Does it take a good argument to be persuaded? How manipulating quality of evidence affects message persuasiveness

Marija Branković[1, 2] & Iris Žeželj[2]

[1]*Department of Psychology, Faculty of Media and Communications,
Singidunum University, Belgrade*
[2]*Department of Psychology, Faculty of Philosophy, University of Belgrade*

Two experiments investigated the effects of manipulating quality of evidence that supports arguments on message persuasiveness. The evidence quality was systematically manipulated by violating one or two of the relevant normative criteria. In experiment one, participants were presented with arguments embedded within a persuasive message. All supportive evidence was of either high, medium or low quality (between-subjects design). In experiment two, each argument was presented separately and was supported with high, medium and low quality evidence (within-subjects design). The recipients were insensitive to manipulations of evidence quality in the first, but sensitive to it in the second experiment. The findings are discussed with reference to conceptual and methodological issues in the study of attitude change.

Keywords: attitude change, argument quality, evidence quality, perceived persuasiveness, actual persuasiveness

How do people evaluate arguments in everyday persuasive settings? Formal and informal logic define distinct criteria that a valid argument should fulfill, but are these criteria relevant for an average message recipient? These issues are of particular importance to persuasion and attitude change researchers. Within these disciplines, argument quality operates as a central methodological tool to test its hypothesis, but is usually established empirically, i.e. it is not based on considerations of what specifically makes a strong or weak argument. Argumentation researchers, on the other hand, seek to define criteria of argument quality, independent of their persuasive effects; e.g. they analyze specific evidence used in support of arguments and define a set of critical questions one should asked about this evidence to establish its acceptability or logical soundness, which is termed the *normative* strength or quality of evidence. This study employs findings from argumentation studies to make a more reliable

Corresponding author: marija.brankovic@fmk.edu.rs

distinction between high and low quality arguments. We are investigating if lay perceptions of arguments' persuasiveness mirror that distinction i.e. whether lay people would be differently persuaded by arguments supported by strong or weak evidence, defined in this manner. By integrating these two research lines, we aim to improve tools for designing arguments in persuasion research. A finer grained analysis of argument quality is needed if we want to reliably distinguish between central modes of message elaboration that include careful scrutiny of arguments and the more peripheral ones that do not.

Argument quality in Persuasion Research

Within persuasion research, argument quality is most widely used for assessing the depth of cognitive processes in message elaboration (e.g., Cacioppo & Petty, 1980; Petty & Cacioppo, 1986; Petty & Wegener, 1999; Rucker, Briñol, & Petty, 2011). According to the Elaboration likelihood model of persuasion (ELM, Petty & Briñol, 2007; Petty & Cacioppo, 1986; Petty & Wegener, 1999), people employ one of the two alternative strategies when confronted with a persuasive message. Central route processing includes a high level of cognitive effort aimed at a thoughtful and systematic scrutiny of message contents, i.e., the arguments the message offers. Although people are motivated to maintain accurate attitudes, because of their limited cognitive capacities they more frequently process the message via peripheral route. Peripheral processing is considerably less cognitively demanding and encompasses a range of different processes, such as classical conditioning or the use of heuristics. Consequently, if a person is more persuaded by strong arguments than weak arguments, one may reasonably assume that the person has processed the message centrally. If both strong and weak arguments are equally successful in persuasion, one assumes that the message has been processed peripherally and without much scrutiny.

The arguments routinely used in ELM experiments usually differ in form, not only in quality; week arguments are often obviously flawed. Strong arguments rely on statistical data or the results from empirical studies, whilst weak arguments are supported by quotations, examples or personal opinions (Petty & Cacioppo, 1986; Petty, Cacioppo, & Goldman, 1981; Petty, Harkins, & Williams, 1980, Study 2). For example, one of the most popular persuasion topics is the ostensible introduction of comprehensive senior exams at universities. A strong argument in favor of exams would state that graduates from universities with exams have higher average starting salaries, whereas a weak argument would state that graduate students thought it would be fair that undergraduates also took comprehensive exams (for more examples, see Petty et al., 1980, Study 2).

The problem of defining argument quality is usually evaded by posing empirical criteria, i.e. by relying on the cognitive responses they elicit (whether they elicit thoughts that are favorable or unfavorable when participants are presented with the arguments). Arguments that elicit predominantly positive

cognitive responses (e.g., 65% favorable and 35% unfavorable thoughts) are classified as strong, whereas those that elicit predominantly negative responses are classified as weak (Petty & Cacioppo, 1986). A number of experiments utilized an identical argument pool originally devised for the topic of exam introduction (e.g., Briñol, Petty, & Barden, 2007; Cacioppo, Petty, & Morris, 1983; Petty et al., 1981; Priester & Petty, 1995). In some of the studies in which new arguments were constructed, this was performed intuitively and the quality of arguments was not empirically pre-tested (Smith & Petty, 1996; Wheeler, Petty, & Bizer, 2005).

The very idea of relying on cognitive responses in assessing argument quality has been criticized for its circularity (O'Keefe, 1995; O'Keefe & Jackson, 1995). In this view, the distinction between strong and weak arguments should rely on independent, normative standards rather than on their perceived effects in a persuasion attempt. What is more, the differences between strong and weak arguments tend to be rather obvious, even for a less cognitively engaged recipient. For instance, one can safely assume that heuristic reasoning is sufficient to designate the arguments that mention statistical data as strong and the arguments that mention an irrelevant experience from one's neighbor as weak. This is a point acknowledged by the authors themselves (See, Petty, & Evans, 2009).

