An exploratory test of an intuitive evaluation method of perceived argument strength

Jos Hornikx*, Radboud University Nijmegen, Centre for Language Studies, Netherlands
Annemarie Weerman, Radboud University Nijmegen, Centre for Language Studies, Netherlands
Hans Hoeken, Utrecht University, Department of Languages, Literature & Communication, Netherlands

*Corresponding author: jos.hornikx@ru.nl

Abstract

According to Mercier and Sperber (2009, 2011, 2017), people have an immediate and intuitive feeling about the strength of an argument. These intuitive evaluations are not captured by current evaluation methods of argument strength, yet they could be important to predict the extent to which people accept the claim supported by the argument. In an exploratory study, therefore, a newly developed intuitive evaluation method to assess argument strength was compared to an explicit argument strength evaluation method (the PAS scale; Zhao et al., 2011), on their ability to predict claim acceptance (predictive validity) and on their sensitivity to differences in the manipulated quality of arguments (construct validity). An experimental study showed that the explicit argument strength evaluation performed well on the two validity measures. The intuitive evaluation measure, on the other hand, was not found to be valid. Suggestions for other ways of constructing and testing intuitive evaluation measures are presented.

Keywords

perceived argument strength, intuitive inferences, argument quality, evaluation method

1 Introduction

Methods for evaluating arguments are important in pretesting or evaluating health behaviour campaigns, policy issues and public service announcements. The pretesting of persuasive documents (De Jong, 1998) and campaign materials (Whittingham, Ruiter, Zimbile, & Kok, 2008) is indeed very useful to improve their effectiveness (Cappella, 2018; Dillard, Weber, & Vail, 2007; Noar, Bell, Kelley, Barker, & Yzer, 2018; O'Keefe, 2018, 2020). To obtain strong and lasting effects, the campaign should contain strong and compelling arguments. It is hard to predict how a given argument will be evaluated by the audience. In argumentation theory, a central strand of research has focused on how to evaluate the quality of an argument. In the perspective of argumentation schemes (prototypical types of arguments), critical questions have been developed that serve to assess the quality of an argument. Depending on the type of argument, different critical questions have been proposed (Hastings, 1962; Van Eemeren & Grootendorst, 1992; Walton, Reed, & Macagno, 2008). Arguments that may be considered to have a high argument quality following the critical questions could be candidates for a strong and compelling argument for a given claim. However, what is normatively strong is not necessarily persuasive (O'Keefe, 2007, for a discussion). One reason is that the perspective of critical questions cannot take into account the audience's acceptance of the claim nor that of the argument itself. Therefore, it is imperative to have argument evaluation instruments that are able to predict whether the use of a certain argument, which may be constructed with the critical questions in mind, will lead to a higher acceptance of a claim.

Most current argument strength evaluation methods in persuasion research, such as the thought-listing procedure (Cacioppo & Petty, 1981) or (explicit) perceived argument strength instruments

https://doi.org/10.24434/j.scoms.2022.02.003

(Zhao, Strasser, Cappella, Lerman, & Fishbein, 2011), have participants reflect on, and deliberately evaluate the arguments under consideration. Such deliberate evaluations do not seem to reflect actual reasoning in everyday life. People generally reason rather intuitively than deliberately. Mercier and Sperber (2009, 2011, 2017) state that people have an immediate and intuitive feeling about the strength of an argument, which is the basis for most real-life argument evaluations. Current evaluation instruments do not seem to capture these intuitive evaluations, whereas such evaluations could be important for predicting an argument's persuasiveness. Therefore, an interesting avenue for research in this area would be to develop measures that capture people's *intuitive* perceptions of argument strength. In the present study a new intuitive method for evaluating arguments will be compared to an explicit, deliberate measure of perceived argument strength with respect to their construct and predictive validity.

1.1 Existing argument strength evaluation methods

Several measurement techniques for assessing argument strength exist, such as the classic thought-listing technique where participants are asked to write down any thoughts they had when reflecting upon an argument (Cacioppo & Petty, 1981). If an argument evokes mainly positive thoughts, it is considered strong; if it evokes mainly negative thoughts, it is considered weak. Since this method is labour intensive and coding can be subjective, Zhao et al. (2011) have developed and validated an alternative instrument to assess argument strength. It should be observed that, in some areas of research, the term *argument strength* is mentioned as one of the dimensions of *argument quality* (Areni & Lutz, 1988; Johnson, Smith-McLallen, Killeya, & Levin, 2004), with *argument valence* being the other dimension. In this paper, the term *argument strength* is used to refer to how an audience evaluates an argument in order to signal the similarity to Zhao et al.'s (2011) instrument that plays a central role in the current research. *Argument quality*, in contrast, refers to the intrinsic characteristics of an argument (for a discussion of argument strength and argument quality, see Hoeken, Hornikx, & Linders, 2020).

Zhao et al.'s (2011) perceived argument strength (PAS) scale consists of nine questions tapping into different aspects of argument strength, among which participants' self-report of positive or negative thoughts they had, plausibility and importance of the arguments, and overall quality of the arguments. The PAS scale's reliability and validity were tested and demonstrated using anti-smoking and anti-drugs public service announcements. Scores on thought-listing and the PAS scale were highly correlated, and greater perceived argument strength consistently turned out to be related to more favourable attitudes towards the position advocated (Bigsby, Cappella, & Seitz, 2013; Zhao et al., 2011).

