# Modeling the Ellsberg Paradox by Argument Strength

**Niki Pfeifer (niki.pfeifer@lmu.de)**
Munich Center for Mathematical Philosophy, LMU Munich, Germany

**Hanna Pankka (hanna.pankka@helsinki.fi)**
Department of Philosophy, History, Culture and Art Studies, University of Helsinki, Finland

**Abstract**

We present a formal measure of argument strength, which combines the ideas that conclusions of strong arguments are (i) highly probable and (ii) their uncertainty is relatively precise. Likewise, arguments are weak when their conclusion probability is low or when it is highly imprecise. We show how the proposed measure provides a new model of the Ellsberg paradox. Moreover, we further substantiate the psychological plausibility of our approach by an experiment ($N = 60$). The data show that the proposed measure predicts human inferences in the original Ellsberg task and in corresponding argument strength tasks. Finally, we report qualitative data taken from structured interviews on folk psychological conceptions on what argument strength means.

**Keywords:** argument strength; coherence; Ellsberg paradox; probability logic

## Introduction

### Measuring Argument Strength

Probabilistic models of argumentation became popular in cognitive science and its subfields including psychology, philosophy, and computer science in recent years (see, e.g., Hahn & Oaksford, 2006; Haenni, 2009; Zenker, 2013). Like logic-based nonmonotonic approaches for defeasible argumentation (see, e.g. Prakken & Vreeswijk, 2002), probabilistic approaches allow for dealing with exceptions and retracting conclusions in the light of new evidence. However, in contrast to qualitative logical approaches, probability allows for managing *degrees of belief* in the sentences involved in common sense argumentation. Moreover, degrees of belief can be used to model the strength of arguments (Hahn & Oaksford, 2006; Oaksford & Hahn, 2007; Pfeifer & Kleiter, 2006; Pfeifer, 2007, 2013b).

The concept "argument" is ambiguous. In logic, it denotes a triple consisting of a (possibly empty) premise set, a conclusion indicator, and a conclusion set. Consider, for example, the following argument, which is an instance of modus ponens:

(P1) If I take the train at five ($T$), I'll be home at six ($H$).
(P2) I take the train at five ($T$).
(C) Therefore, I'll be home at six ($H$).

Here, (P1) and (P2) are the premises, "Therefore" the conclusion indicator and the sentence "I'll be home at six" is the conclusion. In argumentative contexts, "argument" may also denote a premise which speaks for or against a conclusion. For example "The train conductors are on strike", can serve as an argument for concluding that it is better to take the bus. In what follows, however, we will focus on arguments in the logical sense only.

How can we measure the strength of an argument? There are at least two formal approaches to study (probabilistic) argument strength. In the first approach argument strength is based on *uncertain consequence* relations, i.e., by presupposing that the conclusion follows to some degree from the premises. Usually, this is modeled by a conditional probability of "the conclusion *given* (some combination of) the premises" of the argument (see, e.g. Hahn & Oaksford, 2006; Oaksford & Hahn, 2007). As pointed out by Osherson, Smith, Wilkie, López, and Shafir (1990), measures of confirmation can serve as models for argument strength (for an overview of measures of confirmation see Crupi, Tentori, & Gonzales, 2007). Measures of confirmation and previous attempts to model argument strength by uncertain consequence relations are problematic when arguments involve conditionals, like the modus ponens above (see premise (P1)): it is far from clear to give a precise meaning of conditionalizing on a combination of premises, when the premise set contains conditional events. There is ample formal and experimental evidence that uncertain conditionals are best modeled by conditional probabilities (see, e.g., Evans & Over, 2004; Oaksford & Chater, 2007; Over & Cruz, in press; Pfeifer, 2014, 2013a). Therefore, conditionals should be modeled by conditional probabilities. However, this requirement would imply to measure the uncertainty of a *conclusion given (some combination of) the premises*. Unfortunately, satisfactory semantics of expressions like $\overbrace{C}^{\text{conclusion}} \,|\, \overbrace{(A \text{ and } (C|A))}^{\text{premises}}$ do not exist yet. Such semantics would, however, be necessary to capture the underlying logical structure of the modus ponens (for an approach which deals with nested conditionals and which avoids Lewis' triviality results, see Gilio, Over, Pfeifer, & Sanfilippo, 2017; Sanfilippo, Pfeifer, & Gilio, 2017). Modus ponens is just a relatively simple example here: there are, of course, many other argument forms involving conditionals. The inability to deal with conditionals seems to us to be one of the main reasons, why currently no formally satisfactory measure of argument strength exists within the first approach: measures based on uncertain consequence relations do not seem to be able to deal with the logical form of the argument.

