Evaluating arguments from the reaction of the audience

Hugo Mercier

University of Pennsylvania

hmercier@sas.upenn.edu

Brent Strickland

Yale University

brent.strickland@yale.edu

In press *Thinking and Reasoning*

Abstract.

In studying how lay people evaluate arguments, psychologists have typically focused on logical form and content. This emphasis has masked an important yet underappreciated aspect of every day argument evaluation: social cues to argument strength. Here, we focus on the ways in which observers evaluate arguments by the reaction they evoke in an audience. This type of evaluation is likely to occur either when people are not privy to the content of the arguments or when they are not expert enough to appropriately evaluate it. Four experiments explore cues that participants might take into account in evaluating arguments from the reaction of the audience. They demonstrate that participants can use audience motivation, expertise and size as clues to argument quality. By contrast, we find no evidence that participants take audience diversity into account.

Keywords: argument evaluation, argumentation, audience, social cues.

Arguments can be evaluated in several ways. Typically, their form and content is the subject of scrutiny. Their logical validity can be assessed, and this has been the topic of logic. Their content also matters: minimally, if the premises are known to be false, the argument loses its appeal. Straddling form and content, argumentation schemes—appeal to authority, argument from example, etc.—can be evaluated by a variety of means, such as their fit with what is deemed to be the proper scheme in each case (e.g. Walton, Reed, & Macagno, 2008). There is however another *kind* of way to evaluate arguments: by gauging the reaction of their audience. An argument that has convinced its audience is, *ceteris paribus*, more likely to have been sound than an unsuccessful one. This criterion is hardly new: universal acceptance is for Kant the strongest grounds to accept an argument: "If something is valid for anybody in possession of his reason, then its grounds are objective and sufficient" (Kant, 1999 cited in Popper, 2002, p.22).

Most attention in experimental psychology has focused on the first way to evaluate arguments. The psychology of reasoning has mostly focused on testing people's ability to detect the logical validity of arguments, or to follow other normative guidelines such as Bayes' rule (Evans, 2002; Hahn & Oaksford, 2007). Studies of persuasion and attitude change have evaluated the impact of numerous factors, including the strength of different arguments (Petty & Wegener, 1998). By contrast, little attention has been devoted to the evaluation of arguments by way of the reaction of their intended audience. Yet we are often unable to evaluate the form or content of arguments. In some cases, we are simply not privy to the arguments and the reaction of the audience is the only clue we have. In other cases the arguments are available, but we don't have the required expertise. Lack of expertise does not entail a very technical or abstruse topic, but merely lack of

knowledge, as when the argument bears on a vague acquaintance. And even if we can rely on the form and content of the argument, gauging the audience reaction could provide complementary clues to its strength.

In this article, we provide a first exploration of the different ways in which audience responses may be considered in evaluating an argument.

How might people evaluate arguments based on the reaction of the audience?

Arguments are always addressed *at* someone, namely the recipient of the argument, and that recipient may be an individual or a group. However, the reach of an argument will frequently also extend beyond the intended recipient, affecting other individuals not directly addressed by the proponent. All of those people reached by an argument may, in a sense, be considered to be the 'audience'. Argumentation theorists have even considered hypothetical audiences, such as Perelman and Olbrechts-Tyteca's idea of the universal audience, a construct invoked in seeking to provide a rational foundation for rhetoric: "Argumentation addressed at a universal audience must convince the reader that the reasons adduced are of a compelling character, that they are self-evident, and possess an absolute and timeless validity, independent of local or historical contingencies" (Perelman & Olbrechts-Tyteca, 1958, p.32).

Deciding which audience matters in assessing an argument bears on the normative question: "what is a good argument?" Indeed, this is what Kant does in the citation above: he defines a good argument by reference to its effect on an audience: an argument is deemed to be good if it convinces an extremely encompassing audience comprising "anybody in possession of his reason." By contrast with this *Kantian* view, a *pragmatic*

view would focus purely on the intended recipient of the argument and her or his reaction. On this view, the only goal of reasoning is to convince "the ensemble of those whom the speaker wishes to influence by his argumentation" (Perelman & Olbrechts-Tyteca, 1958, p.19). Consequently, an argument is deemed to have been good as long as this intended audience is convinced. The assent or dissent of any other individual, or group of individuals is irrelevant.