Argument quality: Perspectives from Argumentation Studies

Argumentation can be defined as providing support for or criticizing a certain claim that is questionable, with the aim of convincing a reasonable critic of the acceptability of the claim (Van Eemeren & Grootendorst, 2004; Walton, 2006). The type of argumentation most often used in persuasion research is pragmatic argumentation, in which an act is evaluated based on desirability of its consequences (Honikx, 2008; Van Eemeren, & Grootendsorst, 2004; Walton, 2006; Šorm, 2010). For instance, people could be advised to adopt a kitten since pet care positively affects children's mental health. According to argumentation theorists, a pragmatic argument is normatively strong if one can affirmatively answer two critical questions: a. is the consequence (e.g. favorable mental health status of children) really desirable? and b. does the advised behavior (e.g. pet care) really lead to the desirable consequences (or, in other words, is the consequence really probable in case of the advised behavior).

Desirability of consequences (also called *argument valence*) appears to be much easier to assess, while the probability of consequences (also called *argument strength*) usually need to be further supported (Areni & Lutz, 1988; Hornikx, 2008). This additional support – any facts or opinions presented in support of a claim – is referred to as *evidence*[1]. Argumentation scholars distinguish between different types of evidence (Hornikx, 2008; Hornikx, Starren, & Hoeken, 2003):

---

1 Not all theorist make this distinction between argument quality and evidence quality (e.g. Walton, 2006), but we opted for this perspective since it is well suited for the analysis of argumentation in persuasive contexts.

statistical evidence, anecdotal evidence (evidence by analogy), causal evidence and source (expert) evidence, among others. As with the strength of arguments, there are also several critical questions that can be asked for each of the evidence types, in order to establish their normative strength (terms normative strength and normative quality will be used interchangeably in the manuscript). For instance, in case of expert evidence, one can ask if an expert is truly competent, reliable or impartial. Advice on pet care coming from a veterinarian would be considered normatively more compelling compared to advice coming from a non-expert (e.g. a grandparent or a neighbor). However, an advice coming from a veterinarian employed in a company producing pet products would not be considered impartial.

Of central importance for the present study are the attempts aimed at investigating whether quality of supporting evidence, based on a normative analysis, affects the arguments' persuasiveness. Several studies addressed this issue using different types of evidence and some of the relevant criteria.

In one study, a group of expert judges conducted a normative analysis of the arguments used in previous research within the ELM (van Enschot-van Dijk, Hustinx, & Hoeken, 2003). Contrary to earlier findings (e.g., Areni & Lutz, 1988), strong and weak arguments differed significantly in both argument valence and strength. The authors then designed strong and weak arguments that differed exclusively in argument strength, whereas valence was held constant. In the weak version of the argument one of the relevant normative criteria was violated. For instance, a strong version of an argument cited the Ministry of Education, whereas a weak version cited an anonymous university lecturer. Both strong and weak arguments elicited equally favorable attitudes, and less than one percent of the cognitive responses directly addressed the weaknesses of the arguments.

In a similar study, the authors varied the relevance of source expertise or the percentages and sample sizes on which the percentages were based, e.g., "78% of 314 persons" and "35% of 46 persons" (Hornikx & Hoeken, 2007, p. 453). Unlike the previous study, the arguments were presented as separate claims and not within persuasive messages. Arguments that were supported by evidence of higher normative quality were rated as more probable.

A study conducted by Šorm (2010) tested whether arguments that violate different normative criteria tend to lose their actual persuasiveness. The participants were presented with short texts that contained an argument and a piece of supporting evidence and were asked to indicate the degree to which they agreed with the conclusion. For each argument type, several levels of normative argument quality were designed: the strong arguments met all relevant criteria, whereas each weak argument violated one of the criteria. Participants were not consistently sensitive to the normative violations: normatively strong arguments from cause to effect and from example proved more persuasive, whereas normative violations did not diminish persuasiveness of arguments from authority.

The Present Study

Based on the approach taken within argumentation studies and the existent findings, in the current study we wished to investigate whether people are differently persuaded by arguments supported by normatively good or bad evidence. Borrowing from studies of argumentation, we speak of normative strength of evidence assessed through a set of critical questions to establish their logical soundness. To provide a finer distinction of the quality, we compared arguments supported with evidence that fulfilled two criteria (high quality), violated one of the two (medium quality) or violated both (low quality). In addition, we varied (a) type of evidence provided (statistical, causal, authority or anecdotal) and (b) opportunity to directly compare evidence quality (in between-subjects design they were exposed to either low, medium or high quality evidence, in within-subjects they were exposed to all three).

If normative criteria were relevant for recipients, arguments supported by normatively weaker evidence should be less persuasive; normative violations should be easier to spot if respondents had the opportunity to directly compare evidence of different quality. Apart from being an important issue in its own right, this way of manipulating argument quality could provide a useful methodological tool for future persuasion studies.

