"Damned by Faint Praise": A Bayesian account

Adam Harris (harrisaj@cardiff.ac.uk)
School of Psychology, Cardiff University,
Tower Building, Park Place,
Cardiff, CF10 3AT, UK.

Adam Corner (corneraj@cardiff.ac.uk)
School of Psychology, Cardiff University,
Tower Building, Park Place,
Cardiff, CF10 3AT, UK.

Ulrike Hahn (hahnu@cardiff.ac.uk)
School of Psychology, Cardiff University,
Tower Building, Park Place,
Cardiff, CF10 3AT, UK.

Abstract

"Damned by Faint Praise" is the phenomenon whereby weak positive information leads to a negative change in belief. However, in a Bayesian model of belief revision positive information can seemingly only exert a positive change in belief. We introduce a version of Bayes' Theorem incorporating the concept of epistemic closure. This reformalization is able to predict the conditions under which a 'damned by faint praise' effect is observed. Moreover, good, parameter-free fits are observed between the Bayesian model and the experimental data. This provides further support for the Bayesian approach to informal argumentation (e.g., Hahn & Oaksford, 2007).

Keywords: Argument from ignorance; Bayesian probability; Epistemic closure; Evidence

Introduction

'James is always punctual and polite'

The above sentence clearly identifies positive aspects of James' character. Were this however the only information you were given about James within the context of a reference letter, it is likely that your overall impression of James would not substantially improve. In fact, it seems more likely that your impression of him might be lowered through the receipt of this information. In colloquial English, one might say that James was 'damned by faint praise'.

The 'Boomerang effect', by which a very weak positive argument can actually lead to a negative change in belief, through the internal generation of stronger counter-arguments, is already well documented within social psychology (e.g., Petty & Cacioppo, 1996). We, however, are specifically concerned with the effect of weak arguments that clearly *exclude* important information relating to the issue in question. Following the familiar colloquial expression 'damned by faint praise', we shall refer to negative belief change following the receipt of positive evidence in these cases (which are the focus of the present paper) as the Faint Praise effect.

Upon first consideration, the Faint Praise effect represents a considerable challenge for the Bayesian theory of belief revision – how can positive evidence ever lead to negative belief change? Here we present and test a Bayesian formalization within which the Faint Praise effect is readily explained.

Formalising the Faint Praise Effect

The Bayesian framework provides a normative theory for belief revision. On receipt of new evidence, people should update their beliefs in a hypothesis in line with Bayes' Theorem:

$$P(h \mid e) = \frac{P(h)P(e \mid h)}{P(e)}$$

where *P(h)* is a person's prior belief in the truth of the hypothesis under scrutiny and *P(e|h)* is the likelihood term, which captures the sensitivity of the test in signal detection theory terms. This conditional probability is the individual's subjective belief in the probability that the provided piece of evidence would be found given that the hypothesis were true. *P(e)* corresponds to the base rate of the evidence item, and *P(h|e)* is the individual's posterior degree of belief in the hypothesis given this new piece of evidence. Assuming

$$P(e \mid h) \geq P(e \mid \neg h)$$

that is, assuming the evidence is not known to be misleading, evidence in favor of the hypothesis, no matter how weak, can never decrease the person's degree of belief in the hypothesis. Bayes' theorem stipulates that positive evidence should only have a positive impact on belief change (see e.g., Lopes, 1985). Yet the Faint Praise effect is seemingly a demonstration of the opposite: weak positive evidence resulting in a negative change in belief.

A consideration of a well-known type of informal argument, however, the so-called "argument from

ignorance", suggests that the Faint Praise effect may actually be amenable to a Bayesian formalization.

Arguments from ignorance are arguments based on the absence of evidence, such as:

> 'Ghosts exist because nobody has proved that they don't.'