The reaction of the audience can also be used to determine what is a good argument from an evolutionary perspective. The argumentative theory of reasoning posits that reasoning evolved to convince others, and to only be convinced when appropriate (Mercier & Sperber, 2011). In this *evolutionary* view, the goal of argument production is to convince the intended audience, as in the pragmatic view. However, it also specifies that reasoning has to be efficient in its *normal conditions*, "the conditions to which the device that performs the proper function is biologically adapted" (Millikan, 1987, p.34). We surmise that the normal conditions for reasoning to produce arguments are as follows: a dialogue with an audience that shares the arguer's reasoning abilities, as well as most of her beliefs, but who disagrees about one point. A good argument is one that convinces such an audience. It is therefore less exigent than the Kantian view (assent does not have to be universal), but more exigent than the pragmatic view (assent has to be gained from someone in a good position to evaluate the argument).

The Kantian, pragmatic and evolutionary views, however, do not aim at describing how people actually take the audience's reaction into account when evaluating an argument. This is apparent from the counter-intuitive predictions they make. Any

synthetic argument[1] is very likely to be rejected at some point, and would therefore be deemed not to be a good argument following a strict Kantian view, leaving very few good arguments. If the Kantian view is overly strict, treating as deficient arguments that many would say are good, the pragmatic view suffers from the opposite flaw: it accepts too many arguments—such as arguments accepted by an incompetent audience for instance. The evolutionary view is less immediately counter-intuitive, but it is still likely to be overly inclusive in its definition of a good argument, accepting for instance historical arguments that have long been refuted.

While these different views address an interesting philosophical question about what, in an abstract sense, makes an argument 'good', they are of little help to someone seeking to gauge the extent to which they should be convinced by a particular argument, given the reaction of its audience. For any individual trying to evaluate a specific argument from the audience's reaction, the question is: what cues can be gathered that will help me decide whether *I* should accept the conclusion? From an argumentation-theoretic perspective, such cues may be viewed either from a scheme-based approach (see for example, Hoeken, Timmers, & Schellens, current issue), or from a probabilistic perspective (see Harris, Hsu & Madsen, in press).

Our focus in this paper is simply on exploring whether or not people actually make use of a number of intuitively plausible clues. The first clue is the reaction of the intended audience, but the reactions of other people beyond the intended audience can also be informative. Moreover, the reactions of both the recipients and of potential third

[1] A synthetic argument relies on propositions that are not true purely by virtue of the terms they contain, by opposition with analytic propositions such as "all squares have four sides."

parties should, intuitively, be modulated by how carefully they evaluated the argument. If they were too stubborn or too distracted to evaluate the argument, rejection is not diagnostic. Likewise, if they already agreed with the arguer, and were therefore unlikely to cast a critical eye on her arguments, acceptance is not diagnostic of argument strength. In the context of a motivated audience other factors can lead to more fine-grained distinctions. The first is the relative expertise of the proponent, the recipient, and potential third parties. If the arguer is the expert, her ability to convince less expert people may be less diagnostic of how good her arguments are. If there is a disagreement within the audience, or between the audience and third parties, the reaction of the most expert group should take precedence (on expertise and arguments from authority see also, Hoeken et al., current issue; Harris et al., current issue; van Eemeren, Garssen, & Meuffels, current issue). Audience diversity could also be a valuable cue. The more diverse the audience, the more diagnostic its acceptance of an argument should be. Audience size should also matter: the more people accept an argument, the more likely it is that the argument is sound. It should be noted, however, that for size to matter some degree of diversity must exist: being able to convince several people who have exactly the same knowledge and ability is not more diagnostic than convincing one of them. Finally, we might expect in reasoners a general bias towards caution. In general, the dangers of accepting too much information are larger than that of rejecting too much (Mercier, in press), and so people should not easily say that an argument is good when they are in doubt.

