2009) predictions, how convincing we find the argument ad hominem is not affected by the soundness of these arguments. People appear as convinced by the argument ad hominem when the freedom rule is violated (*circumstantial*, *tu quoque*) than when it is not (*control*). It could be argued that this result arises

because of the way we have set up these arguments means that the participants interpret the two interlocuters as in the argumentative stage and not in the confrontational stage. In the argumentative stage, the freedom rule is not violated. One might argue that both A and B have heard C produce her argument for or against the conclusion and so for both of them the argument has been admitted as a legitimate topic of debate. However, presumably an argument only begins when two interlocuters disagree on a matter. Even for the pro argument, A and B differ in the degree to which the conclusion should be believed. Despite C's argument, A may still strongly believe the conclusion, but B's ad hominem argument is designed to take A to being totally convinced. The first two lines of each dialogue set up, this disagreement and hence must be interpreted as at the confrontational and not at the argumentative stage. Moreover, if all these arguments are at the argumentative stage, why is the abusive type less convincing as no pragma-dialectical rule has been violated (although post hoc *t*-tests did not support this difference). So there is good reason to argue that these dialogues are indeed in the confrontational stage and that despite the consequent violation of the freedom rule, two versions of the argument ad hominem were as convincing as a control even though they violated the freedom rule.

A possible explanation for the lack of difference between ad hominem argument types is that we have used too few exemplars and that those we have used, as it happens, are all similar in reasonableness because they cross cut discourse context. We used van Eemeren et al's (2009, Table 3.3, p. 68) domestic context abusive type (mean reasonableness: 3.29), their political context circumstantial type (mean reasonableness: 4.19), and their scientific context *tu quoque* type (mean reasonableness: 3.66). The *tu quoque* is only rated low in reasonableness in this context, which might explain the lack of differences. We would need to introduce more exemplars to truly rule out any difference in convincingness between the circumstantial and the *tu quoque*. However, the differences in reasonableness between the arguments we used predicts a linear trend such that *circumstantial* > *tu quoque* > *abusive* if soundness also affects how persuasive people find these arguments. And ignoring the control, this linear trend was significant in our data, $F(1, 31) = 5.82$, $MSe = .01$, $\eta^2 = .16$, $p < .025$, providing some evidence that how persuaded people are by these arguments is sensitive to their perceived reasonableness in different contexts. Nonetheless, the principle result of no differences with the control remains, which indicates that these judgements may not be sensitive to violations of the freedom rule.

We now turn to modelling the overall pattern of results using the Bayesian source reliability model. Given the very minimal differences between the types of ad hominem argument and the lack of differences between these argument types and the control, we modelled the aggregate data shown in the final row of Table 1. There are quite a number of parameters that could vary to model this data but to guard against over fitting we chose to fix most of these to the values used to illustrate the model in Fig. 1. The argument likelihood was set to 4 ($P(a|C, R) = .8$; $P(a|\neg C,R) = .2$); $P(a|C,\neg R)$ and $P(a|\neg C,\neg R)$ were set to .5; and $P(R)$

