

## Chapter 11

# The Biased Use of Argument Evaluation Criteria in Motivated Reasoning: Does Argument Quality Depend on the Evaluators' Standpoint?\*

*Hans Hoeken<sup>1</sup> & Mariecke van Vugt*

Utrecht Institute for Linguistics OTS, Utrecht University, Utrecht, The Netherlands, <sup>1</sup> j.a.l.hoeken@uu.nl

**Abstract.** People without a background in argumentation theory possess several criteria to distinguish strong from weak arguments. The fact that people have these criteria does not imply that they will use them to objectively assess the quality of an argument. Research on motivated reasoning suggests that people take a more critical stance toward arguments that go against their opinions compared to arguments that are in accordance with these opinions. In this study, the question was addressed whether people employ criteria to evaluate arguments in a biased way. Forty participants were told that they would take part in a debate and either had to defend the claim that mixed schools (that is, schools attended by children with different ethnic backgrounds) were desirable or the claim that these were undesirable. All participants received sixteen (strong and weak) arguments and were asked to prepare themselves for the debate while thinking aloud. Analysis of the think aloud protocols showed that people almost exclusively used criteria to boost the quality of arguments supporting their claim while disqualifying arguments that went against it. These results provide important insights into the nature of motivated reasoning because they show how people deploy argument criteria in this process.

### 1. Introduction

Mercier and Sperber (2011) claim that there is a big difference between situations in which people have to evaluate the acceptability of a claim and its supporting arguments and situations in which they themselves have to defend a certain claim or already have an opinion on the issue at hand. In the former situation, people are hypothesized to assess the quality of the arguments in a relatively objective way whereas in the latter situation, they are predicted to evaluate the arguments in a more biased manner. Biased evaluation implies that arguments in accordance with their opinion or preference are perceived as stronger whereas counterarguments are considered weaker, regardless of the arguments' objective

---

\* A version of this paper has been published in Dutch: Hoeken, H., & Vugt, M. van (2014). Het bevooroordeelde gebruik van argumentatieschema-specifieke criteria. *Tijdschrift voor Taalbeheersing*, 36, 87-105. ISSN: 1573-9775, DOI: 10.5117/TVT2014.1.HOEK. The study reported here was conducted by the second author as part of her master thesis for the master Communication & Persuasion at the Radboud University, the Netherlands. The first author served as her supervisor.

characteristics. Mercier and Sperber explain this asymmetry by stating that this leads to an optimal division of labor: Supporters of a certain claim have to develop arguments in favor of their claim, opponents have to come up with counterarguments, and the ultimate judge of whose arguments are the most convincing ones is the audience. If the audience members do not already have an opinion on the issue, they are well-equipped to assess and weigh the reasonableness of the arguments provided by the debaters.

Schellens and De Jong (2004) state that for people to evaluate argument quality, they need to possess argument scheme specific criteria. That is, to assess the quality of an argument from authority, other criteria are relevant than to assess the quality of an argument from analogy or an argument from example. Recent studies have shown that lay people without special training in the field of argumentation theory possess and apply a number of such criteria when distinguishing strong from weak arguments. These findings appear to support Mercier and Sperber's (2011) claim that people are relatively well versed in argument evaluation. In this paper, we focus on the second part of their prediction, namely that people will interact with arguments in a biased way when they have a preference for a claim. In the next paragraphs, we will discuss studies on the extent to which people are capable of evaluating arguments as well as the research on motivated reasoning. The latter research shows how people go about when evaluating arguments that go against their own opinions.

### *1.1 Evaluating Arguments Employing Argument Scheme Specific Criteria*

According to dual-process models of the persuasion process, such as the Elaboration Likelihood Model (ELM: Petty & Cacioppo, 1986) or the Heuristic-Systematic Model (HSM: Chaiken, 1987), people can form (or change) their attitudes as a result of a careful evaluation of arguments. If people arrive at an attitude in such a way, this attitude is believed to be more stable, less susceptible to counter persuasion attempts, and a better predictor of attitude related behavior (Petty, Haugtvedt, & Smith, 1995). These models state that people need to be both motivated to and capable of carefully evaluating arguments in order for this to occur. Scrutinizing arguments takes time and effort and people will only be willing to invest these if they find it important to have a correct attitude. People will be more inclined to carefully evaluate the arguments if accepting the claim has far reaching consequences (e.g., when buying a house), compared to decisions with only minor implications (e.g., when buying toothpaste). In general: the more people consider an issue as relevant to their own and/or their loved ones' well-being, the more they will be inclined to set out for a careful evaluation of the arguments.