In the experiments below, we provide an exploration of these cues in situations where they are the sole evidence available to participants for argument evaluation.

Although this situation reflects some real life settings, in most cases people would use social cues—the audience reaction—in conjunction with some evaluation of the content of the argument. We leave open the question of how this integration is made, and what weight social cues receive in different contexts.

Overview of the experiments

In the following experiments, participants had to evaluate arguments purely on the basis of the reaction of audiences and third parties. Different characteristics of the arguer, the audience and third parties were varied in order to test which characteristics have an impact on argument evaluation. All experiments were carried out online using a similar template. Accordingly, the characteristics of the participants as well as the outline of the procedures are provided here for all experiments.

Participants

All participants were recruited through the Amazon Mechanical Turk website. They were recruited using the same procedure, in close temporal proximity and so will be treated as a single group (N = 491). They were paid $0.1 for their participation, a normal rate for this type of task in Mechanical Turk. All participants had to be in the US at the time of the experiment. A mechanism stops most participants from taking part in the same experiment, or in several experiments of the same study, several times. However, it is fallible, so the IP addresses of all the participants were checked. Five results were deleted because the same IP had already been used in a previous experiment of the study.

The average participant age was $M = 34.1$, $SD = 12.4$. Sixty-three percent of the participants were female. Finally, 85% of participants had at least some college education. Several published studies already rely on this sample (e.g. DeScioli & Kurzban, 2009), and specifically designed experiments have established its reliability (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010).

Procedure

Participants registered on the Amazon Mechanical Turk Website. They were then redirected to the website of the experiment itself, programmed using the Qualtrics software. They went through the questions one by one, without possibility of going back, and provided some demographic information at the end of the experiment.

Experiment 1: Motivation

The goal of Experiment 1 is to test the impact of audience motivation on the assessment of arguments. In the High Motivation condition the recipient accepts the argument after previously disagreeing with the proponent and should therefore have been motivated to evaluate the argument. In the Low Motivation condition the recipient accepts the argument after previously agreeing with the proponent and should therefore not have been motivated to evaluate the argument very thoroughly. In the Mixed Motivation condition the recipient is not convinced by the argument, but she has paid no attention to the argument; a third party pays attention and accepts the argument. If audience response is viewed as a cue that varies in diagnosticity as a function of audience characteristics, then the argument should be seen to be strong only where a motivated audience finds it acceptable, that is in the High and Mixed Motivation conditions.

Design

The three conditions were varied between participants.

Materials

In the High Motivation condition, participants read the following text:

> John and Michael are talking about something, and they realize they have a different opinion. John makes an argument supporting his opinion, and Michael finds it convincing.

In the Low Motivation condition, the text was:

> John and Michael are talking about something, and they realize they have the same opinion. John makes an argument supporting this opinion, and Michael finds it convincing.

In the Mixed Motivation condition, the text was:

> John and Michael disagree about something. John makes an argument but Michael does not pay any attention to the argument and, as a result, is not convinced by it. Paul, who overhears the conversation, is convinced by the argument.

After reading the text, participants answered the following question: “Do you think that John's argument is:” which they could answer by “Likely to be good,” “Likely to be poor,” or “I cannot tell.”

Results and discussion

The results are presented in Table 1. They indicate that participants consider audience responses to be informative about argument quality. The Low Motivation condition was significantly different from both other conditions (χ^2, *p*s < .0001), with more "I cannot tell" and less "Likely to be good" answers. This means that in evaluating audience responses participants take audience characteristics that are likely to influence depth of processing into account, whether they are characteristics of the recipient or a third party.

	Likely to be good	Likely to be poor	I cannot tell
High Motivation (N=48)	96%	2%	2%
Low Motivation (N=45)	36%	11%	53%
Mixed Motivation (N=94)	72%	6%	21%

Table 1: Evaluation of arguments by condition in Experiment 1.

Experiment 2: Expertise

The goal of Experiment 2 is to evaluate the impact of expertise on the evaluation of arguments. In all conditions, the proponent produced an argument that convinced the recipient but failed to convince a third party. The expertise of the three characters was varied, allowing examination of the way expertise and agreement/disagreement combine.