In this paper, we advocate the second approach to argument strength. It satisfies the requirement of doing justice to the logical form of arguments involving conditionals (Pfeifer, 2007, 2013b). Specifically, we define argument strength based on the following ideas: (i) keep the consequence relation deductive, (ii) assign probabilities to the premises, and

then (iii) define the measure of argument strength based on the propagated coherent lower and upper probability bounds on the conclusion (Pfeifer, 2007, 2013b). Probability propagation from the premises to the conclusion is governed by *coherence based probability logic* (see, e.g. Coletti & Scozzafava, 2002; Pfeifer & Kleiter, 2009; Gilio, Pfeifer, & Sanfilippo, 2016). The *coherence approach* to probability was originated by Bruno de Finetti (de Finetti, 1970/1974). It conceives probabilities as subjective degrees of belief. Conditional probabilities ($p(C|A)$) are primitive. This allows for zero probabilities of the conditioning event ($A$). Note that in standard approaches to probability, $p(C|A)$ is undefined if $p(A) = 0$, which is problematic in many argument forms (see, e.g. Pfeifer, 2014; Gilio et al., 2016). Moreover, coherence allows for managing *imprecise probabilities* (set-valued probabilities involving lower and upper probability bounds), which is relevant for formalising arguments under incomplete probabilistic knowledge. The above mentioned modus ponens, for example, is formalised as follows (see, e.g. Pfeifer & Kleiter, 2009, Example 1, p. 209):

(P1') $p(H|T) = x$
(P2') $p(T) = y$
(C') Therefore, $z' \leq p(H) \leq z''$, where $z' = xy$ and $z'' = xy + 1 - y$ are the best possible coherent probability bounds on the conclusion.

Following Pfeifer (2013b), we define the measure of argument strength $\mathfrak{s}$ on an argument $\mathcal{A}$ as follows:

> Let $z'$ and $z''$ denote the coherent lower and upper probability bounds, respectively, on the conclusion of argument $\mathcal{A}$. Then,
>
> $$\mathfrak{s}(\mathcal{A}) =_{\text{def.}} \overbrace{(1-(z''-z'))}^{\text{precision}} \times \frac{\overbrace{z'+z''}^{\text{location}}}{2}. \qquad (1)$$

Intuitively, measure $\mathfrak{s}$ combines the *precision* and the *location* of the coherent conclusion probability interval. Specifically, strong arguments are arguments with *low imprecision* of the conclusion probability (measured by the one-complement of the distance between the upper and the lower probability bounds, $1-(z''-z')$) and with conclusion probabilities *close to one* (measured by the mean of the lower and upper probability bound, $(z'+z'')/2$). Weak arguments are characterized by a large conclusion interval (i.e., high imprecision) or by a low-probability conclusion (i.e., the center point of the conclusion interval is close to 0). For a discussion of how logical validity relates to whether the degree of belief in the conclusion is constrained by the assessment of the premises see, e.g., Pfeifer and Kleiter (2009). Of course, precision and location could be modeled differently (e.g., by using the geometric or the harmonic mean instead of the arithmetic mean). Moreover, in contexts where the location is more important than the precision of the conclusion probability interval (or *vice versa*), adding suitable weights to formula (1) can adjust the measure for such cases. However, for the purpose of our paper it is sufficient to keep the measure as simple as possible.