Being motivated is not sufficient in itself; people also need to be capable of evaluating arguments. Schellens and De Jong (2004) have shown what capacities this requires. They analyzed twenty Dutch public information brochures for the general public in which certain behaviors were either discouraged (e.g., smoking, gambling) or encouraged (e.g., using sun tan lotion, exercising). They have shown that for a careful evaluation of the arguments, people need to take three steps. First, in the majority of the analyzed brochures, there are no explicit arguments. Arguments are presented implicitly in the guise of factual information. For instance, in a brochure on exercising, positive consequences of exercising are presented without explicitly stating that these consequences are arguments in support of the claim that exercising is good. As a result, people should be able to identify which information serves as an (implicit) argument. Second, Schellens and De Jong (2004) show that different types of argument are used in the brochures, such as arguments from authority, arguments from example, and arguments from analogy. For a critical evaluation of such arguments, argument scheme specific criteria are needed. For instance, when scrutinizing an argument

from analogy, the similarity between the two cases that are being compared is important, whereas the number and representativeness of examples is at stake when evaluating an argument from example. A careful evaluation of arguments therefore requires two additional skills: identification of the argument scheme and the possession of and ability to apply scheme specific criteria.

People appear to be able to distinguish strong from weak arguments. In many experiments, argument quality has been manipulated and the attitude after reading the message has been measured. Meta-analyses have revealed that, in general, people are more convinced by messages containing strong arguments than by messages containing weak arguments (Carpenter, 2015; Johnson, Smith-McLallen, Killeya, & Levin, 2004; Park, Levine, Kingsley Westerman, Orfgen & Foregger, 2007). Both the meta-analyses and the individual studies reported on in these meta-analyses, do not shed light on the criteria people use to distinguish strong from weak arguments. The manipulation of argument quality is usually based on the intuitions of individual researchers rather than on theory guided normative criteria. As a result, strong arguments differ on many dimensions from their weak counterparts. For instance, Van Dijk-Van Enschot, Hustinx, and Hoeken (2003) have shown that Petty and Cacioppo's (1986) manipulation of argument quality has led to their strong arguments differing on various dimensions compared to their weak arguments. Different consequences are referred to in the strong arguments compared to the weak ones, and the evidence in support of the likelihood that the consequences will occur, does also differ. As a result, it is not possible to link differences in claim acceptance or attitudes to specific characteristics of these arguments. Therefore, it is not possible to identify the criteria participants in these studies have employed to distinguish strong from weak arguments.

There are two lines of research that provide insights into the extent to which people are capable of applying argument specific criteria. Van Eemeren, Garssen, and Meuffels (2009) report on a series of experiments in which they showed that arguments that do not meet the standards of reasonableness were perceived as less reasonable by ordinary language users. For instance, in one of their studies, they had participants rate the reasonableness of an argumentative move within a conversational context. One of the conversational partners attacked the credibility of the other person. In one condition, this could be considered a reasonable move whereas in other conditions, it could be considered as an instance of an 'ad hominem' fallacy. In the former condition, participants rated the move as more reasonable compared to the reasonableness of the moves in the latter conditions. This research shows that people can distinguish a fallacious move from a non-fallacious one.

In a different line of research, it is studied to what extent people are more convinced by arguments that meet specific argument scheme criteria than by those that do so to a lesser extent. For instance, in case of an argument from example, the number of examples is relevant. Several studies have shown that a claim is accepted more strongly when supported by a statistical summary of a number of cases compared to a single case (Hoeken & Hustinx, 2009; Hornikx & Hoeken, 2007). Even adding one additional example led to a stronger claim acceptance (Hoeken, Šorm, & Schellens, 2014). In the latter study, it was also shown that the representativeness of the example in the argument, had a positive impact on claim acceptance. These studies thus show that people are sensitive to two normative criteria for the quality of an argument from example: number and representativeness.

A similar sensitivity to normative criteria is found for the argument from analogy. The most important criterion for this type of argument is the extent to which the two cases that are being compared can be considered similar. Both Hoeken and Hustinx (2009) and Hoeken, Timmers, & Schellens (2012) have shown that the more similar the two cases are, the more people are willing to accept the claim supported by the argument from analogy.

Sensitivity to argument scheme relevant criteria also has been documented for the argument from authority (Hoeken *et al.*, 2012, 2014; Homikx & Hoeken, 2007), the argument from cause to effect (Hoeken *et al.*, 2014), and the argument from consequences (Hoeken *et al.*, 2012). For all of these argument schemes it was found that people accepted the claim more strongly, if the argument met relevant, normative criteria.

## *1.2 Biased Evaluation of Arguments*

The fact that people possess relevant criteria for argument evaluation does not guarantee that they will use those to objectively assess the merits of the argument at hand. The concept of motivated reasoning (Kunda, 1990; Molden & Higgins, 2005; Westen, Blagov, Harenski, & Hamann, 2006) refers to the biased evaluation of arguments: arguments that are in line with one's opinion are evaluated less critically than those that go against it. It is an important part of the so-called confirmation bias phenomenon (see, for a review, Nickerson, 1998). This broader phenomenon not only includes the biased evaluation of arguments but also the selective seeking of information that corroborates one's opinion.