Examples like this have led to the view that arguments from ignorance are fallacious (e.g., Copi & Cohen, 1990; Evans & Palmer, 1983; Van Eemeren & Grootendorst, 2004). Oaksford & Hahn (2004), however, noted that not all arguments from ignorance seem equally weak (see also Walton, 1992). There are examples of this argument form which seem far more plausible and, indeed, form the basis of modern science. For example:

> 'This drug is safe because the British Medical Association have not found any evidence of side effects'

One factor that influences the strength of arguments from ignorance is the notion of *epistemic closure* (Hahn & Oaksford, 2007; Hahn, Oaksford, & Bayindir, 2005; Walton, 1992). Epistemic closure is perhaps best understood by means of an example: Upon consulting a railway timetable to determine whether the 12:45 Cardiff to London train stops at Oxford, one assumes that the timetable provides a complete list of all the stations, *and only those stations*, at which the train will stop (i.e., the timetable is a database that is epistemically closed). Consequently, if Oxford is not included on the timetable then one can confidently conclude that the train will not stop at Oxford. Hence, the following argument from ignorance seems entirely reasonable:

> 'The train will not stop in Oxford, because an Oxford stop is not listed in the timetable'

How, then, can this concept of epistemic closure explain the 'damned by faint praise' phenomenon? We propose that when a weak argument results in a change in belief in the opposite direction to that intended by the argument, it is as a result of implicatures drawn by the argument's recipient relating to evidence *not* included in the argument. Consequently, the change in belief brought about by such a weak argument results from an *implicit* argument from ignorance. If James' maths teacher writes a reference to support James' application for a university mathematics course that reads 'James is punctual and polite', then he his flouting the conversational maxim of quantity: "Make your contribution as informative as is required" (Grice, 1975/2001, p. 171). Through a recognition that James' maths teacher would surely know more about James than these two facts (such as, for example, his maths ability) and presumably be motivated to include this information were it true, the reader can imply that the referee must "be wishing to impart information that he is reluctant to write down" (Grice, 1975/2001, p. 171). Thus, the possible impact of such evidence is best understood as an example of an argument from ignorance – that is, the effect of *not* saying "James is good at maths".

Hahn and colleagues (Hahn & Oaksford, 2007; Hahn et al., 2005) proposed a Bayesian formalization of the argument from ignorance which included the concept of epistemic closure. In addition to e and $\neg e$, a third possibility, represented by the term n is included in this Bayesian formalization. Here, n (as in 'nothing') refers to a lack of evidence (i.e., not explicitly saying either "e" or "not e") whilst $\neg e$ refers to explicit negative evidence. Such an approach is familiar from Artificial Intelligence where one might distinguish three possibilities in a database search regarding a proposition (h): the search can either respond in the affirmative (e), the negative ($\neg$e), or it can find nothing on the subject (n). Epistemic closure has been invoked here to license inferences from search failure (i.e., a query resulting in nothing) to non-existence, given that the database is assumed to be complete.

The Bayesian formalization of epistemic closure is directly analogous, except that it is probabilistic, and hence accommodates arbitrary degrees of closure (Hahn & Oaksford, 2007). For example, one might be certain that one has lost a red sock after looking throughout the house, but also be fairly certain if one has only looked in several key locations, such as the drawer and the washing machine.

This three-valued approach to evidence, which allows one to capture degree of closure by varying the probability of a 'no response' *($P(n|h)$)*, helps capture two kinds of arguments from ignorance:

(a) *not* (Database says: *not* exists), therefore exists

e.g., Ghosts exist, because nobody has proven that they don't, and

(b) *not* (Database says: exists), therefore *not* exists

e.g., I cannot find my sock, hence it is lost.

We propose that the "damned by faint praise" phenomenon stems from an inference of type (b) above. To return to our earlier example, the referee does not say that James is good at maths, so the inference that is subsequently made is that he is not good at maths. The reader's degree of belief in the falsity of a given hypothesis having not received a specific item of evidence is therefore given by:

$$P(\neg h \mid n) = \frac{P(\neg h)P(n \mid \neg h)}{P(n)}$$

where $P(n|\neg h) = 1 - [P(e|\neg h) + P(\neg e \mid \neg h)]$ and $P(n) = 1 - [P(e) + P(\neg e)]$. Thus, this account incorporates a result of Gricean pragmatics of conversation into a normative Bayesian framework.