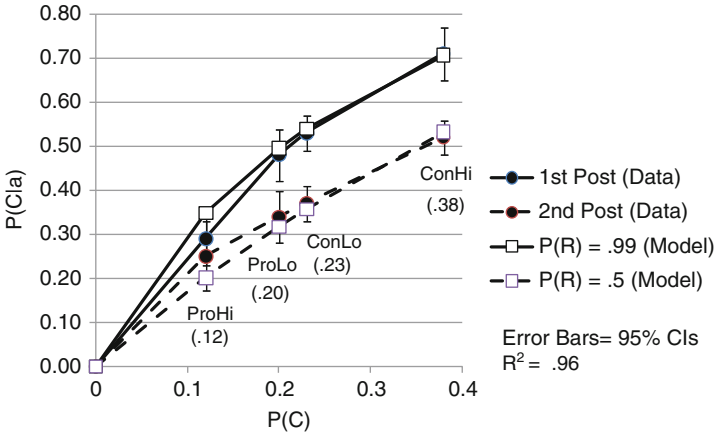

Fig. 2 Fit of the model to the data aggregated across ad hominem argument type (data) showing the Bayesian source reliability models predictions (model) at two levels of $P(R)$ fixed to .99 and .50 (as in Fig. 1) with only $P(C)$ free to vary (*Note*. The fitted value of $P(C)$ for the pro argument high initial degree of belief condition (ProHi) was .12; for the pro argument low initial degree of belief condition (ProLo), it was .20; for the con argument low initial degree of belief condition (ConLo), it was .23; and for the con argument high initial degree of belief condition (ConHi), it was .38. 1st post = 1st posterior judgement and 2nd post = 2nd posterior judgement; CIs = confidence intervals; R^2 is the coefficient of determination which indicates the proportion of variance in the data accounted for by the model)

was set to .99 for the 1st posterior and to .5 for the 2nd posterior. As we argued in the introduction, with these factors fixed then the differences between the 1st posterior for the different conditions—pro argument high initial degree of belief condition (ProHi), pro argument low initial degree of belief condition (ProLo), con argument low initial degree of belief condition (ConLo), con argument high initial degree of belief condition (ConHi)—must be due to differences in the prior, $P(C)$. $P(C)$ was the only parameter varied to capture the data in each of the four conditions. As it took on four different values, there were effectively four degrees of freedom against which to assess the model fit. In Fig. 2, we have plotted the best fit values of the prior $P(C)$ on the x-axis and the 1st and 2nd posteriors on the y-axis, for the data and for the model. Figure 2 shows good eye-ball fits, which were confirmed using the coefficient of determination, R^2, which indicates the proportion of variance in the data accounted for by the model. We have not attempted to formally optimise these fits but nonetheless an R^2 value of .96 was easily obtained by adjusting $P(C)$ by hand to get as close as possible to the 1st posterior and then using this value to calculate the 2nd posterior and then iterating once or twice. This R^2 value indicates that at least 96% of the variance in the data can be accounted for by the model. Figure 2 also shows that each predicted value fell within the 95% confidence interval for the data. In sum, the Bayesian source reliability model can account for most of the variation in our data on how convincing people find the ad hominem argument.

7 Conclusion

In this chapter, we have presented a Bayesian source reliability model of how the argumentum ad hominem should modify people's degrees of belief in a conclusion. According to this model, someone judges how convinced they are by an argument by initially assuming a principle of charity, that is, they assume that their informant is maximally reliable. The Bayesian model separates out the contribution of the argument and the reliability of the informant so that the likelihood ratio that maps their prior degree of belief into their posterior degree of belief is sensitive to both factors. The ad hominem argument represents a direct attack on the charity assumption, suggesting that the informant is unreliable. The effect on the Bayesian model is to reduce the likelihood ratio, so reducing someone's posterior degree of belief in the conclusion. We presented an experiment that tested this model using both pro and con versions of the argumentum ad hominem and materials derived from van Eemeren et al. (2009). That experiment revealed most of the effects predicted by the Bayesian model, which moreover provided excellent overall fits to the data.

This was also the first experiment we can find directly testing the effects of the argumentum ad hominem on people's degrees of belief in an argument, that is, on argument convincingness. While van Eemeren et al. (2009) showed that judgements of reasonableness appeared sensitive to violations of the pragma-dialectical freedom rule, they also speculated that how convincing people found the argumentum ad hominem would be sensitive to a variety of other factors. Our experiment seems to show that it is these other factors related to the content of the argument that primarily drive how convincing people find the argumentum ad hominem. We found no differences between different types of the argumentum ad hominem, where the freedom rule was violated, and a control, which introduced no violation of the freedom rule. However, consistent with the pragmatics of politeness we did find a trend over the different types of the argumentum ad hominem consistent with van Eemeren et al's (2009) results on reasonableness judgements and a difference between the abusive type and the other types of ad hominem argument. Further experiments are of course needed to determine what role if any pragma-dialectical discourse rules may have in determining how convincing we find this and a variety of other arguments.