Mercier and Sperber (2011, p.76) state that "people are good at assessing arguments and are quite able to do so in an unbiased way" but also attach a precondition for such an unbiased assessment: "provided they have no particular axe to grind". Mercier and Landemore (2012, p. 251) claim that "the confirmation bias mostly affects the production, and not the evaluation of arguments". The production of arguments is guided by a persuasive goal, that is, to have another person accept one's standpoint (Mercier, 2012, p. 317). As a result, the production is geared to finding arguments that will support the individual's standpoint (or rebut the opponent's arguments). Argument evaluation, on the other hand, "aims at distinguishing good arguments from bad ones, and hence genuine information from misinformation" (Mercier & Sperber, 2011, p. 72). However, if people do have an axe to grind, that is, they already have an opinion on a topic, the evaluation of arguments may display motivated reasoning as well.

Lord, Ross, and Lepper (1979) were one of the first to provide empirical evidence for the existence of motivated reasoning. They had participants read two studies on the deterrence effect of capital punishment. In one study, the results appeared to provide support for this effect whereas the other study's results failed to do so. Some of their participants were in favor of capital punishment whereas others were against it. Those in favor of capital punishment were much more critical about the quality of the study that did not find positive effects of capital punishment on crime rates compared to the study reporting such an effect. Exactly the opposite pattern of results was obtained for the opponents of capital punishment. In addition, despite the fact that both groups were presented exactly the same information, the supporters felt more certain about the positive effect of capital punishment and the opponents about the absence of such an effect. Several other studies replicated these findings (e.g., Ditto & Lopez, 1992; Edwards & Smith, 1996).

In the studies described above, participants were already in favor of, or against, capital punishment before they entered the experiment. As a result of this difference in attitude, supporters and opponents probably also differed with respect to their knowledge on the issue. This prediction follows from the existence of a confirmation bias, which would lead those in favor of capital punishment to seek out and remember information that is in line with their opinion on the capital punishment's deterrence effect whereas the opposite will hold for the opponents of capital punishment. One of the strategies that people apply when evaluating arguments is coherence checking. That is, people activate their previously held beliefs and check to what extent the new information is accordance with these beliefs (Mercier & Sperber, 2011, p. 60). If incoherencies come to the fore, people have to choose

between revising their previous beliefs or rejecting the new information. If an opponent of capital punishment reads a study reporting deterrence effects of the death penalty, incoherencies are evoked and he or she is well-equipped to bring up counterarguments against the assumptions and interpretation of this study; the same holds for supporters of capital punishment who can use their reservoir of facts and beliefs to tackle the study reporting the absence of a deterrence effect. In sum, the difference in knowledge accompanying the difference in attitude may have made it more easy for both supporters and opponents of the death penalty to criticize the studies reporting results that go against their views.

Jain and Maheswaran (2000) conducted two experiments in which they manipulated instead of observed the participants' standpoint. That is, in the first stage of the experiment, participants received either positive or negative information about a telephone answering machine. In this way, prior attitude was manipulated instead of observed. Because the experiments were conducted immediately afterwards, the participants could not seek out and store information relevant to their preferences. As a result, they differed in prior attitude, but the differences with respect to prior knowledge were limited. In both experiments, it was found that participants generated more counterarguments when confronted with information that went against their initial attitude compared to when new information was in line with that attitude. So, even if people's preference is manipulated instead of observed, it still leads to a biased evaluation of subsequent arguments that go against that preference. Still, the participants differed in their knowledge about the product which may have enabled them find more flaws in the arguments that went against their attitude.

The results reported by Jain and Maheswaran (2000) do not enable an assessment of the extent to which people use argument scheme specific criteria to critically evaluate arguments. Klaczynski, Gordon, and Fauth (1997) provided some insight into this issue. They studied to what extent participants employed normative criteria when evaluating arguments that went against their interests. Students received arguments in support of the claim that the teaching program they were in either provided excellent or very meager prospects at the job market. The arguments were not very strong. For instance, claims about the job prospects of the program as a whole were backed up by the case of a single alumnus who was either very successful or very unsuccessful in getting a job. Participants were asked to rate the arguments and to provide an explanation of their ratings. The results showed that arguments were rated as stronger when they were in support of positive prospects on the job market whereas they were rated as weaker when they were in support of negative prospects. Analyses of the explanations revealed that participants employed argument scheme specific criteria when evaluating these arguments. For instance, they referred to the danger of hasty generalization when the meager job prospects claim was supported by the experiences of a single alumnus. In four experiments, Klaczynski *et al.* showed that people thought longer, deeper and more critical when confronted with arguments that went against their vested interest compared to when the arguments were in favor of their interests. Apparently, people use normative criteria to refute displeasing arguments.