Within the present conceptualization, the faint praise effect will occur wherever $P(n|h) < P(n|\neg h)$. Hence, it should be observed where a motivated (or positively inclined), but non-lying source is presenting an argument. By contrast, there is no reason for a faint praise inference in the case of a maximally uninformed source, $P(n|h) \approx P(n|\neg h)$, who simply knows nothing on the topic, or given a source who prefers, where possible, to provide negative information $P(n|h) > P(n|\neg h)$.

Returning to our example of the reference letter, consider that instead of being written by James' maths teacher, it has been written by James' tutor who does not teach maths and who only rarely meets with James. In this instance, it could reasonably be assumed that his tutor would not possess any specialist knowledge concerning James' maths ability. With this assumption, the reader of the reference letter could not make any inferences pertaining to James' intelligence on the basis of the letter. Consequently, our theory would not predict the occurrence of a "damned by faint praise" effect in this instance as $P(n|h) \approx P(n|\neg h)$.

It should also be clear from the above, and our focus upon what is *not* being said as the explanation for the 'damned by faint praise' effect, that material that has preceded the weak argument will affect its influence. Returning again to our example of the reference letter, we suggest that inferences will be made about James' maths ability as a result of the referee not mentioning his maths ability within the letter. Now consider the case in which the referee has already stated that the candidate is a very good mathematician. Following this reference it is likely that the reader has raised his opinion of James' maths ability. If the referee now adds that the candidate is punctual (our previous weak argument), what effect will it have upon the reader's newly revised degree of belief? Because our theory assumes that a weak positive argument has a negative impact through an implicit argument from ignorance (as a result of not including important positive information), a negative impact is no longer predicted (as the important positive information *is* included). A strong argument in favour of the candidate has already been presented, and so this new evidence will have a persuasive impact based solely upon the importance that the reader places upon punctuality. Hence, his opinion of the candidate will either remain unchanged or increase, but it will not decrease.[1]

The hypotheses of the present study are therefore that a weak argument from an expert source will result in a decrease in degree of belief of the hypothesis being advocated (a Faint Praise effect), whilst a weak argument from a non-expert source will not have this effect. In addition, a weak argument preceded by a strong argument will also not have this effect. The first prediction is consistent with findings of Birnbaum and colleagues (Birnbaum & Mellers, 1983; Birnbaum & Stegner, 1979) who found that effects of source bias were greater for sources of greater expertise.

[1] Assuming, of course, that the reader of the reference letter does not consider punctuality to be a negative trait.

Method

Participants

95, predominantly female, Cardiff University undergraduates participated in the experiment in exchange for course credit. They were randomly assigned to one of three experimental conditions.

Design, Materials and Procedure

We manipulated two independent variables: the perceived expertise (i.e., the presumed knowledge) of the source, and the type of argument presented. There were three experimental conditions, such that participants either read a strong followed by a weak argument from an expert source, only a weak argument from an expert source, or only a weak argument from a non-expert source. The remaining possible condition (a strong argument, followed by a weak argument from a non-expert source) was not included as it did not impact directly on our experimental predictions.

Participants were presented with a fictional UCAS application[2] containing background information about James Driver (date of birth, address, GCSE[3] grades obtained). Based on this information, participants were required to indicate on a scale from 0 (definitely not) – 100 (definitely) whether they thought the applicant should be accepted to study mathematics at Newcastle University. Depending on the experimental condition they were assigned to, they were then presented with either one or two arguments from an expert or a non-expert source. The expert source was an A-level mathematics teacher, who had 'taught James maths throughout his AS and A-level course.' The non-expert source was the personal tutor of the applicant, who only met him 'once a term to discuss any concerns James has'. The weak argument stated that James was:

'*punctual, and always did his best to look smart'*.

The strong argument stated that the applicant was:

'*sharp and clever with an ability to critically analyze others' proofs and theories. When new material is introduced he is quick on the uptake. On the odd occasion when he has failed to grasp a concept straight away, he has demonstrated considerable maturity in his use of the library's resources to help him understand the topic in question.'*

[2] UCAS is the organization through which British school pupils apply to university.