Study 1

Overview

In Study 1 we investigated whether offering normatively strong or weak evidence in support of arguments in a persuasive message would make messages differently persuasive for participants. As argumentation research makes a distinction between perceived and actual persuasiveness of arguments, we decided to measure both**.**

Perceived persuasiveness is defined as a subjective impression of how good an argument is, whereas actual persuasiveness is a measure of whether an argument induces an attitude change (Hoeken, 2001; Hornikx, 2008). A recent meta-analysis of 40 studies revealed a mean correlation of .40 between the perceived persuasiveness of the message and the actual persuasiveness measured through attitude measures. The authors concluded that perceived persuasiveness could be considered as a necessary but not sufficient condition for actual persuasiveness (Dillard, Weber, & Vail, 2007, p 625).

Importantly, sensitivity to the normative differences in evidence quality appears to depend on how persuasiveness is measured, as either perceived or actual persuasiveness, because these persuasiveness measures are not completely interchangeable (Dillard et al., 2007). We therefore manipulated the quality of supporting evidence and investigated whether it affected both perceived and actual persuasiveness of arguments.

Method

Design. We conducted an experimental study with a **4 x 4** mixed design, with evidence quality as between-subject factor (four levels: strong (fulfilling both criteria), medium (violating one criterion of the two) or weak (violating two criteria)) and evidence type as within-subject factor (four levels: statistical, authority, anecdotal and causal).

We created two messages with medium quality evidence: one violated the first and the other violated the second of the criteria. From the perspective of quality, these conditions are equivalent, but we opted to distinguish between them for logical reasons (i.e. it is not logically the same whether one or the other criterion is violated), and also to test empirically whether they would produce the same effects (since previous research suggests that some criteria might be more salient or important for laymen than others; see Šorm, 2010).

Participants and Procedure. A total of 83 psychology students from Belgrade, Serbia participated in the study in exchange for course credit. The participants were randomized to the levels of the between-subject factor, which resulted in 19–21 participants per level.

Participants were administered experimental booklets containing persuasive messages, and were asked to carefully read a short text elaborating on the idea of high-school examination reform, one of the "hot" topics in education in the country and therefore relevant for psychologists. It was first explained that the reform, encompassing several comprehensive examinations, is about to be introduced in all high schools in the country. This was followed by the arguments in favour of the examinations. Students were then asked to fill a number of dependent measures. Afterwards, they were thanked and debriefed about the true purpose of the experiment.

Materials. Participants read one of four versions of a persuasive message that advocated a new form of high school examination. Each message comprised the following four arguments: a. the examinations would bolster long-term learning, b. the final grades would be more objective, c. the students would not be required to take entrance exams and d. the new system would be harmonized with the EU regulations. Each of the arguments was supported by a piece of different type evidence (statistical, causal, authority or anecdotal). All the evidence within the same message was of the same quality, so that each participant read four arguments supported by either high, medium (two conditions) or low quality evidence. The arguments were embedded in a newspaper article type text between 307 and 320 words long (arguments and evidence used in the materials are presented integrally in Appendix A).

Independent Variables.

Evidence quality. For each of the evidence types introduced, we chose two of the most relevant critical **questions used to establish normative strength of the evidence** in argumentation literature (Hornikx, 2008; Šorm, 2010; Van Eemeren & Grootendorst, 2004; Walton, 2006). These questions referred to the **probability aspect of argument quality**. The choice was based on two considerations: a. the possibility to manipulate the criteria and b. the demonstrated relevance of the criteria in previous studies. Each participant was presented with evidence from only one level of quality and was thus unable to compare the different levels. This allowed us to test *how poor* the supporting evidence for an argument could normatively be for a psychological effect to occur, i.e., for an average communicator to perceive that there is "something wrong" with the argument.

For example, one of the arguments stated: „*Comprehensive exams would motivate the students to work more regularly and more thoroughly, since they would expect a long-term testing of the knowledge.* „This argument was supported by statistical evidence, which was varied according to two relevant criteria (as presented schematically in Table 1): the sample size (one school in Denmark vs. Scandinavian schools) and the objectivity of the measures employed (objective test results vs. self-reported interests).

Table 1
An example of evidence quality manipulation

Evidence quality by criteria	Sample size	Type of measure
Strong	+ large sample	+ objective outcome measure
Medium I	- small sample	+ objective outcome measure
Medium II	+ large sample	- subjective outcome measure
Weak	- small sample	- subjective outcome measure

Evidence (by level of quality) read as follows:

Strong

The evaluation programmes have demonstrated that **PISA results** of the students **in Scandinavian countries have improved by 15%** since the introduction of the state high-school exams.

Medium(1)

The evaluation programmes have demonstrated that **PISA results** of the students **in one high school in Denmark have improved by 15%** since the introduction of the state high-school exams.

Medium (2)

The evaluation programmes have demonstrated **that the number of students interested in continuing education at universities increased by 15% in Scandinavian countries** since the introduction of the state high-school exams.

Weak

The evaluation programmes have demonstrated that **the number of students interested in continuing education at universities increased by 15% in one high school in Denmark** since the introduction of the state high-school exams.

Similarly, we varied two critical aspects in case of other types of evidence. In the case of expert evidence, we distinguished between more or less relevant sources (education researchers vs. high school teachers) and more or less impartial sources (those involved in the reform vs. those who were not). Anecdotal evidence presented examples that were more or less relevant for the present argument (similar reforms in the neighboring Czech Republic vs. distant New Zealand) and more or less recent (in the last decade vs. in the sixties). Finally, causal evidence varied according to the relevance of the cause (objective criteria for assessment vs. a similar context of test taking) and the relevance of the outcome (the elimination of the "halo effect" which biases the grading process vs. teachers being more relaxed) (complete materials are given in Appendix A).