Nienhuis, Manstead, and Spears (2001) showed that under certain conditions people who have to defend a claim, do evaluate arguments in an objective manner. In three experiments, they had participants read a text in which a certain claim was defended, for instance, "hard drugs should be legalized". There were two versions of the text: one containing strong and one containing weak arguments. Some participants were told that in a subsequent part of the study, they would have to convince another participant of the claim's acceptability and that they themselves would be evaluated with respect to their task performance. After they had read the text, participants were asked to write down all

thoughts they had while reading the text. Participants who had read the text containing strong arguments generated more thoughts that were favorable with respect to the issue than those who had read the text containing weak arguments. The favorableness of the thoughts listed by participants who had not received this instruction did not differ as a result of the quality of the arguments in the text. In addition, the attitude toward the claim of the former group was more positive after reading the version with the strong arguments. An important difference between the Nienhuis *et al.* study and the one by Klaczinsky *et al.* (1997) was that in the former study participants did not have a strong opinion on the issue whereas in the latter study they immediately had one (i.e., it was about their chances on the job market). This raises the question to what extent motivated reasoning is mainly evoked when people have a vested interest in an issue.

### 1.3 Research Questions

Several studies have shown that people possess and use normative criteria to distinguish strong from weak arguments (see, e.g., van Eemeren *et al.*, 2009; Hoeken & Hustinx, 2009; Hoeken *et al.*, 2012, 2014; Hornikx & Hoeken, 2007). In addition, research has shown that people respond more critically toward arguments that go against their existing opinions or vested interests (Ditto & Lopez, 1992; Edwards & Smith, 1996; Lord *et al.*, 1979). Klaczinsky *et al.* (1997) provided evidence that people use normative criteria when evaluating arguments in a biased fashion. In these studies, the participants already held these opinions or had a vested interest when they entered the experiments. That is, they were already in favor (or against) the death penalty and they had an interest of being in a program that would give them excellent opportunities on the job market. As a result of these existing preferences, they may have also differed in their knowledge about these issues which may have made them better equipped to spot the weaknesses in (strong) arguments that ran counter their opinion and repair the flaws in weak arguments in favor of their position. Jain and Maheswaran (2001) manipulated instead of observed the participants' opinion. However, their manipulation consisted of providing participants with different information which may also have enabled participants to spot weaknesses or incompatibilities between the first batch of information and the second. In addition, it is unclear whether the participants in the Jain and Maheswaran study employed normative criteria when evaluating the arguments in a biased way. Finally, Nienhuis *et al.* (2001) have shown that people who have to defend a claim are not necessarily blind to differences in argument quality. In their study, however, participants were only exposed to arguments in support of the claim they had to defend.

In this paper, a study is reported in which we had participants evaluate arguments ostensibly as a preparation for a discussion in which they had to defend a certain claim. This instruction resembles the one used by Nienhuis *et al.* (2001) but in this study we provided participants with normatively weak and strong arguments in favor of the claim as well as weak and strong arguments against the claim the participant had to defend. In addition, we asked participants to verbalize their thoughts while preparing for the debate in order to assess whether or not they employed argument specific criteria when evaluating the arguments. This set up enabled us to address the question to what extent people respond more critically towards arguments that go against the claim they have to defend compared to arguments that are in line with that claim. This general question can be subdivided in a number of more specific questions. First, it is interesting to assess how people evaluate arguments in such a context. It is unclear whether they employ specific criteria in this process. Therefore, the first question is:

1. Do people use argument scheme specific criteria while evaluating arguments in preparation of a debate?

Research on motivated reasoning has shown that people pay more attention to arguments that go against their opinion compared to arguments that are in line with it (see, e.g., Klaczynski *et al.*, 1997). It is unclear whether more attention implies a more frequent use of criteria. This raises the second question:

2. To what extent do people deploy argument scheme specific criteria more often when evaluating arguments that go against the opinion they are instructed to defend compared to arguments that are in line with it?

Nienhuis *et al.* (2001) showed that people distinguish strong from weak arguments when they expect to have to convince someone else of an opinion; research on motivated reasoning claims that people are more likely to seek weaknesses in arguments against their opinion compared to arguments that are in line with it. This discrepancy leads to the third question:

3. To what extent do people deploy argument scheme specific criteria to distinguish strong from weak arguments in an objective way or to what extent they use them in a biased way, that is, disqualify arguments that go against the opinion they have to defend and boost the quality of arguments that are in line with that opinion?

With this study, we hope to shed more light into the exact nature of motivated reasoning and the conditions under which it is likely to occur.