[3] GCSEs are the first public examinations taken by pupils in the UK.

Following each argument, participants were asked to provide updated ratings of whether they thought James should be accepted onto the course.

After providing their numerical ratings, participants were also asked to indicate why they did (or did not) change their ratings.

As an additional test of our Bayesian account we then obtained participants' estimates of the relevant conditional probabilities for a Bayesian formalization of the argument. Through such data we were able to provide a direct, quantitative test of the validity of a Bayesian formalization of the Faint Praise effect. As our formalization predicts that the effect occurs because of a lack of information, at the end of the experiment participants gave 6 conditional probabilities, three conditional on the state of affairs that James is an intelligent and resourceful mathematician, and three conditional on the state of affairs that James is NOT an intelligent and resourceful mathematician. The format in which these questions were asked is given below:

In your opinion, if James ***is*** an intelligent and resourceful mathematician, what is the chance that his maths teacher, who has taught James all his AS-level and A-level maths, would state in his UCAS reference for James:

(a) that James ***is*** an intelligent and resourceful mathematician
(b) that James is ***not*** an intelligent and resourceful mathematician
(c) he would make ***no mention*** of this information

Please report your answers as numbers between 0 (absolutely no chance) and 100 (would be certain to report this information)

(a) _____
(b) _____
(c) _____

Participants in all experimental groups completed these questions, but for those in the non-expert group, the words 'maths teacher, who has taught James all his AS-level and A-level maths' were replaced by: 'tutor, who meets with James once a term to discuss any concerns he might have'.

Results

Both experimental predictions were supported by the data. Figure 1 shows the effect of the weak argument on participants' ratings of whether James should be accepted to read mathematics at Newcastle University. It is clear from Figure 1 that a weak argument presented on its own by an expert had a significant negative effect on participants' judgments of James' suitability, $t(31) = 2.93$, $p<.01$, while the very same argument presented after a strong argument had a positive effect, $t(31) = 2.98$, $p<.01$. A t-test confirmed that the difference between the two conditions was significant, $t(62) = 3.67$, $p=.001$. Additionally, the direction of this effect negates the possibility that these data could be an artifact of floor or ceiling effects. Having received the strong argument, participants' ratings of James' suitability (M = 77.50; SD = 11.29) were higher than their priors in the weak argument only condition (M = 58.44; SD = 10.88), $t(62) = 6.88$, $p<.001$.

In addition, the weak argument from the non-expert source had a significantly weaker negative effect than that from the expert source on ratings of James' suitability, $t(61) = 1.83$, $p<.04$ (1-tailed). Indeed, the weak argument from the non-expert source exerted no reliable effect on people's judgments of James' suitability, $t(30) = 0.78$, $p>.05$.

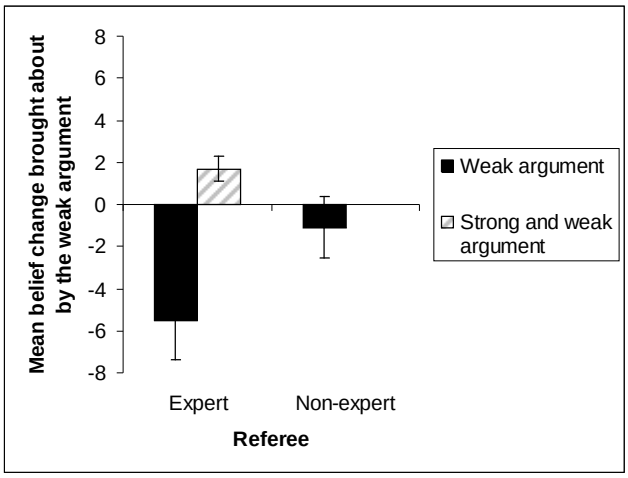

Figure 1: Mean effect of weak argument on ratings of applicant's suitability in each experimental condition. Error bars are plus and minus 1 standard error.