Intuitively, one might expect a counterargument by an expert to outweigh agreement by a novice, irrespective of the expertise of the arguer. An expert convincing another expert should be a good clue that the argument is good, even if a novice disagrees, since the novice is more likely to be mistaken. When the level of expertise of the audience and the third parties are similar, it should be harder to tell who is right, and therefore it is more cautious to remain undecided.

Design

The design is a partial 2*2*2, with the first variable being Arguer expertise (Novice or Expert), the second Audience expertise (Novice or Expert) and the third Third Party expertise (Novice or Expert). Three conditions were not conducted. Expert-Expert-Expert did not add anything to Novice-Novice-Novice since it is the relative level of expertise that matters. Novice-Expert-Novice and Novice-Expert-Expert were not conducted because it was not felicitous to have a novice convince an expert. The conditions formed two groups that were conducted between participants.

Materials

We used the following template (here the Expert-Novice-Expert) for all but one of the variants:

> A physicist and a layman disagree about a problem in physics. The physicist makes an argument and convinces the layman to change his mind. Another physicist, who overhears the conversation, thinks of counterarguments and is not convinced.

The roles of the laymen and physicists were permutated to create the different conditions. The only exception was the case of identical level of expertise for which we did not specify the degree of expertise in order to have a natural middle ground between novices and experts:

> John and Michael disagree about something. John makes an argument and convinces Michael to change his mind. Paul, who overhears the conversation, thinks of counterarguments and is not convinced.

The question asked was similar to that of Experiment 1: "Do you think that [proponent]'s argument is:" "Likely to be good," "Likely to be poor," or "I cannot tell."

Results and discussion

The results are laid out in Table 2. We can distinguish three cases. The first is that of equal expertise among all three protagonists, here represented by the Novice-Novice-Novice condition. When two people with equal expertise disagree on their evaluation of an argument, it makes intuitive sense to withhold judgment, which is what a majority of participants did. This condition differs significantly from all the others (four χ^2 with df = 2, *p*s < .05). The second case is when the third party, who rejects the argument, has more expertise than the recipient who accepts it—here the Expert-Novice-Expert and Novice-Novice-Expert conditions. This is a good indication that the argument may be poor, and indeed a majority of participants in these two conditions gave the "Likely to be poor" answer. These two conditions (Expert-Novice-Expert and Novice-Novice-Expert) are not significantly different from one another but differ from the other three conditions (six χ^2 with df = 2, *p*s < .001). Finally, the third case involves a third party who rejects an

argument from a more expert proponent, an argument accepted by the recipient (here Expert-Expert-Novice and Expert-Novice-Novice). A majority of participants (in the Expert-Expert-Novice condition) and a plurality of participants (in the Expert-Novice-Novice) answered that the argument was "Likely to be good." These two conditions differ from the other three (six χ^2 with df = 2, *p*s < .001). It is noteworthy, however, that when the protagonist who agreed with the proponent was only as expert as the one who disagreed (Expert-Novice-Novice), equally many participants answered "I cannot tell" and "Likely to be good," reflective a more cautious stance overall.

Arguer expertise	Audience expertise	Third party expertise	Likely to be good	Likely to be poor	I cannot tell
Expert (N=94)	Expert	Novice	53%	22%	25%
Expert (N=91)	Novice	Expert	15%	53%	32%
Expert (N=94)	Novice	Novice	37%	25%	37%
Novice (N=91)	Novice	Expert	13%	63%	24%
Novice (N=94)	Novice	Novice	18%	30%	52%

Table 2: Evaluation of arguments by condition in Experiment 2.

Experiment 3: Diversity

Experiment 3 aims at evaluating the impact of audience diversity on argument evaluation. If an audience accepts an argument, the more diverse the audience is, the more it constitutes a clue that the argument is good.

Design

This experiment had a between participants design with two conditions: Diversity High and Diversity Low.