## 2. Method

### 2.1 Materials

Sixteen arguments were developed relevant to the issue of “mixed schools”. This issue pertains to the question as to whether it is desirable or undesirable that pupils in primary schools come from (many) different ethnic backgrounds. Eight arguments supported the claim that mixed schools are desirable while the other eight were in support of the claim that they are undesirable. For each of these sets of eight arguments, half did meet argument scheme specific criteria, which made them strong from a normative point of view, whereas the other half did not meet these criteria, which made them weak from a normative point of view. In summary, there were four strong and four weak arguments in favor of mixed schools, as well as four strong and four weak arguments against mixed schools.

Four different argument schemes were employed: argument from analogy, argument from authority, argument from example, and argument from cause to effect. Argument quality was manipulated by having arguments meet to a stronger or lesser extent criteria for which it had been established in previous research that people are sensitive to. For the argument from analogy, cases were selected that either were more similar to the Dutch situation at hand (e.g., “In Sweden, the introduction of mixed schools has led to better results”) or less similar (e.g., “In France, mixing people with different ethnic backgrounds on the work floor has led to better results.”). For the argument from authority, sources were selected that had no vested interest in the issue (e.g., Peter den Boer, renowned child



pedagogic) or ones who had such an interest (e.g., Annieck ten Haven, director of a mixed school). For the argument from examples, the representativeness of the example was manipulated (e.g., "Lisa de Bruin performs better since she moved to a mixed school" vs. "Metab Rakkech performs better since she moved from an all black school to a mixed school"). Finally, for the argument from cause to effect, the plausibility of the causal relation was manipulated. A more plausible relation would be "Mixing will lead to a reduction of social and cultural differences" compared to "Mixing will lead the management of the school to pay attention to the educational policy". The quality of arguments in favor of mixed schools as well as those against mixed schools were manipulated in this way.

## ***2.2 Research Design and Procedure***

Participants were approached by the second author to ask whether they were willing to take part in a study. If the participant agreed, the experimenter told the participant that the study was about the way in which people prepare for a debate and how they debate. Next, they received the sixteen arguments about the issue of mixed schools. Participants were randomly assigned to either the condition in which they were told that they had to defend the claim that mixed schools are desirable or the condition in which they had to defend the opposite claim. The experimenter explained that in the first stage of the study, the participant would have to prepare for the debate by evaluating the provided arguments while thinking aloud. This part of the session would be recorded. After transcribing the recording, the data would be anonymized. Upon completing this part, the participants were told the aim of the study and it was explained to them that there would be no debate. Any remaining questions were answered. A session lasted 30 minutes on average.

## ***2.3 Participants***

Forty participants took part in the study. Age varied from 21 to 60 with a mean of 35 years. Slightly more women (22) than men (18) participated. Level of education varied from vocational studies (13), applied university (19), to a completed master's degree (8). There were no significant differences between the two conditions with respect to the participants' age, gender, and level of education ( $p$ 's > .52). At the end of a session, participants were asked whether they already had an opinion on the issue of mixed schools. The majority (28) held a neutral stance towards the issue, five held a strong favorable attitude, six held a slightly favorable attitude on the issue; only one participant held a (slightly) negative attitude towards mixed schools. Participants holding a prior attitude were almost equally distributed over the conditions.

## ***2.4 Data Processing***

The think aloud protocols were analyzed from the perspective of whether participants used argument scheme specific criteria and if they did, whether that led to a positive or a negative qualification of the argument's quality. For each participant, it was established how often he or she applied a criterion to qualify an argument as strong and how often application led to qualify an argument as weak. For each of the four sets of (four) arguments (i.e., in favor & strong, in favor & weak, against & strong, against & weak), the number of criterion based evaluations by a participant were computed. The resulting scores were analyzed using a 2 (Instruction: defend desirability, defend undesirability) x 2 (arguments in favor of claim to be defended, arguments against) x (strong arguments, weak



arguments)  $\times$  2 (criterion application leads to positive qualification or leads to negative qualification) Analysis of Variance was conducted with the final three factors being within-participants factors.

### 3. Results

The first question was whether participants would use argument scheme specific criteria when evaluating the arguments. This proved to be the case for each and every participant. With a minimum of 4 and a maximum of 13, the average number of criteria used by a participant was 6.58 ( $SD = 1.92$ ). First, the results of the quantitative analyses of the data will be presented, followed by presenting quotes from the think aloud protocols which reveal how participants employed these criteria. Table 1 contains the mean scores (and standard deviations) for the application of criteria for the different conditions.