1.2 Deliberate versus intuitive processing

For a better understanding of the distinction between intuitive and deliberate argument evaluations and of the relevance of intuitive evaluations, it is useful to consider the distinction between intuitive and deliberate processing, which is a hallmark of Dual-Process Theories of reasoning (DPT; Chaiken & Trope, 1999; De Neys & Pennycook, 2019; Evans, 2010; Evans & Stanovich, 2013; Gawronski & Creighton, 2013). DPT distinguishes between fast and intuitive *Type 1* processes on the one hand, and slow and reflective *Type 2* processes on the other. The difference between these two types has been examined in the context of logical reasoning and judgment and decision-making (for reviews, Evans, 2010; Evans & Stanovich, 2013). Many problems that participants are asked to solve in these fields require *Type 2* processing, but the (wrong) answers provided by people indicate that they often base their answer on *Type 1* processing.

The general idea is that intuitive processing happens automatically when encountering reasoning problems whereas analytical *Type 2* processing only kicks in later. Especially under time pressure, *Type 1*

processing is more relied upon to come up with an answer. Evans and Curtis-Holmes (2005) provided participants with a number of syllogistic reasoning problems and asked them whether the premises warrant the conclusion. Participants were asked to either respond quickly or were allotted unlimited time to respond. The reasoning problems varied on believability and validity. A claim should be accepted when the reasoning is valid, regardless of its believability, like in the following example: "No healthy people are unhappy; some astronauts are unhappy, therefore some astronauts are not healthy people" (Evans & Curtis-Holmes, 2005, p. 384). The claim should be rejected when the reasoning is invalid, even though it can still be believable, for example: "No healthy people are unhappy; some astronauts are unhappy, therefore some healthy people are not astronauts". The intuitive *Type 1* processing was expected to be sensitive to the believability of claims, while the analytical *Type 2* processing should be more sensitive to the validity of the reasoning, as validity is often more difficult to determine. The sensitivity to believability of *Type 1* processing was expected to be stronger under time pressure.

The results obtained confirmed this prediction: In the rapid response condition, participants accepted more invalid but believable conclusions than valid but unbelievable conclusions. This indicates that *Type 1* processing is more dominant than *Type 2* processing under time pressure. These results suggest that intuitive processing can lead to faulty conclusions.

In the unlimited time condition, participants provided more correct conclusions which suggests that *Type 2* processing can correct the output of *Type 1* processing. However, this correction effect is limited as well. A series of experiments by Thompson, Prowse Turner, and Pennycook (2011) showed that *Type 2* is often used to rationalise initial intuitive answers, even if these are incorrect. Participants were requested to answer different kinds of reasoning problems twice: Participants had to provide an answer immediately, their first initial response, and subsequently had to provide their final deliberate answer after a period of thinking. For both answers, participants were asked to indicate how confident they felt about their answer: the Feeling of Rightness (FOR). What Thompson et al. (2011) consistently showed was that less (*Type 2*) thinking time was allotted to problems for which the intuitive answers were accompanied by a stronger FOR. FOR judgments themselves were related to the fluency with which the answers came to mind. If the answer came to mind quickly and easily, participants felt more confident about their answer and assumed it was correct more often. Interestingly, changing an answer on the basis of *Type 2* processes was unlikely and did not necessarily lead to correct answers. This suggests that *Type 2* processes were mostly used to confirm the initial response, especially when participants felt confident about their intuitive answer.

1.3 Intuitive versus deliberate argument evaluation: Predictive validity

Type 1 processes may also be relevant for argument evaluation. According to Mercier and Sperber (2011, p. 59), "intuitions about arguments have an evaluative component: Some arguments are seen as strong, others as weak. […] These evaluation[s] and preferences are ultimately grounded in intuition." These intuitive inferences could be construed as similar to *Type 1* processing (for a more elaborate discussion, see Mercier & Sperber, 2009). Cognitive energy that is subsequently spent on evaluating arguments is often aimed at confirming this intuition, rather than a more objective and deliberate reflection on the argument's merits (similar to the effect found by Thompson et al., 2011). Current methods for argument strength evaluations, such as the PAS scale (Zhao et al., 2011), appear to stimulate participants to evaluate arguments in a deliberate manner. If Mercier and Sperber (2009, 2011) are correct, such methods may fail to identify the more intuitive responses. The aim of the current paper is to develop and test a measure that taps into these more intuitive responses. Based upon the studies of Evans and Curtis-Holmes (2005) and Thompson

et al. (2011) concerning reasoning under time pressure, the current paper will report on a study in which people have to rate arguments under time pressure which would force them to rely on their intuitive responses. The current paper compares the extent to which intuitive and deliberate argument evaluations predict subsequent claim acceptance, to establish the methods' predictive validity.

1.4 Intuitive versus deliberate argument evaluation: Construct validity

Apart from comparing the intuitive and deliberate argument evaluations with respect to their predictive validity, the construct validity will be assessed as well. An argument evaluation measure has adequate construct validity if it can discriminate between arguments with varying degrees of strength. A number of empirical studies have examined arguments with varying degrees of strength, usually for one type of argument, which is typical for persuasive messages: the pragmatic argument, also known as the argument from consequences (O'Keefe, 2013; Schellens & De Jong, 2004; Walton, 1996). An argument from consequence advocates taking a certain action or introducing a certain measure on the basis of its consequences. From a normative perspective, the strength of these arguments should depend on two main criteria (Walton et al., 2008). First, the consequence should be considered desirable. Second, the occurrence of the consequence as a result of the advocated action should be likely or probable.