Materials

In the Diversity High condition, participants read the following text:

> Rob, a scientist, has a new idea to improve the way science is conducted. He presents this idea and makes an argument for it in front of an audience of twenty scientists—physicists, biologists, chemists, linguists, psychologists. After the argument nearly all of them are convinced that Rob's idea is right.

In the Diversity Low condition, "scientists—physicists, biologists, chemists, linguists, psychologists" was replaced by "physicists." People were then asked a single question: "Do you think that Rob's argument is," which they could answer on a scale from 1 to 7 (1 Bad, 2 Average, 3 Good, 4 Very good, 5 Powerful, 6 Very powerful, 7 Incredibly powerful) or they could answer "I cannot tell." The scale was used to obtain a finer grained assessment of participants' evaluations, and it was skewed towards positive evaluation in order to avoid ceiling effects.

Results and discussion

First, there was no significant difference in the number of 'I cannot tell' answers across conditions (Fisher exact test, *ns*). Looking only at the other answers, the answers in both conditions averaged around 6 (very powerful): Diversity High, $M = 5.9$, $SD = 1.00$; Diversity Low, $M = 6.0$, $SD = 0.82$ ($t(83) = 0.61$, $p = .35$). It seems as if participants do not take diversity into account when evaluating arguments. Yet this result could stem from a ceiling effect, with most participants being reluctant to describe an argument as 'Incredibly powerful'. Another possible confound is that participants could have judged the physicists to be the most competent group, and their competence would have offset the lack of diversity.

Experiment 4: Diversity, motivation and audience size

Experiment 4 is an attempt at correcting the ceiling effect that render the interpretation of the results from Experiment 3 difficult, as well as the potential confound linked to competence. To do so, it introduces a Low Motivation condition in which the audience agrees with the arguer to start with. As a result, the argument should be deemed to be less good overall, keeping the competence of audience members neutral. It also introduces a new variable, audience size, which could be used as a clue by participants to evaluate arguments.

Design

The design was a 2*2*2. The first two factors were Diversity (High or Low), Motivation (High or Low), both varied between participants, and the third Audience Size (Large or Small), varied within participants.

Materials

Participants read variants on the following text (here the Audience Size Large-Diversity High-Motivation High condition):

> Rob, an Australian politician, makes an argument in front of an audience of a hundred people. They are members of several political parties, all of which oppose Rob's position. After the argument nearly all of them are convinced that Rob is right.

In the Audience Size Small conditions, the audience comprised 10 people. In the Diversity Low condition, "They are members of several political parties" was replaced with "They are members of the same political party." Finally, in the Motivation Low condition, "all of which oppose Rob's position" was replaced by "all of which agree with Rob's position." Question format was the same as Experiment 3.

Results and discussion

Experiment 4 is consistent with the results of Experiment 1: significantly more participants answered 'I cannot tell' in the Motivation Low (26%) than in the Motivation High (4%) condition (Fisher exact test, df = 1, $p < .001$). Regarding the 'I cannot tell' answer, no other variation by condition or interaction was significant. That only a minority of participants chose to answer 'I cannot tell' in this experiment, as opposed to a majority in Experiment 1, can be explained by the increased difficulty of the task facing the arguer, as he has to convince a much larger audience.

If we look only at the results excluding the 'I cannot tell' answers (Table 3 and Figure 1), we observe two main effects. The first is the effect of motivation: participants judged the argument to be less good when the motivation of the audience was low ($t(157) = 4.6$, $p < .001$). This effect complements the increase in 'I cannot tell' answers, and confirms that audience motivation is an important factor for argument evaluation. The second significant main effect is that of size: participants judged the argument that convinced a larger audience to be better ($t(157) = 2.7$, $p < .01$). Diversity again did not give rise to a main effect ($t(157) = 0.6$, $p = .53$). It failed to interact with motivation (2-way ANOVA, $F(1,158) < 1$) and it also failed to interact with group size (2-way ANOVA, $F(1,158) = 3.7$, $p = .056$). These results confirm those of Experiment 3 that participants fail to take audience diversity into account when gauging the value of arguments. The ceiling effect that rendered the interpretation of Experiment 3 difficult was removed here, as even in the Motivation Low condition, in which the average argument rating was 4.7 out of 7, diversity had no effect on argument rating. It should be noted however that the significant effect of size, as opposed to diversity, could be due to the fact that size was manipulated within participants while diversity was manipulated between participants. This may have made the size variable more salient to participants. More experiments will be necessary to tease out the effects of size and diversity.