	Supportive of claim		Opposing of claim	
	Strong	Weak	Strong	Weak
<b>Positive Evaluation</b>				
<i>In favor</i>	1.30 (0.73)	1.35 (0.93)	0.05 (0.22)	0.00 (0.00)
<i>Against</i>	1.50 (0.76)	0.80 (0.70)	0.00 (0.00)	0.00 (0.00)
<b>Negative Evaluation</b>				
<i>In favor</i>	0.00 (0.00)	0.00 (0.00)	1.90 (0.72)	2.15 (1.14)
<i>Against</i>	0.00 (0.00)	0.00 (0.00)	2.20 (0.77)	1.90 (0.91)

**Table 1.** The means and standard deviations for the number of times criteria for the evaluation of arguments were applied as a function of the claim the participant had to defend (in favor, against), argument relation with claim (supporting claim to be defended, opposing claim to be defended), argument quality (strong, weak), and evaluation of argument (positive, negative) (Minimum = 0, Maximum = 4)

Research question 2 was about whether people would employ argument scheme specific criteria to a greater extent when evaluating arguments that went against the claim they had to defend compared to arguments that were in line with that claim. This proved to be the case ( $F(1, 38) = 40.89, p < .001, \eta^2 = .52$ ). When the arguments went against the claim to be defended, participants applied normative criteria more often ( $M = 1.03, SE = .050$ ) compared to evaluating arguments that supported this claim ( $M = 0.62, SE = .049$ ). Participants were also more likely to deploy criteria to evaluate an argument as weak ( $M = 1.01, SE = .051$ ) than to evaluate an argument as strong ( $M = 0.63, SE = .050; F(1, 38) = 36.34, p < .001, \eta^2 = .49$ ). The nature of the claim (mixed schools are desirable vs. mixed schools are undesirable) had no effect on the number of criteria they applied ( $F(1, 38) < 1$ ), nor was there a main effect for argument quality ( $F(1, 38) = 1.79, p = .19$ ).

Research question 3 was about whether participants would use normative criteria to assess argument quality in an objective or in a biased way. If they had used them in an objective way, participants would use normative criteria mainly to classify strong arguments as strong and weak arguments as weak. This would result in an interaction between argument quality and evaluative judgment. However, this interaction was not significant ( $F(1, 38) = 1.79, p = .19$ ). If, on the other hand, participants used normative criteria in a biased way, they would use them to qualify arguments that went against the

claim they had to defend as weak, and the arguments that were in line with that claim as strong. In that case, the interaction between relation to claim (in favor, against) and evaluative judgment should become significant. This interaction was indeed highly significant ( $F(1, 38) = 445.24, p < .001, \eta^2 = .92$ ) and explained almost all of the variance in the data. Without exception, criteria were deployed to qualify arguments in support of the claim to be defended as strong ( $M = 1.24, SE = .098$ ), and – with one exception – to qualify arguments as weak that went against the claim to be defended ( $M = 2.04, SE = .10$ ).

Two other effects were significant as well. First, participants who had to defend the claim that mixed schools are desirable deployed normative criteria in equal numbers for strong and weak arguments whereas those who defended the claim that mixed schools are undesirable were more likely to use these criteria to evaluate strong arguments. This resulted in a significant interaction between claim (mixed schools are desirable, undesirable) and argument quality ( $F(1, 38) = 4.96, p = .03, \eta^2 = .12$ ). Given that application of criteria could lead to both a positive or a negative evaluation of the argument, this effect is not relevant for the research questions. Finally, the four way interaction also proved significant ( $F(1, 38) = 5.47, p = .03, \eta^2 = .13$ ). Three-way analyses of variance were conducted separately for people who had to defend the claim that mixed schools are desirable and people who had to defend the opposite claim. For the former group, no three-way interaction arose ( $F(1, 19) < 1$ ) whereas it did for the latter group  $F(1, 19) = 5.76, p = .03, \eta^2 = .23$ ). Given the small sample size, this interaction should be interpreted with caution, but it appears that the opponents deployed criteria more often to evaluate strong supporting arguments as strong than to evaluate weak supporting arguments as strong whereas they used criteria equally often to evaluate contradicting arguments as weak, regardless of argument quality.

The think aloud protocols provided more detailed information on how participants used criteria to evaluate arguments. A good example of how participants went about in tailoring the arguments to their purpose, is provided by the different evaluations of the same argument from analogy. One of the arguments in support of the desirability of mixed schools was that Sweden had favorable experiences with mixed schools whereas one of the counterarguments was that Russia had unfavorable experiences with mixed schools. Given that most people regard Sweden as more similar to the Netherlands than Russia, the first argument should be considered stronger than the second. Participants who had to defend the claim that mixed schools are desirable made remarks such as “If you look at how things work in other countries, then that will work for the Netherlands as well. Sweden makes a good comparison for the Netherlands” and “Sweden is in many respects comparable to the Netherlands, so it is a good comparison”. When evaluating the Russia argument, participants made remarks such as “The organization of the educational system in Russia is so saddening, that whatever their experience is, it says nothing about the Netherlands”. So in both cases, the criterion of “sufficient similarity” is used to evaluate these arguments. Participants who had to defend the opposite claim, also applied the same criterion but reached opposite evaluations. For the Sweden argument, they pointed to dissimilarities: “Sweden is not a good example because there are far fewer people with a different ethnic background, therefore the mix will also be less extreme” and “Sweden is very different from the Netherlands. Just because mixed schools are effective in Sweden does not imply that it will work in the Netherlands as well”. They also used this criterion to evaluate the Russia argument as strong: “Russia is a country with many ethnic differences and therefore a perfect gauge for the Netherlands” and “It has been proven that it doesn’t work in Russia so why would it work in the Netherlands? Russia has, similar to the Netherlands, many different cultures”.