Previous studies have shown that the desirability of the consequence is a strong determinant of argument strength (Areni & Lutz, 1988; Johnson et al., 2004; Hoeken, Timmers, & Schellens, 2012; O'Keefe, 2013) and that an argument from consequence is more persuasive if the consequence referred to is considered more desirable (O'Keefe, 2013). Changing the desirability of an outcome can be achieved in two ways: by using different outcomes where one is more desirable than the other (e. g., "Introducing senior comprehensive exams led to an increase in *grade point average* vs. an increase in *student anxiety*", Petty & Cacioppo, 1986), or by using the same outcome and varying the strength or positivity of that outcome (e. g., "The VCR provides *exceptional* picture quality vs. *adequate* picture quality", Wheeler, Petty, & Bizer, 2005). The latter option may create smaller differences in desirability, but ensures similar content in both versions, keeping the arguments more comparable.

In addition, the probability or likelihood of the occurrence of the consequences can be manipulated. Presenting more probable outcomes – outcomes that are more likely to occur – should make for stronger arguments as well. A strategy to manipulate probability is by varying the evidence used to support a claim. It has been shown in several studies that people are sensitive to these variations in argument strength, since claims have been found to be accepted to a larger degree depending on the strength of the supportive argument (Hoeken & Hustinx, 2009; Hoeken, Šorm, & Schellens, 2014; Hoeken et al., 2012; Hornikx & Hoeken, 2007). These studies also showed that probability variations have resulted in smaller differences in claim acceptance than desirability variations.

1.5 Research questions

This study aims at evaluating an intuitive evaluation method of argument strength, which could be an addition to explicit argument evaluation methods. This method is only valid and useful if differences in argument evaluations are predictive for differences in claim acceptance (predictive validity) and if arguments with varying degrees of strength are correctly discriminated (construct validity). To test whether the two argument evaluation methods meet these requirements, the research questions addressed in the current study are:

RQ1: To what extent can an intuitive evaluation method and an explicit evaluation method predict claim acceptance? (predictive validity)
RQ2: To what extent are these methods sensitive to manipulations of argument quality? (construct validity)

As not much is known about the nature of intuitive evaluations and how to measure them, this is an exploratory study.

2 Method

Participants indicated the extent to which they accept claims supported by arguments, and they provided intuitive evaluations of argument strength for these arguments as well as evaluated arguments explicitly employing the PAS scale. The intuitive measures consisted of argument strength items that needed to be answered under time pressure.

2.1 Participants and procedure

121 participants from the Netherlands took part in the study, 72.7 % female, mean age of 30.4 years (SD = 17.5; range: 13–78). Level of education varied from high school (22.3 %) to a completed MA degree (43.8 %).

Participants were approached in the centre of a city in the Netherlands and were asked if they were willing to participate in a study on news facts and societal issues. If they agreed, they were led to a quiet and private room in a coffee house, where four laptops were installed.

The current study was combined with another study on the use of narratives in news messages (topically unrelated). Participants filled out the questionnaires of the other study on paper, while this experiment was conducted on the computer. The order of participation was randomised: Roughly half of the participants started with the questionnaire, the other half with the experiment on the laptops. The total duration of the experiment was 35–45 minutes.

After completing the experiment, participants were thanked and were offered a gift certificate of 10 Euros for their participation. Participants were asked to leave their e-mail address for debriefing, as other participants might still be working on the experiment in the same room.

2.2 Material

Pragmatic arguments were developed in support of 24 different claims. For each of these claims, four variants of a supporting argument were developed. Claim-argument combinations consisted of 23 to 40 words, with a mean of about 33 words. The claims were about new policy proposals that were relatively neutral or that did not concern participants directly (e.g., creating a skate park or using more electrical cars in another city) to minimise the chance that participants would already have strong attitudes towards the claims beforehand.

The desirability and probability of the consequences referred to in these arguments were manipulated to create the four versions of the same argument. Both factors had two levels: high and low. For desirability, the degree to which the measure had a positive consequence was manipulated (high vs. low desirability). For instance, it was stated that a measure had "very large positive effects" or "a significant influence" versus "some effects" or "a marginal influence".

For the probability dimension, the quality of the evidence in support of the claim about the likelihood of the consequence was manipulated, following previous manipulations that were based on normative criteria from argumentation theory (Hoeken et al., 2012, 2014; Hoeken & Hustinx, 2009; Hornikx & Hoeken, 2007). Three types of evidence were used: expert evidence, statistical evidence, and anecdotal evidence. The quality of expert evidence was manipulated by ascribing information to either an impartial or a biased expert. Statistical evidence was manipulated by providing (fictitious) data that captured the likelihood of the consequence with a larger versus smaller number. Anecdotal evidence was manipulated by providing a case that was either highly similar or dissimilar with the case in the claim. Examples of arguments can be found in Table 1.