Evidence type. Four types of evidence supported the presented arguments (as in Hornikx, Starren, & Hoeken, 2003): statistical evidence, anecdotal evidence (evidence by analogy), causal evidence and source (expert) evidence. Each participant read all four types of evidence. This factor was held constant across the arguments (each argument was always supported by identical types of evidence and presented in the same order) to provide comparability. This procedure was chosen as the most ecologically valid because a message varying different types of evidence for an identical claim would not seem genuine in this context. It is clear from this procedure that we opted to optimize the design for testing the effects of evidence quality, while different evidence types were distinguished because criteria of quality are type specific.

Dependent measures.

Perceived persuasiveness. The quality of the message was assessed using a 7-point semantic differential type scale with five attribute pairs (anchored: *poorly supported – well-supported, not persuasive – persuasive, logically incoherent – logically coherent, poorly worded – well worded, meaningless – meaningful*). The scale demonstrated a good internal consistency (α=.89), which allowed for the computing of total scores ranging from 7 to 35.

The quality of the arguments stated in the message was assessed on 7-point scales ranging from *very weak* to *very strong*. This measure was included to enable testing the effects of different evidence types, since the messages contained four arguments, each of which was supported by a different type of evidence. The participants were initially instructed to list the arguments they could identify in the text and then rate their strength. A strong argument was defined as "one that is logically justified to use in support of the issue in question and that is difficult to counter-argue". We opted for this procedure to test whether participants were able to identify the arguments presented. The results showed that most participants were capable – 86.74% of participants identified at least three out of four arguments correctly.

Actual persuasiveness. General attitude measure: The participants were asked to indicate how strongly they supported the introduction of examinations on a 7-point Likert-type scale, ranging from *not at all* to *very strongly*.

The likelihoods of the four outcomes presented in each of the arguments were assessed on 5-point scales, ranging from *not at all likely* to *very likely*. Specifically, we asked participants to assess how probable they thought each of the purported positive results of the introduction of state exams stated in the arguments would be (i.e. that the introduction would bolster long-term learning, give more objective grades, alleviate the stress related to entrance examinations and improve competitiveness of the students at European universities). These ratings served as a more specific measure of the actual persuasiveness of the arguments, in addition to the general attitude measure.

Results

Perceived Persuasiveness. Analysis of variance revealed a marginally significant effect of evidence quality (F (3, 42.12) = 2.67, p = .06) on the ratings of the message quality (M_{strong} = 5.04, SE = 0.89; M_{med1} = 5.67, SE = 0.90; M_{med2} = 4.78, SE = 1.41; M_{weak} = 5.10, SE = 0.91). Because the assumption of equality of variances was violated, we employed a robust test of equality of the means (Welch F ratio). Since we attempted to distinguish between different levels of evidence quality, we tested for a linear trend in the ratings. Planned comparisons did not reveal a significant linear trend, which means that participants' ratings were not affected by the strength of evidence quality, as we defined it through normative criteria. The data were also subjected to a mixed-design ANOVA to test for potential effects on the ratings of argument quality[2]. No significant effects of evidence quality were observed on the ratings of the argument quality

---

2 Since the messages contained four different evidence types, it was not possible to test for effects of evidence type using the global ratings of the message. We therefore used the ratings of isolated arguments, each of which was supported with a different type of evidence, and included only the participants who successfully stated all of the arguments (*n*=40). Ratings of message quality and mean ratings of quality of the four arguments were strongly correlated (r = .69, p = .00).

either. We can therefore conclude that the messages supported by normatively stronger evidence were not perceived as more persuasive by participants.

The effects of evidence type did not reach statistical significance either, although anecdotal evidence tended to be rated somewhat less favorably than the remaining three evidence types (the means are detailed in Table 2). The registered tendencies were consistent with previous findings; however the design was not optimal for testing the effects of evidence type because of a possible confound with the contents of the claims so we will not address this issue in more depth (Hoeken, 2001; Hornikx, 2008). No significant interactions between the two factors were observed.

Table 2
Perceived Persuasiveness of Evidence Differing in Quality and Type

Evidence quality	Evidence type				
	Statistical	Source (expert)	Anecdotal	Causal	Mean
Strong	5.40 *(1.71)*	5.50 *(.97)*	4.10 *(1.52)*	5.00 *(1.89)*	5.04
Medium 1	5.45 *(1.81)*	5.27 *(1.35)*	5.00 *(1.79)*	5.45 *(1.29)*	5.29
Medium II	5.00 *(1.31)*	4.88 (1.36)	4.75 *(1.67)*	4.63 *(1.85)*	4.81
Weak	4.82 *(1.40)*	4.73 *(1.28)*	4.00 *(1.61)*	4.91 *(1.61)*	4.61
Mean	5.16	5.10	4.46	5.00	4.94

Note. Means and standard deviations of ratings of perceived persuasiveness (ranging from 1 to 7)

Actual Persuasiveness. Analysis of variance revealed no significant effects of evidence quality on either of the measures of actual persuasiveness. These findings indicate that participants neither perceived nor were differentially persuaded by evidence differing in normative quality.