A similar strategic employment of criteria was also observed for the other argument schemes. For instance, for the argument from authority, proponents reacted very differently

compared to opponents when evaluating the argument that a renowned child pedagogic held a positive opinion on mixed schools. Whereas proponents were positive ("The fact that a child pedagogic says something about it, that it's good, that proves something. It's not a lay man who says so."), opponents questioned the source's authority ("On what grounds does he claim that? He can be a child pedagogic, but that doesn't make him automatically an expert on the issue of mixed schools"). When a director of a mixed school expressed her opinion on the desirability of mixed schools, opponents pointed out that she had a vested interest in expressing that claim: "Director of a mixed school, so she's definitely biased. She will never say that a mixed school is bad. You never say about your own school that it performs worse." Opponents, on the other hand, qualified this argument as strong, not through focusing on the reliability criterion but by focusing on the criterion of relevant expertise: "A director has knowledge about the results and thus knows a lot about the effects of a mixed school. She can provide a well-founded statement about mixed schools."

#### 4. Conclusion and Discussion

Without exception, all participants in this study used argument scheme specific criteria to evaluate arguments. Previous research has shown that people spend more time evaluating arguments that went against their initial opinion than arguments that were in line with it (Klaczynski *et al.*, 1997). The results of our study replicated that finding and in addition showed that this additional attention also implied the more frequent application of normative criteria. Finally, the results clearly provided a pattern consistent with the concept of motivated reasoning: rather than using normative criteria to distinguish strong from weak arguments, participants used them to disqualify arguments that went against the claim they had to defend and, to a lesser extent, glorify the arguments that were in line with this claim. The results of the think aloud protocols showed how people who had to defend different claims used the exact same criterion on the exact same argument to reach opposite conclusions about the argument's quality.

This study provides an important addition to our understanding of motivated reasoning for four reasons. First, we manipulated the claim the participants had to defend instead of using already established opinions. As a result, differences in prior knowledge could not serve as an alternative explanation for the results. Second, we provided participants with both supporting and counterarguments. Because participants differed systematically with respect to the claim they had to defend, the exact same argument was a supporting argument in one condition while being a counterargument in the other condition. As a result of this design, we could establish for each argument how it was treated when considered as a supporting argument and when considered as a counterargument. Third, the quality of the arguments was manipulated systematically employing criteria for which it had been established in previous research that lay people could use them. This enabled a much more controlled sample of arguments for participants to react upon. Finally, by having participants express their thoughts verbally while evaluating the arguments, we were able to assess online the extent to which participants used normative criteria and identify which specific criteria they used.

As cited in the introduction, Mercier and Landsmore (2012) predict that the confirmation bias is more likely to occur in the production of arguments than in their evaluation. In our study, participants were evaluating arguments instead of producing them, and still their evaluation pattern revealed a strong confirmation bias. However, this finding does not go against the predictions of Mercier and Sperber's (2011) argumentative theory of reasoning. Mercier (2012) describes how people may go about when finding arguments

to defend their position in a discussion. From his description it becomes clear that there is a lot of evaluation going on when arguments are produced: when looking for arguments to use in a discussion, potential arguments are evaluated in a serial manner. This evaluation process is geared by a strong confirmation bias in the sense that “positive consequences of one’s proposition, as well as negative consequences of one’s interlocutor’s proposition, often make good arguments, while the converses don’t” (Mercier, 2012, p. 319). In our study, participants did not have to generate arguments themselves; these were presented to them. However, Mercier (2012) suggests that the confirmation bias manifests itself in the evaluation of the arguments, not in their generation. The evaluation patterns of the participants in our study confirm Mercier’s prediction.

Participants were instructed to defend a claim on an issue the vast majority did not have strong opinions on. This context may have created a mindset in which they considered the task a game without real-life consequences. As a result of this playful mindset, their evaluation patterns may have become more extreme and less representative of how they would act under normal circumstances. It would therefore be interesting to see how participants would evaluate these same arguments when instructed to form their own opinion on this issue, an opinion that they would have to defend in a debate with another person. Under those conditions, it is more likely that participants would use the criteria to distinguish strong from weak arguments. Mercier and Sperber (2011) claim that we have immediately intuitions about the validity of a belief. These intuitions may subsequently guide our evaluation of the arguments. To assess the validity of this prediction, one could unobtrusively manipulate the participants’ attitude towards an issue, for instance, through conditioning or priming and subsequently assess whether participants evaluate the arguments in the same biased way as reported in this study.