It is possible that the order in which the desirability and probability information appears in the argument has an influence on participants' evaluations. Participants may only read the first or the last

Table 1: Example arguments for the three evidence types, with probability (bold) and desirability (italic) manipulations

Expert	Claim	"Eco-friendly washing machines should be subsidised."
	Argument	"This will lead to a *significant (strong) / slight (weak)* decrease of domestic water bills, according to research by the **European Parliamentary Committee for Industry, Research and Energy (strong) / washing machine manufacturer Miele (weak)**"
Statistical	Claim	"There should be standardised protocols for school medical officers to help recognise child abuse and neglect."
	Argument	"In *most (strong) / some (weak)* cases, this will prevent long term damages. Early detection of abuse reduces the chance of long term damages by **83 % (strong) / 7 % (weak).**"
Anecdotal	Claim	"Leiden University should create more group workspaces."
	Argument	"Working together increases the chance of good grades *a lot (strong) / somewhat (weak).* Since the **University of Amsterdam (strong) /high school in Doetinchem (weak)** did this, performance of students has improved."

part of the sentence thoroughly. Therefore, the order of the desirability and probability information in the argument was manipulated, with half of the participants receiving the arguments with desirability first (see Table 1), the other half with probability first.

In total, there were 192 unique claim-argument combinations: 24 claims with 4 supporting arguments, with 2 versions for each argument that only differed with respect to the order in which the probability and the desirability information was presented.

2.3 Design

The design of the study was a 2 (Argument desirability: high, low) x 2 (Argument probability: high, low) x 2 (Order: desirability-probability, probability-desirability) mixed design with desirability and probability as within-subject factors and order as a between-subject factor.

A Latin square design was employed, to ensure that each claim-argument combination was rated by an equal number of participants (see Appendix, Table A1). There were 24 different versions of the experiment, which was programmed in Inquisit version 4.0.0.

Participants were randomly assigned to a version, with the restriction that there had to be at least 5 participants per version. Participants first rated claim acceptance for 8 claims and supporting arguments, then provided intuitive argument evaluations for 8 different claims and arguments, and finally gave their explicit argument evaluations of again 8 different claims and arguments. The tasks were always presented in this order to make sure that the influence the tasks could have on each other was kept to a minimum. Starting with the explicit evaluation task would familiarise participants with the construct of argument strength too much, making their intuitive evaluations less intuitive. In a similar vein, if participants were to end with the claim acceptance task after extensively evaluating arguments, they would probably become much more critical of the arguments than they normally would be.

2.4 Dependent variables

The three dependent variables of interest in this study were claim acceptance (the degree to which participants agree with a claim), explicit argument evaluation (a thorough examination of the strength of an argument), and intuitive argument evaluation.

2.4.1 Claim acceptance

Claim acceptance was measured with two items on a 7-point Likert scale, ranging

from 1 (not at all) to 7 (very much), $r = .84$. The items were "How much do you agree with the statement?" and "How sensible do you think the proposed measure is?" These statements were customised for each claim. For example, if the claim was: "There should be more rest areas along the highway", the items were: "How much do you agree with providing more rest areas along the highway?" and "How sensible do you think it is to create more rest areas along the highway?"

2.4.2 Explicit evaluation

Explicit evaluation was measured with eight items on a 7-point Likert scale, ranging from 1 (not at all) to 7 (very much), $\alpha = .87$. The items, translated into Dutch, were taken from the PAS scale from Zhao et al. (2011). The original scale consists of 9 statements with 5-point Likert scales. In the current study, a 7-point scale was used in order to better detect subtle differences in evaluations and to keep the scale length of all three dependent variables equal. Item 5 of the original PAS scale was omitted since this item was irrelevant to the arguments used in the current study. Item 5 measures perceptions of the behaviour of others, such as "The statement would help my friends quit smoking".

2.4.3 Intuitive evaluation

The intuitive evaluation task was a forced-choice task. Participants were asked to evaluate the strength of the arguments by categorising them as either "strong" or "weak" as quickly as possible. The claim-argument combinations appeared in the middle of the screen and participants' only task was to press the "e" key on the keyboard if they thought the argument was weak, and the "i" key if they thought the argument was strong. After pressing the "e" or the "i" key, the next argument appeared immediately. To familiarise participants with this procedure, participants were presented with four practice trials, responses of which were not taken into account in the analyses. To ensure that participants answered quickly, a time constraint was added. The time constraint applied to the entire set of arguments participants had to evaluate as either weak or strong. Based on the study by Evans and Curtis-Holmes (2005), the time necessary to read and categorise an argument was estimated to be 10 seconds. A pretest was conducted on a student sample, where participants were asked to respond to 10 claim-argument combinations within 80 seconds, to see if they would be able to answer 8 or more items within this time frame. The results from the pretest showed that on average, participants were able to categorize 8.3 arguments within 80 seconds and the average response time per argument was 9.1 seconds. However, 30 % of participants were not able to respond to eight arguments in time. Therefore, the maximum amount of time for the intuitive measure was increased to 100 seconds in total to prevent missing data. The keys were locked for the first 3 seconds the arguments were on the screen, which prevented participants from pressing the keys without reading. A reminder to answer appeared after 10 seconds. After the intuitive evaluations were recorded, participants were asked to report the extent to which they experienced time pressure, and the extent to which they felt they had been able to read the arguments on 7-point Likert scales ranging from 1 (not at all) to 7 (very much).