	Motivation High		Motivation Low	
	Diversity High	Diversity Low	Diversity High	Diversity Low

Size Large	6.23 (1.27) (N=22)	6.14 (0.77) (N=24)	5.71 (1.10) (N=21)	4.64 (1.86) (N=14)
Size Small	5.79 (1.18) (N=22)	5.95 (1.50) (N=21)	3.88 (1.71) (N=16)	4.32 (1.97) (N=19)

Table 3: Mean evaluation of arguments by condition in Experiment 4, showing potential interactions (*SD* in parentheses).

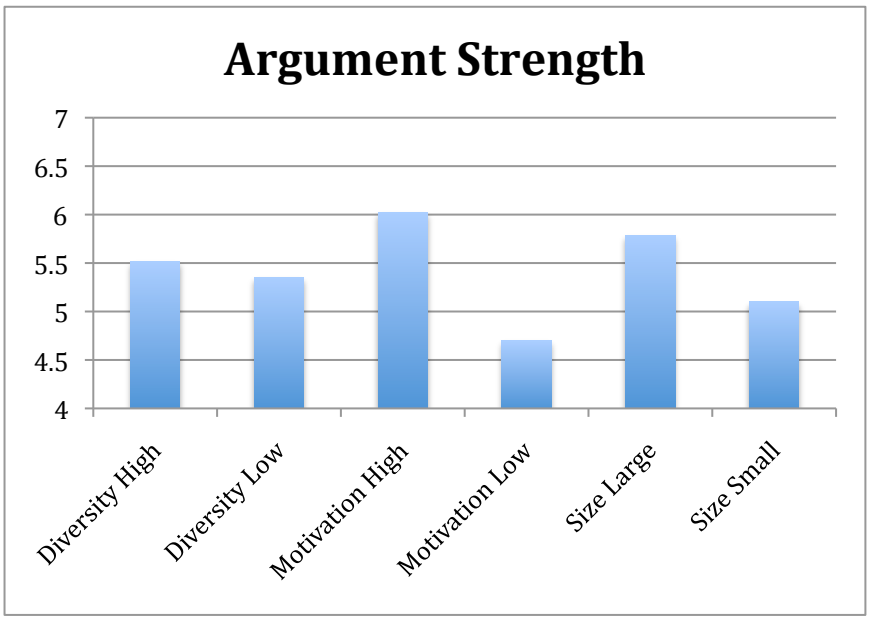

Figure 1: Evaluation of arguments by condition in Experiment 4, showing main effects.

Conclusion

The main way in which people evaluate arguments is through their form and content. It is therefore natural that experimental psychology has dedicated most of its efforts to understanding the processes that underlie such evaluation. But arguments can also be evaluated by gauging the reaction of their audience. For instance if a motivated,

expert audience is swayed by an argument, it is a reasonably good clue that the argument was a good one, and that maybe we should also accept its conclusion. The experiments above show that people are able to take such clues into account.

Experiment 1 demonstrated that participants are sensitive to the motivation of the audience: if the audience agrees with the arguer to start with, or if it does not pay attention to the argument, its reaction is discounted. Experiment 2 showed that participants take into account the relative expertise of the arguer, the audience, and third parties. A majority of participants found the argument to be good only when a novice found a counterargument to an argument that had convinced an expert. In all other cases participants found the arguments to be poor or refused to judge them. This fits with the idea that people are cautious: a reason to doubt that the argument is good—such as a disagreement between two equally competent listeners—is enough to stop them from accepting it. In Experiment 3, participants failed to take into account audience diversity in their evaluation of arguments. Two potential confounds were ruled out in Experiment 4, which confirmed that diversity is not easily taken into account. This null effect does not reflect lack of power for Experiment 4 replicated the effect of motivation observed in Experiment 1 and demonstrated that participants can take audience size into account. It should be noted, however, that if a modicum of diversity is necessary for audience size to make a difference in argument evaluation, this result could mean that diversity is in fact taken into account by participants, even if by a circuitous route.