Measure $\mathfrak{s}$ has a number of plausible consequences: it ranges always from zero to one (i.e., $0 \leq \mathfrak{s} \leq 1$, since $z'$ and $z''$ are probability values, which are also in the unit interval, $[0,1]$). The extreme "0" denotes weak arguments and "1" denotes strong arguments. Arguments with conclusion probability 1, are strong arguments, since $\mathfrak{s} = 1$ if $z' = z'' = 1$. Arguments with conclusion probability 0 (i.e., $z' = z'' = 0$) are weak arguments, since $\mathfrak{s} = 0$. Likewise, probabilistically non-informative arguments (i.e., $z' = 0$ and $z'' = 1$) are weak arguments, since $\mathfrak{s} = 0$.

Interestingly, measure $\mathfrak{s}$ also provides a new solution to the Ellsberg paradox (Ellsberg, 1961),[1] which we describe in the next section.

## Modeling the Ellsberg Paradox by Measure $\mathfrak{s}$

Ellsberg described the following situation (Ellsberg, 1961):

> An urn contains 90 balls, of which 30 are red ($R$) and 60 are black or yellow. The ratio of the black and yellow balls is unknown—there might be anything between 0 to 60 black (or yellow) balls. One ball is drawn at random from the urn and you are asked to choose a bet between two bets. If you take **Bet 1**, you will win \$100, if the ball drawn from the urn is red. If you take **Bet 2**, you will win \$100, if the ball drawn from the urn is black.

Ellsberg predicted that most people choose **Bet 1** when asked to decide which of the two bets they prefer. Then, considering again the same urn, Ellsberg predicted that people will choose **Bet 4**, when they are asked to decide between the following two alternative bets:

> If you take **Bet 3**, you will win \$100, if the ball drawn from the urn is red or yellow. If you take **Bet 4**, you will win \$100, if the ball drawn from the urn is black or yellow.

Ellsberg's predictions create a well-known paradox as they violate the independence axiom of rational choice (see, e.g., Briggs, 2016). Moreover, Ellsberg's predictions were experimentally confirmed in many studies (see, e.g., Becker & Brownson, 1964; Slovic & Tversky, 1974; MacCrimmon & Larsson, 1979).

We propose to frame the Ellsberg paradox in terms of probability logical arguments. Specifically, the *premises* represent the probabilistic information given in the description of the urn, and the *conclusions* represent the respective bets involved in the Ellsberg paradox. Thus, we obtain four arguments. Each argument speaks for choosing the corresponding bet. The associated argument to **Bet 2**, for example, is argument $\mathcal{A}_2$ (where "$\vee$" denotes *disjunction* ("or") and $R$, $B$ and $Y$ are mutually exclusive):

$p(R) = .33$
$p(B \vee Y) = .67$
Therefore, $0 \leq p(B) \leq .67$ is coherent.

[1] We thank Kevin T. Kelly for pointing us to the Ellsberg paradox.

The strength of this argument is denoted by $\mathfrak{s}(\mathcal{A}_2)$ and by applying equation (1) equal to .11 (i.e., $\mathfrak{s}(\mathcal{A}_2) = .11$). Table 1 lists the conclusions and the argument strengths $\mathfrak{s}$ for each argument for the corresponding four bets involved in the Ellsberg paradox.