3 Results

The two argument strength evaluation methods should provide an indication of the strength of an individual argument. Therefore, for each argument the average score was computed of all participants rating that argument. This resulted in an average indication of argument strength using an explicit scale, an indication of argument strength using an intuitive scale, as well as the extent to which the claim supported by this argument was accepted.

All analyses were conducted on average scores per individual argument, based on approximately 5 participants per item. Intuitive evaluations are expressed as the proportion of participants that indicated

that the argument was strong. A higher proportion reflects higher perceived argument strength. The choice for the argument as the basic unit of analysis is in line with recent critical observations made by O'Keefe (2020). He argues that the comparison between perceived message effectiveness (i. e., perceived argument strength) and actual message effectiveness (i. e., claim acceptance) can only be made by comparing a message's score on both types of effectiveness – and not by comparing the two scores for a set of individual participants, as is often done in research (Dillard, Weber, & Vail, 2007).

The order of the probability and desirability information did not affect the main results (interaction order and desirability on claim acceptance: $F(1, 23) = 1.09$, $p = .308$, on the explicit measure: $F(1, 23) = 1.09$, $p = .308$, on the implicit measure: $F(1, 23) < 1$; interaction order and probability on claim acceptance: $F(1, 23) < 1$, on the explicit measure: $F(1, 23) = 3.16$, $p = .089$, on the implicit measure: $F(1, 23) = 1.16$, $p = .293$). Therefore, order was not analysed as a separate factor in the results below. All subsequent analyses that are reported below were conducted on the entire sample.

3.1 Manipulation check intuitive measure

Based on scores on the two relevant measures, the manipulation of the intuitive measure was considered successful. First, participants were asked if they felt time pressure, which they did ($M = 4.93$, $SD = 1.62$, significantly higher than the scale midpoint, $t(120) = 6.36$, $p < .001$). Second, they were asked how well they were able to read the arguments, which was reasonably well ($M = 4.48$, $SD = 1.58$, significantly higher than the scale midpoint, $t(116) = 3.33$, $p = .001$). Taken together, there was time enough to read the arguments, as well as enough limitation of time to feel time pressure.

3.2 Predicting claim acceptance from intuitive and explicit evaluations (RQ1)

The predictive validity of the two types of evaluations was tested in a regression analysis, with claim acceptance as outcome variable. Prior to this analysis, it was established that, while explicit evaluation correlated with claim acceptance ($r = .77$, $p < .001$), intuitive evaluation did not ($r = –.14$, $p = .510$), and that there was not a significant correlation between explicit and intuitive evaluations ($r = –.135$, $p = .528$). Next, both explicit and intuitive evaluation scores were entered as predictors at the same time. In response to RQ1, the model with both explicit and intuitive evaluations explained a significant part of the variance, $F(2, 93) = 22.09$, $p < .001$, $R^2 = .32$. Explicit evaluations were a significant predictor ($\beta = .55$, $p < .001$), meaning that higher claim acceptance was predicted by more positive explicit evaluations. Intuitive evaluations were found not to be a significant predictor ($p = .10$).

Table 2: Claim acceptance, intuitive evaluations and explicit evaluations as a function of desirability and probability

	Desirability	Probability			
		High		Low	
		M	*(SD)*	*M*	*(SD)*
Claim acceptance	High	4.89	(0.73)	4.80	(0.79)
	Low	4.92	(0.81)	4.80	(0.71)
Intuitive evaluations	High	.64	(.22)	.61	(.20)
	Low	.58	(.24)	.48	(.25)
Explicit evaluations	High	4.47	(0.82)	3.96	(0.58)
	Low	4.06	(0.53)	3.93	(0.63)

3.3 Differences in sensitivity to dimensions of argument strength (RQ2)

To examine whether the two dimensions of argument strength had an influence on the dependent variables (RQ2; construct validity), three separate repeated measures ANOVAs were conducted with desirability and probability as within-argument factors, and claim acceptance, intuitive evaluations and explicit evaluations as dependent variables, respectively (see Table 2 for mean scores of the dependent variables for the different levels of desirability and probability).

For claim acceptance, the main effects of desirability ($F(1, 23) < 1$) and probability ($F(1, 23) = 2.36$, $p = .14$) as well as the interaction effect between the two were not significant ($F(1, 23) < 1$).

For intuitive evaluations, there was a significant main effect of desirability, $F(1, 23) = 8.15$, $p < .01$, $\eta^2 = .26$: Intuitive evaluations were more positive when desirability was high ($M = .63$, $SD = .19$) than when desirability was low ($M = .53$, $SD = .21$). There was also an effect of probability on intuitive evaluations, $F(1, 23) = 4.35$, $p < .05$, $\eta^2 = .16$: Intuitive evaluations were more positive when probability was high ($M = .61$, $SD = .21$) than when probability was low ($M = .54$, $SD = .19$). The interaction between desirability and probability was not significant: $F(1, 23) = 1.23$, $p = .28$.