The negative result regarding the role of diversity may be a promising avenue for future research. Diversity has been shown to be a crucial determinant of collective intelligence (Hong & Page, 2004). Through simulations, Hong and Page have shown that

diversity often trumps competence: as long as group members are minimally competent, an increase in diversity brings more benefit to the problem solving potential of the group than an increase in competence. The counterintuitiveness of this result may reflect a general failure to appreciate the cognitive benefits of diversity, a failure that may also explain the results of Experiments 3 and 4. But before drawing too bleak a conclusion from our results, it should be pointed out that they clash with a large literature showing that adults and even children routinely take diversity into account in a variety of tasks (see, for instance, Hahn, Bailey, & Elvin, 2005; Osherson, Smith, Wilkie & Lopez, 1990; Heit & Hahn, 2001). A possible explanation for our discrepant findings is that participants may have inferred that when the audience was more diverse, the scope of the arguments was wider, which would make the expertise of the various audience members potentially less relevant to the evaluation of the argument.[2] More experiments will have to be conducted before we can conclude that participants really discount diversity when they evaluate arguments from the reaction of the audience.

Whatever clues they take into account, people are often called to use audience reaction in the evaluation of arguments. This article is a first step towards a better understanding of the processes underlying this type of evaluation.

Acknowledgements

We thank Frank Keil and the members of his lab meeting for a most useful discussion and Matt Fisher for his feedback. We also thank the editors, Ulrike Hahn in particular, who considerably improved the manuscript.

[2] We thank an anonymous Reviewer for this suggestion.

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, *6*(1), 3.

DeScioli, P., & Kurzban, R. (2009). The alliance hypothesis for human friendship. *PloS one*, *4*(6).

Evans, J. S. B. T. (2002). Logic and human reasoning: an assessment of the deduction paradigm. *Psychological bulletin*, *128*(6), 978-996.

Hahn, U., Bailey, T. M., & Elvin, L. B. C. (2005). Effects of category diversity on learning, memory, and generalization. *Memory & cognition*, *33*(2), 289–302.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704-732.

Harris, A. J. L., Hsu, A. S., & Madsen, J. K. (in press). Because Hitler did it! Quantitative tests of Bayesian argumentation using *ad hominem. Thinking and Reasoning.*

Heit, E., & Hahn, U. (2001). Diversity-Based Reasoning in Children* 1. *Cognitive psychology*, *43*(4), 243–273.

Hoeken, H., Timmers, R. & Schellens, P. J. (current issue) Arguing About Desirable Consequences: What Constitutes a Convincing Argument? *Thinking and Reasoning.*

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385.

Kant, I. (1999). *Critique of pure reason*. Cambridge: Cambridge University Press.

Mercier, H. (in press). Our pigheaded core: How we became smarter to be influenced by other people. In B. Calcott, R. Joyce, & K. Sterelny (Eds.), *Evolution, Cooperation, and Complexity*. Cambridge: MIT Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57-74.

Millikan, R. G. (1987). *Language, Though and Other Categories*. Cambridge: MIT press.

Osherson, D., Smith, E.E.,Wilkie, O, & L´opez, A (1990) Category-based induction. *Psychological Review*, 97(2):185–200.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5).

Perelman, C., & Olbrechts-Tyteca, L. (1958). *The New Rhetoric: A Treatise on Argumentation*. Notre Dame, IN: University of Notre Dame Press.

Petty, R. E., & Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 323–390). Boston: McGraw-Hill.

Popper, K. R. (2002). *The logic of scientific discovery*. London: Routledge.

Van Eemeren, F. H., Garssen, B. & Meuffels, B. (current issue). The disguised abusive *ad hominem* empirically investigated strategic maneuvering with direct personal attacks. *Thinking and Reasoning*.

Walton, D. N., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge: Cambridge University Press.