Table 1: Conclusions and normative strengths ($\mathfrak{s}$) of Arguments $\mathcal{A}_1, \ldots, \mathcal{A}_4$ associated with the four bets involved in the Ellsberg paradox. The premises are always $p(R) = .33$ and $p(B \vee Y) = .67$.

| | Conclusion | Argument strength |
|---|---|---|
| **Bet 1** | $p(R) = .33$ | $\mathfrak{s}(\mathcal{A}_1) = .33$ |
| **Bet 2** | $0 \leq p(B) \leq .67$ | $\mathfrak{s}(\mathcal{A}_2) = .11$ |
| **Bet 3** | $.33 \leq p(R \vee Y) \leq 1$ | $\mathfrak{s}(\mathcal{A}_3) = .22$ |
| **Bet 4** | $p(B \vee Y) = .67$ | $\mathfrak{s}(\mathcal{A}_4) = .67$ |

The four argument strength values in Table 1 induce the following preference orders in the classical Ellsberg task: **Bet 1** $\succ$ **Bet 2**, since $\mathfrak{s}(\mathcal{A}_1) = .33 > \mathfrak{s}(\mathcal{A}_2) = .11$, and **Bet 4** $\succ$ **Bet 3**, since $\mathfrak{s}(\mathcal{A}_4) = .67 > \mathfrak{s}(\mathcal{A}_3) = .22$ (where $X \succ Y$ denotes *X is preferred over Y*). This preference order corresponds to Ellsberg's predictions and matches the data (see, e.g., Becker & Brownson, 1964; Slovic & Tversky, 1974; MacCrimmon & Larsson, 1979). The functions of the four arguments can be understood in an *epistemic* and in a *persuasive* sense. The epistemic function of the arguments is to *gain knowledge* about which bet should be preferred. The persuasive function of the arguments is to *convince* someone which bet should be preferred.

In the following section we further investigate the psychological plausibility of $\mathfrak{s}$ by an experiment.

## Method

### Participants

In this experiment 60 university students (mean age 25.9 years ($SD = 5.6$), 48 females, 12 males) participated for a compensation of 15€. All of the participants were Finnish native speakers and none of them had studied psychology, mathematics, statistics or philosophy as their major.

### Design and Materials

We used three target task types: argument ranking tasks, argument rating tasks, and the (original) Ellsberg tasks. The *argument ranking tasks* first instructed the participants to *rank* the strength of arguments $\mathcal{A}_1$ and $\mathcal{A}_2$ (see Table 1). Second, the participants were instructed to rank the strength of arguments $\mathcal{A}_3$ and $\mathcal{A}_4$. The *argument rating tasks* instructed the participants to *rate* the strength of each of the four arguments. In the original version of the Ellsberg task, participants had to rank which bets they preferred as described in the Introduction. We investigated the following questions which relate argument strength to the Ellsberg problem:

- Do the results of the argument strength rating tasks predict the responses in the Ellsberg tasks?
- Do the results of the argument strength rating tasks predict the responses in the argument strength ranking tasks?

Moreover, we explored empirically, whether argument strength formulated in epistemic or in persuasive terms impacts participants' reasoning. Finally, we systematically manipulated the information conveyed in the argument rating and in the argument ranking tasks by the following independent variables: (i) only the uncertainty of the conclusion was presented, (ii) only the uncertainties of the premises were presented, and (iii) uncertainties of the premises and the conclusion were presented. The instructions introduced the following symbol for marking not conveyed information in the respective conditions which correspond the variables (i) and (ii): 🔇. By using a $2 \times 3$ between-participant design we fully crossed epistemic versus persuasive formulations and the manipulated information conveyed in the arguments. In the epistemic booklets we used knowledge-oriented phrasings like "Which argument is stronger to know which bet to choose?", whereas in the persuasive booklets we used according phrasings like "Which argument convinces stronger which bet to choose?". The experimental conditions are explained in Table 2.