A scan of participants' explanations detailing why their degree of belief had changed supported our intuition that the Faint Praise effect is a result of an inference from missing evidence – participants' cited reasons such as: "He left out lots of information about his ability." In addition to this qualitative support, we were also able to analyze the predictions derived from Bayes' Theorem using the conditional probabilities supplied by participants . From an analysis of participants' subjective conditional probabilities, it was possible to split the data, for those participants receiving only the weak argument, into groups of participants whose subjective conditional probabilities would predict them to show the faint praise effect (that is, $P(n|h) < P(n|\neg h)$) and those whose probabilities would not.[4] Only four (out of 62) participants reported conditional probabilities that did not satisfy this inequality and none of these demonstrated a Faint Praise effect, supporting a Bayesian formalization of the phenomenon. Participants who had received both a strong and weak argument were not included in any parameter-based analyses, because the

[4] One participant did not provide conditional probabilities and therefore could not be included in this analysis.

relevant conditional probabilities were not provided in this condition.

As a general check of the validity of the Bayesian approach to argument convincingness, a correlation was performed between the responses made by participants, and those predicted from their conditional probabilities (calculated from Bayes' Theorem).[5] The strong correlation between observed ratings and those predicted from their Bayesian conditional probabilities, $r(60) = .43$, $p<.001$, does however provide a good fit of the Bayesian model to the experimental data. Notably, there were no free parameters in the model which could be 'tweaked' to match the data. In addition, the model is compared simultaneously with the individual data points of all eligible participants. Thus, the model evaluation has not benefited from *any* averaging of experimental data (as is usually performed for such analyses) which would reduce the amount of noise in the estimates (see e.g., Ariely et al., 2000; Cohen, Sanborn, & Shiffrin, 2008; Wallsten, Budescu, Erev, & Diederich, 1997).

Discussion

At first glance, the Faint Praise effect would seem to present a challenge to the Bayesian theory of belief revision. How can positive evidence ever lead to negative belief change? We have proposed that the effect is an example of an implicit argument from ignorance – and so the negative belief revision is being driven by inferences about the *absence* of positive information. A Bayesian formalization incorporating the concept of epistemic closure enabled empirical predictions to be made and tested. The experimental data matched Bayesian predictions, and a good parameter free fit was obtained between participants' ratings and the Bayesian model.

Notably, our results cannot be accommodated by averaging models of belief adjustment as these would predict the opposite effect to that observed in our data (e.g., Hogarth & Einhorn, 1992; Lopes, 1985). An averaging model effectively evaluates the polarity of a piece of information with respect to current belief. Thus, such models predict a stronger *negative* effect of positive evidence after the receipt of a strong argument than without the prior receipt of a strong argument. Such a result is often observed in traditional belief updating tasks (see Hogarth & Einhorn, 1992; Lopes, 1985, and references therein). Lopes has suggested that the propensity for people to use averaging models can explain conservatism in traditional belief updating tasks (e.g., Edwards, 1968). Lopes (1985) tentatively suggested that averaging might be less likely to occur in situations where stimuli were more clearly 'marked' in support of or against a given hypothesis. Subsequently, Lopes (1987) succeeded in reducing participants' use of an averaging rule by instructing them to separate their judgments of belief updating into 2 steps, where the first was the labelling of a piece of evidence as either favouring or countering the hypothesis. We believe that our participants did not show the use of sub-optimal averaging strategies as the domain used is familiar to them and hence the evidence is subjectively well 'marked' as to the hypothesis it supports. The failure to observe the use of averaging strategies in this context suggests that previous documentation of their sub-optimal usage (e.g., Lopes, 1985) may result from the unfamiliar and artificial nature of traditional belief updating tasks (e.g., the bookbag and poker chip paradigm).