Ratings of the likelihood of the outcomes supported by different types of evidence did not differ significantly either. However, there appeared a significant interaction effect ($F_{(9)} = 1.95$, $p=.05$, $\eta^2=.07$), which indicates that actual persuasiveness of evidence differing in quality may depend on evidence type. One-way ANOVAs revealed significant differences in ratings of outcomes depending on evidence quality in case of evidence from authority ($F_{(3, 76)} = 3.02$, $p=.03$, $\eta^2 = .11$), roughly, but not entirely in line with the predictions (Table 3).

Table 3
Actual Persuasiveness of Evidence Differing in Quality and Type

Evidence quality	Evidence type				
	Statistical	Source (expert)	Anecdotal	Causal	Mean
Strong	3.10 *(.97)*	3.55 *(1.1)*	3.15 *(.99)*	3.00 *(1.12)*	3.20
Medium 1	3.42 *(1.02)*	2.95 *(1.5)*	3.53 *(1.07)*	3.68 *(1.00)*	3.39
Medium II	3.19 *(1.03)*	3.57 *(1.1)*	2.90 *(1.09)*	3.52 *(1.17)*	3.30
Weak	3.10 *(.97)*	2.65 *(.99)*	2.70 *(1.08)*	3.10 *(1.02)*	2.89
Mean	3.20	3.19	3.06	3.33	3.19

Note. Means and standard deviations of ratings of actual persuasiveness (ranging from 1 to 5)

As an aid in understanding the results, we investigated the relationship between perceived and actual persuasiveness. The correlation between the mean ratings of message quality and attitude measure was significant but modest (r = .29, p = .00). The regression analysis revealed that message quality explained 8.5% of the variance in the attitudes (R^2 = .08, p = .00). The same measure explained approximately 18% of the variance in the mean rating of the outcome probability (R^2 = .18, p = .00). The mean ratings of the arguments were also significant predictors of the attitudes and accounted for 30% of their variance (R^2 = .30, p = .00). Our findings thus demonstrate that a large amount of the variance in the actual persuasiveness of an argument was not a result of its perceived persuasiveness.

Discussion

Our findings suggest that the message recipients perceived persuasive arguments as equally compelling, regardless of the quality of the supporting evidence. The present experiment mirrored a typical persuasion experiment. From the perspective of argumentation analysis, this study confronted recipients with the demanding task to identify and extract relevant arguments and assess their validities with no quality variation that would provide opportunity for comparison.

These findings are consistent with previous studies in which weak arguments that were similar in form to the strong ones (e.g. providing statistical evidence, albeit weak; consulting and expert albeit biased) were equally efficient in attitude change as strong arguments (Branković, M., & Žeželj, I., 2010; Van Enschot-Van Dijk et al., 2003).

Contrary to the present findings, some studies observed that laymen occasionally rely on normative criteria for evaluating evidence (Hoeken & Hustinx, 2009; Hornikx & Hoeken, 2007). The type of task in which the arguments are presented appears to be one of the important dimensions affecting whether recipients use normative criteria. Hoeken (2001) suggested that an important issue is whether the arguments are presented separately or embedded within a text because the processing of irrelevant information also expends cognitive capacities. Another possible explanation is that the differences in sensitivity to normative criteria could depend on the opportunity to compare evidence that differs in normative strength within the same task. The presence of evidence that differs in quality could be assumed to offer implicit standards and therefore facilitate the task. For instance, the participants in previous studies were also able to make distinctions between strong and weak arguments when the entire preliminary pool was pre-tested for quality (Branković, M., & Žeželj, I., 2010).

We designed another experiment to test whether participants would demonstrate sensitivity to the normative strength of the evidence employed in our initial study if different quality evidence appeared *within the same task.*

Study 2

Overview

In this study, we wanted to test whether normative criteria of evidence strength are of any relevance for laymen, and in this aim we sought to make the task as simple as possible for the participants. We wanted to test whether participants will *recognize* differences in evidence strength in a task where they are made highly salient, since evidence of differing quality is presented within the same task and thus directly comparable. We used the same material as in Study 1, but presented each participant with all the levels of evidence strength. We opted to measure only perceived persuasiveness in this study since we were interested in participants' perceptions and also since it would be not ecologically valid to assess actual persuasiveness in a task where evidence of differing quality is presented simultaneously.

Method

Design. In the second study we also applied a **4** (evidence type: statistical, authority, anecdotal and causal) x **4** (evidence quality: strong, medium (two levels), weak) experimental design. The crucial change was that we used a complete *within-subject design*, i.e., each participant was presented with evidence from each level of evidence quality (16 separate claims), which allowed for direct comparisons among evidence levels.

Materials. In this experiment, we used identical arguments with supporting evidence as in the initial study. The arguments were presented as separate claims followed by four pieces of evidence that differed in quality. The students were thus exposed to a total of 16 situations (four types of evidence x four levels of quality). The order of the blocks, as well as the evidence within the blocks was counter-balanced.

Participants and Procedure. A total of 35 psychology students from Belgrade, Serbia participated in exchange for course credit. The participants were presented with written arguments and asked to assess their quality, which took approximately 10 to 15 minutes; they were then thanked and debriefed.

Dependent Measures. The participants indicated the extent that each piece of evidence made the argument strong or weak on a 7-point Likert-type scale, ranging from *very weak* to *very strong*. Similar to experiment 1, a strong argument was defined as logically sound and difficult to refute.