For explicit evaluations, the main effects of desirability ($F(1, 23) = 12.08$, $p < .01$, $\eta^2 = .34$), and probability ($F(1, 23) = 22.81$, $p < .001$, $\eta^2 = .50$) were significant, with high desirability ($M = 4.22$, $SD = 0.49$) leading to higher explicit evaluations than low desirability ($M = 4.00$, $SD = 0.53$), and high probability ($M = 4.27$, $SD = 0.45$) leading to higher explicit evaluations than low probability ($M = 3.95$, $SD = 0.57$). These main effects were qualified by their interaction, $F(1, 23) = 6.26$, $p < .05$, $\eta^2 = .21$. The effect of desirability was only found in the case of high probability.

4 Conclusion and discussion

Methods for evaluating argument strength have participants reflect on the arguments at hand – implying that analytical *Type 2* processing is needed to accurately assess the merits of an argument (Cacioppo & Petty, 1981; Zhao et al., 2011). Mercier and Sperber (2009, 2011, 2017) have argued that people may intuitively be successful at distinguishing strong from weak arguments. The current study therefore compared a novel, intuitive measure of argument strength to a conventional, explicit measure of argument strength. These measures were compared in terms of their construct and predictive validity.

The first research question addressed in this paper was to what extent an intuitive evaluation method, compared to an explicit evaluation method, was able to predict claim acceptance (predictive validity). Intuitive evaluations, measured under time constraint, were intended to capture intuitive inferences as the outcome of a fast and intuitive *Type 1* process (Evans, 2010). Results showed that these intuitive evaluations were not capable of predicting claim acceptance whereas the explicit measures were.

The second research question was whether the intuitive and the explicit methods would prove sensitive to manipulations of argument quality in terms of desirability and probability (construct validity). Variations in desirability and probability of the consequences mentioned in the argument have been shown to be important determinants of argument strength in support of claims in previous empirical research (Areni & Lutz, 1988; Johnson et al., 2004; Van Enschot-Van Dijk, Hustinx, & Hoeken, 2003). Results first showed that explicit argument evaluations were sensitive to these manipulations. Arguments that were manipulated to have a higher probability and / or a higher desirability were perceived by participants as stronger arguments than when they were manipulated to have a lower probability and / or desirability. This finding is consistent with results obtained in other experimental studies, in which claim acceptance was

higher when claims were supported by arguments with a higher probability and / or a higher desirability (Hoeken & Hustinx, 2009; Hoeken et al., 2012, 2014; Hornikx & Hoeken, 2007). The current results further demonstrate that intuitive argument evaluations are also sensitive to these manipulations of argument strength, suggesting support for the construct validity of the intuitive measure.

Taking together, the results regarding RQ1 and RQ2 do not provide evidence for the validity of the intuitive measure as it was operationalised in the current experiment. In fact, based on the results, the validity of the current intuitive measure is questionable, and there are at least two reasons for this conclusion. In the first place, the intuitive measure was not correlated with the explicit measure. Whether participants judged an argument more or less strong under *explicit* evaluation was not associated with a varying *intuitive* evaluation of the same argument. As the explicit evaluation measure was found to have predictive validity, this non-significant correlation raises serious concerns about the validity of the novel intuitive measure. In the second place, there was a surprising non-effect of desirability and probability on claim acceptance. Whereas such an effect has been observed in many studies (Hoeken & Hustinx, 2009; Hoeken et al., 2012, 2014; Hornikx & Hoeken, 2007), it did not reach statistical significance in this study. The consequence of this non-effect is that it qualifies the effect that was observed of desirability and probability on the intuitive measure. Although this effect in itself should be taken as evidence in support of the construct validity of the intuitive measure, the non-effect of desirability and probability on claim acceptance puts reasonable doubt on this construct validity.

4.1 Future research

There are different avenues for future research. First and foremost, it is essential to test other intuitive measures of argument strength. In the current study, the intuitive measure was operationalised in two ways: through a time constraint, and through the dichotomous answer "strong" versus "weak". A potential avenue for future research lies in the operationalisation of the time constraint. Ideally, the time constraint should be based on each individual's personal reading speed, to ensure comparable levels of processing and to prevent missing items. This would imply that a personal reading speed should be reliably determined prior to the actual experiment, and implemented in the experiment. Independent of the time constraint, there are other ways of asking an intuitive judgment of the strength of an argument than to judge it as "strong" or "weak". A more intuitive way of asking strength may be through the use of an emoji scale expressing negative, neutral, and positive emotions (Bai, Dan, Mu, & Yang, 2019).

Next, the current study is limited to the specific manipulations of the arguments. First, this study only manipulated the *magnitude* of the desirability of the consequence but not the *nature* of the consequence in the pragmatic arguments. In other studies, the consequences for the more or less desirable outcome were often qualitatively rather than quantitatively different (e. g., "better grades" vs. "a mental challenge", Petty & Cacioppo, 1986). Using qualitatively different outcomes may be a better way to capture explicit and intuitive measures of argument strength. Second, three types of evidence were employed to support the probability of the consequences: expert evidence, statistical evidence, and anecdotal evidence. Other evidence may have been used, particularly causal evidence, following the four types of evidence generally used in persuasion research (Hornikx, 2005). Third, the manipulation of strong and weak variants of this evidence was identical for each type of evidence. For instance, for expert evidence, the strong expert was impartial whereas the weak expert was biased. Following the critical questions for appeals to expert opinions (Walton et al., 2008), other choices can be made to manipulate the quality of these arguments, such as by varying the level of expertise of the expert or the relevance of this expertise for the topic of the claim. In sum, while a strength of the ex-

perimental design was the use of multiple claims and different types of evidence to manipulate the quality of the arguments, it should be noted that the specific choices may have affected the outcomes.