McKenzie, Lee and Chen (2002) described data, using both legal and informal interpersonal dispute paradigms, demonstrating that evidence in favour of side A sometimes decreased confidence in side A (thus resembling the Faint Praise effect). They propose that such data can be captured by traditional adding and averaging models of belief revision *if* the reference point against which the evidence is compared is dynamic and more demanding than neutrality. McKenzie et al. presented the data from four experiments which supported their proposal of a new standard against which a case's strength should be assessed, its *Minimum Acceptable Strength* (MAS). They propose, and offer empirical support for, three claims related to the MAS. Firstly, they show that it is more demanding than neutrality. Secondly, they show that the MAS for the second case is influenced by the strength of the first case, such that it is more demanding following the presentation of a strong opposing case. Thirdly, they show that it is more demanding than neutrality *because* the cases are presented with a bias. In other words, the evidence presented is not a random selection of all the information available about the hypothesis in question, but is specifically chosen by the discussant to support their own interests. McKenzie et al.'s theory could extend to explain the data reported in this paper through this third claim. The reference point for an argument being presented in a reference letter might be considered more demanding than neutrality because references are normally written to enhance the positive aspects of their subject and as such they are positively biased. If an argument does not meet the MAS then participants' implicit reasoning appears to be along the lines of 'If *that's* the best they can do, then I believe the other side (even) more.'" (McKenzie et al., 2002, p. 14). We, however, propose that the present, Bayesian, account of such empirical findings represents greater parsimony for it is able to account for the result within the existing Bayesian framework without the need to add a new comparative standard (MAS) to an already existing theory. In addition, the finding that the relevant conditional probabilities

[5] Although the hypothesis that James is an intelligent and resourceful mathematician is not the same as the hypothesis that he should be accepted to read mathematics at Newcastle University, this analysis proceeded from the assumption that James being an intelligent and resourceful mathematician is the central concern of whether or not he should be admitted to a mathematics course at Newcastle University. Consequently it is presented only as an indication of the potential power of the Bayesian formalization rather than as direct evidence.

predicted the Faint Praise effect provides quantitative support for this Bayesian account.

Perhaps the most fundamental point to be taken from this experiment is that the relationship between source and message characteristics is a complicated one, and the effect that an argument will have on the recipient of that argument cannot be predicted without information pertaining to the source of that argument. Indeed, in some situations these interactions can result in seemingly counter-intuitive results, such as a weak message being less persuasive from an expert source than from a non-expert source. This work extends that of Hahn and colleagues (e.g., Hahn & Oaksford, 2007; Oaksford & Hahn, 2004) in which a Bayesian approach to argument fallacies has been introduced. A practical problem for this approach is determining how the relevant probabilities are derived. The results of this study suggest that among the factors influencing the conditional probabilities [$P(e|h)$; $P(\neg e|h)$; $P(n|h)$] is the subjective impression of the depth of knowledge and bias of a message communicator, which must be considered within the context of the conversational pragmatics that govern language use (Grice, 1975/2001).

Acknowledgments

Adam Harris and Adam Corner were supported by ESRC studentships. We thank James Close, Carl Hodgetts and Andreas Jarvstad for assisting with data collection and Harriet Over for helpful comments on an earlier draft of the manuscript.

References

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., et al. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6,* 130-147.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792-804.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology, 37,* 48-74.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin and Review, 15,* 692-712.

Copi, I. M., & Cohen, C. (1990). *Introduction to Logic (8th Edition).* New York: Macmillan.

Edwards, W. (1968). Conservatism in Human Information Processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment (pp. 17-52).* New York: Wiley.

Evans, D., & Palmer, H. (1983). *Understanding Arguments.* Cardiff, UK: Department of Extra-Mural Studies, University College, Cardiff.

Grice, H. P. (1975/2001). Logic and conversation. In A. P. Martinich (Ed.), *The Philosophy of Language (4th Edition)* (pp. 165-175). Oxford: Oxford University Press.

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

Hahn, U., Oaksford, M., & Bayindir, H. (2005). How convinced should we be by negative evidence? *Proceedings of the Annual Conference of the Cognitive Science Society, 27,* 887-892.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1-55.

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society, 23,* 509-512.

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica, 64,* 167-185.

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making, 15,* 1-18.

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology, 58,* 75-85.

Petty, R. E., & Cacioppo, J. T. (1996). *Attitudes and persuasion: Classic and contemporary approaches.* Boulder, CO: Westview Press.

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

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243-268.

Walton, D. N. (1992). Nonfallacious arguments from ignorance. *American Philosophical Quarterly, 29,* 381-387.