Results

A repeated measures ANOVA revealed a significant main effect of evidence quality (F (3, 93) = 2.94, p = .04, η^2 = .09). The argument quality ratings depended on the quality of the supporting evidence. A significant linear trend was observed (F (1, 31) = 4.33, p = .05, η^2 =.08), which indicated that the participants were sensitive to the number of relevant normative criteria a piece of evidence fulfilled (Table 4).

Table 4

Means and Standard Deviations of Different Levels of Evidence Quality and Type

Evidence quality	Evidence type				
	Statistical	Source (expert)	Anecdotal	Causal	Mean
Strong	4.91 *(1.63)*	5.25 *(.95)*	4.81 *(1.35)*	5.59 *(1.46)*	5.14
Medium I	4.78 *(1.43)*	4.91 *(1.28)*	4.63 *(1.45)*	5.06 *(1.41)*	4.84
Medium II	5.22 *(1.34)*	4.81 *(1.15)*	4.25 *(1.70)*	5.13 *(1.36)*	4.85
Weak	4.72 *(1.30)*	5.13 *(1.04)*	4.31 *(1.75)*	4.81 *(1.49)*	4.74
Mean	4.91	5.02	4.5	5.15	4.89

Note. Means and standard deviations of the ratings (ranging from 1 to 7)

Pairwise comparisons revealed significant differences between the first and all other levels of quality: second level (MD = .30, p = .04), third level (MD = .29, p = .07) and fourth level (MD = .40, p = .03). The mean differences were computed for each level of quality across the four evidence types – for each type of evidence (4 levels), the ratings for each level of quality were averaged. The differences between the second and third levels and between the third and fourth levels were not significant. The psychological boundary between what constitutes a good and a bad argument thus appears to lie between evidence that fulfills the two criteria and evidence that violates either one or both of the criteria.

The analyses also revealed a significant effect of evidence type (F (3, 93) = 3.05, p = .03, η^2 =.09). Anecdotal evidence was perceived as significantly less persuasive than other evidence types (statistical, causal and authority). No significant interaction effects were observed.

Discussion

In contrast to the previous study, this study demonstrated that people can be sensitive to violations of normative criteria of evidence quality when assessing arguments. Significant differences in ratings support the conclusion that people are sensitive to variations of evidence quality based on the number of relevant normative criteria that the evidence fulfills. We believe that the nature of the task facilitated the recognition of differences in evidence quality, in at least two relevant aspects: a. the arguments and supporting evidence were extracted and b. most importantly, the participants were allowed to directly compare four different levels of evidence quality.

General Discussion

The two studies investigated whether the normative strength of evidence supporting persuasive arguments had psychological relevance, i.e., whether lay people take into account the quality of supporting evidence when assessing

arguments. The findings demonstrated seemingly conflicting tendencies. In the initial study, the participants were not sensitive to the number of criteria met or violated, whereas such sensitivity was observed in the second study. It appears that people do not tend to *spontaneously* apply normatively relevant criteria when assessing the persuasiveness of arguments; however, participants do tend to *recognize* relevant criteria if they are salient in the context, i.e. if arguments differing in quality are simultaneously presented.

What can be concluded with certainty from the present findings is that the task in the second study facilitated recognition and application of relevant normative criteria, or, put differently, that there are situations in which these criteria are relevant for lay people. As the employed design does not allow us to conclude that possibility of comparison was the unique factor leading to sensitivity to normative criteria of evidence quality, future studies are warranted to explore this further.

Mirroring previous findings (Hoeken, 2001; Hornikx, 2008), anecdotal evidence was overall rated slightly less persuasive than statistical evidence. The relation between normative strength and *perceived persuasiveness* did not differ across different types of evidence: evidence of the same quality was judged as equally persuasive irrespective of the evidence type. However, evidence type appears to influence *actual persuasiveness*, as evidenced by the interactions observed in the initial study. This implies that evidence type should be taken into account when testing for potential effects of persuasive evidence. At the same time, it further supports the idea that actual and perceived persuasiveness are partially overlapping but distinct dimensions (e.g., Dillard et al., 2007; Šorm, 2010). In addition, the more specific measures of perceived persuasiveness (i.e., argument ratings) served as better predictors of actual persuasiveness than the global ratings of message quality. The argument quality ratings explained on average 36% of the variance in the ratings of the outcome probabilities presented in the arguments, consistent with previous studies (Hoeken, 2001). The relations between perception and the persuasiveness of different evidence types should be investigated further.

The current studies have several implications for persuasion research. Our study demonstrated one possible method for a normatively founded operationalization of argument quality that would allow its more systematic investigation. We believe that this is a prerequisite for clarifying the cognitive operations involved in the process of attitude change, and specifically, distinguishing between the central and peripheral routes of elaboration (as suggested by O'Keefe, 1995). If message recipients cannot recognize the normatively founded distinctions in argument quality, this implies either that the message elaboration has engaged less complex cognitive operations, or that people do not apply any normatively relevant criteria when assessing arguments (i.e. that non-expert criteria are completely different from normative criteria). Our findings appear to be more in line with the former proposition: as spontaneous application of normative criteria is a cognitively demanding

task, they are used only when recipients are highly motivated to elaborate the message. In other words, "truly central" message processing could be an even less frequently applied mode than presumed. Persuasion research could also benefit from the distinction between perceived and actual persuasiveness when studying the cognitive processes mediating attitude change. Although persuasion researchers do not use this terminology, they sometimes include measures of perceived message quality as indicators of the underlying cognitive processes (e.g., Cacioppo, Petty, & Sidera, 1982; Lavine & Snyder, 1996). However, perceived message quality and attitude change (i.e. actual persuasiveness) can be affected differently by argument quality and may not be in a straightforward relationship.