Appendix: Experimental Materials

Abusive

(A) (Even) After listening to him, I *think it's possible* (strongly believe) that Ford cars *simply* (don't) drive better.
(B) Actually, you should be certain that they don't drive any better.
(A) Why do you think that?
(B) Because how would he know? He doesn't know the first thing about cars.

Circumstantial

(B) I *think it's possible* (highly likely) that her recommendation to use Stelcom Ltd is (not) a good one; (even though) she says that they are the only contractor in the Netherlands that can handle such an enormous job.
(B) Well, you should be absolutely certain that her recommendation isn't a good one.
(A) Why do you say that?
(B) How can we really believe her? Surely, it's no coincidence that the company is owned by her father-in-law.

Tu Quoque

(A) It's *possible* (highly likely) he *was* (wasn't) right to criticise the way in which they processed the data statistically.
(B) Well, you should be convinced that he wasn't right to criticise them.
(A) What do you mean?
(B) Because he said they should have expressed the figures as percentages. But how can he say that when his own statistics are not up to the mark.

Control

(A) I *think it's possible* (strongly believe) that her scientific integrity *is* (isn't) impeccable. She says her research has always been honest and sound.
(B) Well, you should be convinced that her integrity isn't impeccable.
(A) Why do you say that?
(B) Well, how can you really believe her? She has already been caught twice tampering with her research results.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale: Erlbaum.
Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.
Brinton, A. (1995). The *ad hominem*. In H. V. Nansen & R. C. Pinto (Eds.), *Fallacies: Classical background and contemporary developments* (pp. 213–222). University Park: Pennsylvania State University Press.
Cherniak, C. (1986). *Minimal rationality*. Cambridge, MA: MIT Press.
Coady, C. A. J. (1992). *Testimony: A philosophical study*. Oxford: Oxford University Press.
Davidson, D. (1974). On the very idea of a conceptual scheme. In D. Davidson (Ed.), *Inquiries into truth and interpretation* (pp. 183–198). Oxford: Oxford University Press.
Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Belmont: Thompson Wadsworth.
Gillies, D. (2000). *Philosophical theories of probability*. London: Routledge.
Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese, 152*, 207–236.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114*, 704–732.

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic, 29*, 337–367.

Harris, A.J.L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*, 1366–1373.

Ikuenobe, P. (2004). On the theoretical unification and nature of the fallacies. *Argumentation, 18*, 189–211.

Johnson, B. T., Maio, G. R., & Smith-McLallen, A. (2005). Communication and attitude change: Causes, processes, and effects. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes and attitude change: Basic principles* (pp. 617–669). Mahwah: Erlbaum.

Olsson, E. J. (2005). *Against coherence: Truth, probability and justification*. Oxford: Oxford University Press.

Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation*. Notre Dame: University of Notre Dame Press.

Shogenji, T. (1999). Is coherence truth-conducive? *Analysis, 59*, 338–345.

van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation. The pragma-dialectical approach*. Cambridge: Cambridge University Press.

van Eemeren, F. H., Grootendorst, R., & Snoeck Henkemans, F. (1996). *Fundamentals of argumentation theory*. Mahwah: Erlbaum.

van Eemeren, F. H., Garssen, B., & Meuffels, B. (2009). *Fallacies and judgements of reasonableness*. Dordrecht: Springer.

Voss, J. F., & Van Dyke, J. A. (2001). Argumentation in psychology: Background comments. *Discourse Processes, 32*, 89–111.

Walton, D. N. (1990). What is reasoning? What is argument? *Journal of Philosophy, 87*, 399–419.

Walton, D. N. (2000). Case study of the use of a circumstantial *ad hominem* in political argumentation. *Philosophy and Rhetoric, 33*, 101–115.

Walton, D. N. (2009). *Ad hominem arguments*. Tuscaloosa: University of Alabama Press.

Wilson, N. L. (1959). Substance without substrata. *The Review of Metaphysics, 12*, 521–539.