Finally, conducting a study using eye-tracking could teach us more about the sensitivity to desirability or probability information, as eye-tracking makes it possible to see on which words or parts of the sentence people fixate on (or go back to) when reading. This can help gaining more insight into the effects of presentation of information and might indicate which parts of the sentence are mainly used as input for explicit versus intuitive processing.

Altogether, these are interesting avenues for future research that could result in an improved intuitive argument evaluation measure, with the aim of exploring its predictive and construct validity.

4.2 Implications

There are two implications based on the findings in this study. First, the fact that explicit argument evaluations were found to be related to claim acceptance adds to the growing body of literature on the reliability and predictive validity of the perceived argument strength (PAS) scale (Bigsby et al., 2013; Shi, Messaris, & Cappella, 2014; Zhao et al., 2011). The results of this study support the validity of the PAS scale.

Second, although the intuitive argument evaluation measure was not successful in the current experimental study, the aim of developing such a measure is still relevant. Given that intuitive inferences are very common in everyday life and people are not always inclined to deliberately evaluate the strength of the arguments they encounter, a measure that captures these intuitive evaluations is very useful in pretesting or evaluating arguments that are used in health behaviour campaigns, policy issues or public service announcements. As a consequence, more research examining the potential of intuitive argument evaluation methods should be welcomed.

Acknowledgements

The study reported in this paper was completed with the support of the Centre for Language Studies (CLS), Radboud University Nijmegen.

Conflict of interests

The authors declare no conflict of interests.

References

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

Bai, Q., Dan, Q., Mu, Z., & Yang, M. (2019). A systematic review of emoji: Current research and future perspectives. *Frontiers in Psychology, 10*(2221), 1–16. https://doi.org/10.3389/fpsyg.2019.02221

Bigsby, E., Cappella, J. N., & Seitz, H. H. (2013). Efficiently and effectively evaluating public service announcements: Additional evidence for the utility of perceived effectiveness. *Communication Monographs, 80*(1), 1–23. https://doi.org/10.1080/03637751.2012.739706

Cacioppo, J. T., & Petty, R. E. (1981). Social psychological procedures for cognitive response assessment: The thought-listing technique. In T. V. Merluzzi, C. R. Glass, & M. Genest (Eds.), *Cognitive Assessment* (pp. 309–342). New York, NY: Guilford Press.

Cappella, J. N. (2018). Perceived message effectiveness meets the requirements of a reliable, valid, and efficient measure of persuasiveness. *Journal of Communication, 68*(5), 994–997. https://doi.org/10.1093/joc/jqy044

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York, NY: Guilford Press.

De Jong, M. (1998). *Reader feedback in text design: Validity of the plus-minus method for the pretesting of public information*

brochures. Amsterdam, The Netherlands: Rodopi.

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science, 28*(5), 503–509. https://doi.org/10.1177/0963721419855658

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*(4), 613–631. https://doi.org/10.1111/j.1460-2466.2007.00360.x

Evans, J. St. B. T. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry, 21*(4), 313–326. https://doi.org/10.1080/1047840X.2010.521057

Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking and Reasoning, 11*(4), 382–389. https://doi.org/10.1080/13546780542000005

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Gawronski, B., & Creighton, L. A. (2013). Dual-process theories. In D. E. Carlston (Ed.), *The Oxford handbook of social cognition* (pp. 282–312). New York; NY: Oxford University Press.

Hastings, A. C. (1962). *A reformulation of the modes of reasoning in argumentation* (dissertation). Speech-Theater, Northwestern University, Evanston, IL ,USA.

Hoeken, H., Hornikx, J., & Linders, Y. (2020). The importance and use of normative criteria to manipulate argument quality. *Journal of Advertising, 49*(2), 195–201. https://doi.org/10.1080/00913367.2019.1663317

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*(4), 491–510. https://doi.org/10.1111/j.1468-2958.2009.01360.x

Hoeken, H., Šorm, E., & Schellens, P. J. (2014). Arguing about the likelihood of consequences: Laypeople's criteria to distinguish strong arguments from weak ones. *Thinking and Reasoning, 20*(1), 77–98. https://doi.org/10.1080/13546783.2013.807303

Hoeken, H., Timmers, R., & Schellens, P. J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking and Reasoning, 18*(3), 394–416. https://doi.org/10.1080/13546783.2012.669986

Hornikx, J. (2005). A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Studies in Communication Sciences (SComS), 5*(1), 205–216.

Hornikx, J., & Hoeken, H. (2007). Cultural differences in the persuasiveness of evidence types and evidence quality. *Communication Monographs, 74*(4), 443–463.

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.

Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 146–170). Oxford, UK: Oxford University Press.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, MA: Harvard University Press.