There are several limitations to the current studies. We investigated only some of the possible argument and evidence types and a limited number of relevant criteria. There is evidence (e.g., Šorm, 2010) that lay people are not equally sensitive to different normative criteria, which raises the issue of whether some criteria are more important than others in affecting persuasiveness. The issue of the relation between the number of violated criteria and the magnitude of their violation remains to be clarified in future research. It should be noted that translating well-defined normative criteria to experimental manipulations of quality is not a task with a single solution, and different manipulations should not be treated equivalently without investigation.

The present findings encourage further efforts to develop a research approach that combines insights from persuasion and argumentation studies to benefit both disciplines.

References

Areni, C., & Lutz, R. J. (1988). The role of argument quality in the Elaboration Likelihood Model. *Advances in Consumer Research, 15*, 197–203.

Branković, M., & Žeželj, I. (2010). The effects of matching a persuasive message to a recipient's self-concept on attitude change. *Psihologija*, *43*, 233–251. doi: 10.2298/PSI1003233B

Briñol, P., Petty, R. E., & Barden, J. (2007). Happiness versus sadness as a determinant of thought confidence in persuasion: a self-validation analysis. *Journal of Personality and Social Psychology, 93*, 711–727. doi: 10.1037/0022–3514.93.5.711

Cacioppo, J. T., & Petty, R. E. (1980). Persuasiveness of communications is affected by exposure frequency and message quality: A theoretical and empirical analysis of persisting attitude change. In J. H. Leigh & C. M. Martin (Eds.), *Current Issues and Research in Advertising* (pp. 97–122). Ann Arbor: The University of Michigan. doi: 10.1080/01633392.1980.10505295

Cacciopo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion.*Journal of Personality and Social Psychology, 45*, 805–818. doi: 10.1037/0022–3514.45.4.805

Cacciopo, J. T., Petty, R. E., & Sidera, J. A. (1982). The effects of a salient self-schema on the evaluation of proattitudinal editorials: Top-down versus bottom-up message

processing. *Journal of Experimental Social Psychology, 18*, 324–338.doi: 10.1016/0022–1031(82)90057–9

Dillard, J. P., Weber, K. M., & Vail, R. G. (2007). The relationship between the perceived and actual effectiveness of persuasive messages: A meta-analysis with implications for formative campaign research. *Journal of Communication, 57*, 613–631. doi:10.1111/j.1460–2466.2007.00360.x

Hoeken, H. (2001). Anecdotal, statistical and causal evidence: their perceived and actual persuasiveness. *Argumentation*, *15*, 425–437. doi: 10.1023/A:1012075630523

Hoeken, H., & Hustinx, L. (2009). When is statistical evidence superior to anecdotal evidence in supporting probability claims? The role of argument type. *Human Communication Research*, *35*, 491–510.doi:10.1111/j.1468–2958.2009.01360.x

Hornikx, J. (2008). Comparing the actual and expected persuasiveness of evidence types: how good are lay people at selecting persuasive evidence? *Argumentation, 22*, 555–569. doi: 10.1007/s10503–007–9067–6

Hornikx, J., & Hoeken, H. (2007). Cultural differences in the persuasiveness of evidence types and evidence quality.*Communication Monographs, 74*, 443–463. doi: 10.1080/03637750701716578

Hornikx, J., Starren, M., & Hoeken, H. (2003). Cultural influence on the relative occurrence of evidence types. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.), *Proceedings of the fifth conference of the International Society for the Study of Argumentation* (pp. 531–536). Amsterdam: Sic Sat.

Lavine, H., & Snyder, M. (1996). Cognitive processing and the functional matching effect in persuasion: The mediating role of subjective perceptions of message quality. *Journal of Experimental Social Psychology, 32*, 580–604. doi:10.1006/jesp.1996.0026

O'Keefe, D. (1995). Argumentation studies and dual-process models of persuasion. In F. H. van Eemeren, R. Grootendorst, J. A. Blair, & C. A. Willard (Eds.), *Proceedings of the third ISSA conference on argumentation, vol. 1: Perspectives and approaches* (pp. 3–17). Amsterdam: Sic Sat.

O'Keefe, D., & Jackson, S. (1995). Argument quality and persuasive effects: A review of current approaches. In S. Jackson (Ed.), *Argumentation and values: Proceedings of the ninth Alta conference on argumentation* (pp. 88–92). Annandale, VA: Speech Communication Association.

Petty, R. E., & Briñol, P. (2008). Persuasion: From single to multiple to meta-cognitive processes. *Perspectives on Psychological Science, 3,* 137–147. doi: 10.1111/j.1745–6916.2008.00071.x

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York: Springer-Verlag.

Petty, R. E., Cacciopo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41,* 847–855. doi:10.1037/0022–3514.41.5.847

Petty, R. E., Harkins, S. G., & Williams, K. D. (1980). The effects of group diffusion of cognitive effort on attitudes: an information-processing view. *Journal of Personality and Social Psychology, 38*, 81–92. doi:10.1037/0022–3514.38.1.81

Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 41–72). New York: Guilford Press.