In conclusion, the results of this study provide more insight into the exact nature of motivated reasoning by revealing that people employ normative criteria in a biased way. The study’s design, having participants evaluate systematically manipulated arguments while thinking aloud, provides interesting opportunities to further explore the conditions under which motivated reasoning occurs as well as assess predictions that can be derived from Mercier and Sperber’s (2011) argumentative theory of reasoning. Apart from the scientific implications, this study has also practical consequences. It is believed that attitudes and opinions based upon a careful evaluation of relevant arguments yield a more stable attitude that is more predictive of subsequent behavior. Implicitly, the idea is that argument quality will make the difference. This study reveals that people can be very creative in assessing the quality of arguments.

## References

- Carpenter, C. J. (2015). A meta-analysis of the ELM’s argument quality x processing type predictions. *Human Communication Research, 41*(4), 501-534.
- Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario symposium* (Vol. 5, pp. 3-39). Hillsdale, NJ: Erlbaum.
- Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 64*(4), 568-584.
- Edwards, K., & Smith, E.E. (1996). A Disconfirmation Bias in the Evaluation of Arguments. *Journal of Personality and Social Psychology, 71*(1), 5-24.

- Hoeken, H., & Hustinx, L. (2009). When is statistical evidence superior to anecdotal evidence? The role of argument type. *Human Communication Research, 35*, 491-510.
- Hoeken, H., Šorm, E., & Schellens, P.J. (2014). Arguing about beliefs: Lay people's criteria to distinguish strong arguments from weak ones. *Thinking & Reasoning, 20*(1), 77-98.
- Hoeken, H., Timmers, R., & Schellens, P.J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking & Reasoning, 18*(3), 394-416.
- Hornikx, J., & Hoeken, H. (2007). Cultural differences in the persuasiveness of evidence types and evidence quality. *Communication Monographs, 74*(4), 443-463.
- Jain, S.P., & Maheswaran, D. (2000). Motivated reasoning: A depth-of processing perspective. *Journal of Consumer Research, 26*(4), 358-371.
- Johnson, B. T., Smith-McLallen, A., Killeya, L. A., & Levin, K. D. (2004). Truth or consequences: Overcoming resistance with positive thinking. In E. S. Knowles & J. A. Linn (Eds.), *Resistance and persuasion* (pp. 215-233). Mahwah, NJ: Erlbaum.
- Klaczynski, P., Gordon, D., & Fauth, J. (1997). Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology, 89*, 470-485.
- Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin, 108*(3), 480-498.
- Lord, C.G., Ross, L., & Lepper, M.R. (1979). Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology, 37*(11), 2098-2109.
- Mercier, H. (2012). Looking for arguments. *Argumentation, 20*, 305-324.
- Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology, 33*(2), 243-258.
- Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*, 57-111.
- Molden, D. C., & Higgins, E. T. (2005). Motivated thinking. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 295-317). Cambridge: Cambridge University Press.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.
- Nienhuis, A. E., Manstead, A. R., & Spears, R. (2001). Multiple motives and persuasive communication: Creative elaboration as a result of impression motivation and accuracy motivation. *Personality and Social Psychology Bulletin, 27*(1), 118-132.
- Park, H. S., Levine, T. R., Kingsley Westerman, C. Y., Orfgen, T., & Foregger, S. (2007). The effects of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social-judgment predictions. *Human Communication Research, 33*, 81-102.
- Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion. Central and peripheral routes to attitude change*. Berlin: Springer.
- Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In R. E. Petty, & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 93-130). Mahwah, NJ: Erlbaum.
- Schellens, P. J., & Jong, M. D. T. (2004). Argumentation schemes in persuasive brochures. *Argumentation, 18*, 295-323.
- Van Eemeren, F. H., Garssen, B., & Meuffels, B. (2009). *Fallacies and judgments of reasonableness: Empirical research concerning the pragma-dialectical discussion rules*. Dordrecht: Springer.
- Van Enschot – Van Dijk, R., Hustinx, L., & Hoeken, H. (2003). The concept of argument quality in the Elaboration Likelihood Model. In F. H. van Eemeren, J. A. Blair, C.

Hans Hoeken and Mariecke van Vugt

- A. Willard, & A. F. Snoeck Henkemans (Eds.), *Anyone who has a view. Theoretical contributions to the Study of Argumentation* (pp. 319-335). Dordrecht: Kluwer.
- Westen, D., Blagov, P.S., Harenski, K., & Hamann, S. (2006). Neural Bases of Motivated Reasoning: An fMRI Study of Emotional Constraints on Partisan Political Judgment in the 2004 U.S. Presidential Election. *Journal of Cognitive Neuroscience*, *18*(11), 1947-1958.