Noar, S. M., Bell, T., Kelley, D., Barker, J., & Yzer, M. C. (2018). Perceived message effectiveness measures in tobacco education campaigns: A systematic review. *Communication Methods and Measures, 12*(4), 295–313. https://doi.org/10.1080/19312458.2018.1483017

O'Keefe, D. J. (2007). Potential conflicts between normatively-responsible advocacy and successful social influence: Evidence from persuasion effects research. *Argu-*

mentation, 21(2), 151–163. https://doi.org/10.1007/s10503-007-9046-y

O'Keefe, D. (2013). The relative persuasiveness of different forms of arguments-from-consequences: A review and integration. *Annals of the International Communication Association, 36*(1), 109–135. https://doi.org/10.1080/23808985.2013.11679128

O'Keefe, D. J. (2018). Message pretesting using assessments of expected or perceived persuasiveness: Evidence about diagnosticity of relative actual persuasiveness. *Journal of Communication, 68*(1), 120–142. https://doi.org/10.1093/joc/jqx009

O'Keefe, D. J. (2020). Message pretesting using perceived persuasiveness measures: Reconsidering the correlational evidence. *Communication Methods and Measures, 14*(1), 25–37. https://doi.org/10.1080/19312458.2019.1620711

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

Schellens, P. J., & De Jong, M. (2004). Argumentation schemes in persuasive brochures. *Argumentation, 18*(3), 295–323. https://doi.org/10.1023/B:ARGU.0000046707.68172.35

Shi, R., Messaris, P., & Cappella, J. N. (2014). Effects of online comments on smokers' perception of antismoking public service announcements. *Journal of Computer-Mediated Communication, 19*(4), 975–990. https://doi.org/10.1111/jcc4.12057

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Van Eemeren, F. H., & Grootendorst, R. (1992). *Argumentation, communication and fallacies: A pragma-dialectical perspective.* Hillsdale, NJ: Lawrence Erlbaum.

Van Enschot-Van Dijk, R., Hustinx, L., & 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, J. A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.), *Anyone who has a view. Theoretical contributions to the study of argumentation* (pp. 319–335). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning.* Mahwah, NJ: Lawrence Erlbaum.

Walton, D. N., Reed, C., & Macagno, F. (2008). *Argumentation schemes.* Cambridge, MA: 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*(4), 787–797. https://doi.org/10.1086/426613

Whittingham, J., Ruiter, R. A. C., Zimbile, F., & Kok, G. (2008). Experimental pretesting of public health campaigns: A case study. *Journal of Health Communication, 13*(3), 216–229. https://doi.org/10.1080/10810730701854045

Zhao, X., Strasser, A., Cappella, J. N., Lerman, C., & Fishbein, M. (2011). A measure of perceived argument strength: Reliability and validity. *Communication Methods and Measures, 5*(1), 48–75. https://doi.org/10.1080/19312458.2010.547822

Appendix

Table A1: Latin square design: Division of arguments over version and dependent variables

Version	Claim acceptance	Intuitive evaluations	Explicit evaluations
1/13	1a 2b 3c 4d	9a 10b 11c 12d	17a 18b 19c 20d
	5a 6b 7c 8d	13a 14b 15c 16d	21a 22b 23c 24d
2/14	1b 2c 3d 4a	9b 10c 11d 12a	17b 18c 19d 20a
	5b 6c 7d 8a	13b 14c 15d 16a	21b 22c 23d 24a
3/15	1c 2d 3a 4b	9c 10d 11a 12b	17c 18d 19a 20b
	5c 6d 7a 8b	13c 14d 15a 16b	21c 22d 23a 24b
4/16	1d 2a 3b 4c	9d 10a 11b 12c	17d 18a 19b 20c
	5d 6a 7b 8c	13d 14a 15b 16c	21d 22a 23b 24c
5/17	9a 10b 11c 12d	17a 18b 19c 20d	1a 2b 3c 4d
	13a 14b 15c 16d	21a 22b 23c 24d	5a 6b 7c 8d
6/18	9b 10c 11d 12a	17b 18c 19d 20a	1b 2c 3d 4a
	13b 14c 15d 16a	21b 22c 23d 24a	5b 6c 7d 8a
7/19	9c 10d 11a 12b	17c 18d 19a 20b	1c 2d 3a 4b
	13c 14d 15a 16b	21c 22d 23a 24b	5c 6d 7a 8b
8/20	9d 10a 11b 12c	17d 18a 19b 20c	1d 2a 3b 4c
	13d 14a 15b 16c	21d 22a 23b 24c	5d 6a 7b 8c
9/21	17a 18b 19c 20d	1a 2b 3c 4d	9a 10b 11c 12d
	21a 22b 23c 24d	5a 6b 7c 8d	13a 14b 15c 16d
10/22	17b 18c 19d 20a	1b 2c 3d 4a	9b 10c 11d 12a
	21b 22c 23d 24a	5b 6c 7d 8a	13b 14c 15d 16a
11/23	17c 18d 19a 20b	1c 2d 3a 4b	9c 10d 11a 12b
	21c 22d 23a 24b	5c 6d 7a 8b	13c 14d 15a 16b
12/24	17d 18a 19b 20c	1d 2a 3b 4c	9d 10a 11b 12c
	21d 22a 23b 24c	5d 6a 7b 8c	13d 14a 15b 16c

Note. Versions 1–12 had the order desirability-probability, while versions 13–24 had the order probability-desirability. Label "a": high desirability, high probability; label "b": high desirability, low probability; label "c": low desirability, high probability; label "d": low desirability, low probability.