Priester, J. R., & Petty, R. E. (1995). Source attributions and persuasion: perceived honesty as a determinant of message scrutiny. *Personality and Social Psychology Bulletin, 21*, 637–654. doi:10.1177/0146167295216010

Rucker, D. D., Briñol, P., & Petty, R. E. (2011). Metacognition. Methods to assess primary versus secondary cognition. In K. C. Klauer, A. Voss, & C. Stahl (Eds.), *Cognitive methods in social psychology* (pp. 236–264). New York: Guilford Press.

See, Y. H. M., Petty, R. E., & Evans, L. M. (2009). The impact of perceived message complexity and need for cognition on information processing and attitudes. *Journal of Research in Personality, 43,* 880–889. doi:10.1016/j.jrp.2009.04.006

Smith, S. M., & Petty, R. E. (1996). Message framing and persuasion: A message processing analysis. *Personality and Social Psychology Bulletin, 22*, 257–268. doi:10.1177/0146167296223004

Šorm, E. (2010). *The good, the bad and the persuasive*. Utrecht: LOT.

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

Van Enschot-Van Dijk, R., Hustinx, L. M., & Hoeken, H. (2003). The concept of argument quality in the Elaboration Likelihood Model: A normative and empirical approach to Petty and Cacioppo's 'strong' and 'weak' arguments. In: F. H. van Eemeren, A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.) *Anyone who has a view. Theoretical contributions to the study of argumentation* (pp. 320–333). Dordrecht/Boston/London: Kluwer Academic Publishers

Walton, D. (2006). *Fundamentals of critical argumentation: Critical reasoning and argumentation.* New York: Cambridge University Press.

Wheeler, S. C., Petty, R. E., & Bizer, G. Y. (2005). Self-schema matching and attitude change: Situational and dispositional determinants of message elaboration. *Journal of Consumer Research, 31*, 787–797. doi:10.1086/426613

Appendix A
Materials used in the study
(arguments and evidence, varying by level of quality)

Argument 1: Comprehensive exams would motivate the students to work more regularly and more thoroughly, since they would expect a long-term testing of the knowledge.

Evidence:

1. The evaluation programmes have demonstrated that PISA results of the students in Scandinavian countries have improved by 15% since the introduction of the state high-school exams. (strong)
2. The evaluation programmes have demonstrated that PISA results of the students in one high school in Denmark have improved by 15% since the introduction of the state high-school exams. (medium 1)
3. The evaluation programmes have demonstrated that the number of students interested in continuing education at universities increased by 15% in Scandinavian countries since the introduction of the state high-school exams. (medium 2)
4. The evaluation programmes have demonstrated that the number of students interested in continuing education at universities increased by 15% in one high school in Denmark since the introduction of the state high-school exams. (weak)

Argument 2: The comprehensive high-school examination would replace the entrance examinations for universities, and taking a test in a familiar environment entails considerably lower stress levels and better performance.

1. Researchers of the Institute of educational reasearch believe that entrance examinations cause great stress and that test-taking in a familiar environment provides much better conditions for students to demonstrate their knowledge. (strong)
2. State secretary from the Ministry of education, one of the creators of the reform, believes that entrance examinations cause great stress and that test-taking in a familiar environment provides much better conditions for students to demonstrate their knowledge. (medium 1)
3. Aleksandar Miladinović, a high-school teacher and a member of the Higher education council, believes that entrance examinations cause great stress and that test-taking in a familiar environment provides much better conditions for students to demonstrate their knowledge. (medium 2)
4. Aleksandar Miladinović, a high-school teacher and a member of the the team which will implement the reform, believes that entrance examinations cause great stress and that test-taking in a familiar environment provides much better conditions for students to demonstrate their knowledge. (weak)

Argument 3: State high-school examinations are a vital element of the educational reform, that would make our high-school students more competitive at universities throughout the European Union.

1. The Czech Republic makes a very illustrative example, since a similar reform conducted there right before EU accession resulted in very good positioning of Czech students at European universities. (strong)
2. The Czech Republic makes a very illustrative example, since a similar reform conducted there in the mid 1960s resulted in very good positioning of Czhech students at European universities. (medium 1)
3. New Zealand makes a very illustrative example, since a similar reform conducted there several years ago resulted in very good positioning of their students at the universities in USA. (medium 2)
4. New Zealand makes a very illustrative example, since a similar reform conducted there in the mid 1960s resulted in very good positioning of their students at the universities in USA. (weak)

Argument 4: Comprehensive state high-school examinations would provide a more objective assessment of the students' knowledge than school grades.

1. State examinations will be graded according to objective criteria, following the rule that a teacher must not grade the students he or she teaches. This procedure will eliminate biases in assessment connected with previous experience with the student, as halo-effect (to expect better performance from high achievers and lower performance from low achievers). (strong)
2. State examinations will be graded according to objective criteria, following the rule that a teacher must not grade the students he or she teaches. This procedure entails less stress for the teachers who grade the tests. (medium 1)
3. State examinations will provide similar test-taking conditions for all students: they will be administered at the same time and following the same procedure. This procedure will eliminate biases in assessment connected with previous experience with the student, as halo-effect (to expect better performance from high achievers and lower performance from low achievers). (medium 2)
4. State examinations will provide similar test-taking conditions for all students: they will be administered at the same time and following the same procedure. This procedure entails less stress for the teachers who grade the tests. (weak)