Table 2: Experimental conditions (Cd 1–Cd 6; $N = 60$).

| Presented probabilities | Epistemic | Persuasive |
|---|---|---|
| Premise & conclusion | Cd 1 ($n_1 = 10$) | Cd 2 ($n_2 = 10$) |
| Conclusion only | Cd 3 ($n_3 = 10$) | Cd 4 ($n_4 = 10$) |
| Premise only | Cd 5 ($n_5 = 10$) | Cd 6 ($n_6 = 10$) |

**Argument ranking tasks** In these tasks, the participants were instructed to imagine two friends arguing about which bet the participant should choose. Then, argument $\mathcal{A}_1$ for **Bet 1**, and argument $\mathcal{A}_2$ for **Bet 2** were presented to the participant, e.g.:

**Argument 2** for **Bet 2**

> I am 🔇 % sure that the ball drawn from the urn is red.
> I am 🔇 % sure that the ball drawn from the urn is black or yellow.
> Therefore, I am at least 0 % and at most 67 % sure that the ball drawn from the urn is black.

The participants were then presented with the question "Which argument is stronger to know which bet to choose?" (*Kumpi argumentti on vahvempi sen tietämiseen, kumpi veto kannattaisi valita?*) in the epistemic condition. In the persuasive condition, they were asked *"Which argument convinces you stronger which bet to choose?"* (*Kumpi argumentti vakuuttaa sinut vahvemmin siitä, kumpi veto kannattaisi valita?*). Then, the participants were instructed to indicate

their choice by ticking the respective box for Argument 1 (i.e., $\mathcal{A}_1$) or Argument 2 (i.e., $\mathcal{A}_2$). Finally, the participants ranked Argument 3 (i.e., $\mathcal{A}_3$) and Argument 4 (i.e., $\mathcal{A}_4$).

**Argument rating tasks** In these tasks participants were presented with the same four arguments as in the argument ranking tasks. They were asked to carefully reconsider each. Instead of using forced choice response formats, each argument was followed by a question, e.g., "How strong is **Argument 2** for choosing **Bet 2**?" (*Kuinka vahva **Argumentti 2** on **Vedon 2** valitsemiseksi?*; original epistemic formulation) or "How strong is **Argument 2** for convincing to choose **Bet 2**?" (*Kuinka vahva **Argumentti 2** on vakuuttamaan **Vedon 2** valitsemisesta?*; original persuasive formulation). The participants were asked to mark their responses on a ten point rating scale (see Figure 1).

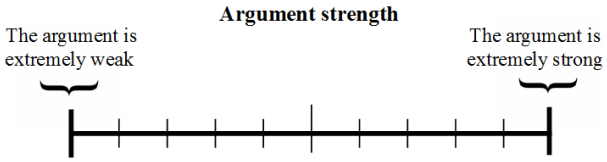


Figure 1: Answer scale used in the argument rating tasks.

**Ellsberg tasks** Here, as explained in the introduction, the participants had to choose which rankings among bets they preferred (**Bet 1** or **Bet 2** and **Bet 3** or **Bet 4**). All participants were presented with the same Ellsberg tasks.

## Procedures

Participants completed the booklets individually in a quiet room. At the beginning of the testing, participants were informed to take as much time as needed for completing the tasks. Furthermore, they were instructed not to look back on their previous responses. After reading the introduction the participants worked on tasks which differed from the Ellsberg problem (and which are not in the scope of the present paper). After that, the target tasks were presented in the following order: (i) argument ranking tasks, (ii) argument rating tasks, and (iii) the Ellsberg tasks. The easier argument ranking tasks (rankings require less cognitive effort than ratings) appeared prior to the argument rating tasks to further help participants to familiarize themselves with the task materials. To avoid any influences of the Ellsberg tasks on the argument strength tasks and to see whether our samples replicate the findings in the literature, the Ellsberg tasks were presented at the end of the booklet. Then, the participants filled in demographic data and rated the difficulty and clearness of the tasks. Participants used 9.6 minutes ($SD = 2.8$) on the average to work on the booklets. Each session concluded by an interview to further explore argument strength from a qualitative point of view: we asked how the participants solved the tasks and what they thought determined the strength of an argument.

Table 3: Percentages of argument preferences in the argument ranking tasks (rnk($\mathcal{A}$)) and in the Ellsberg tasks ($N = 60$).

| % | rnk($\mathcal{A}$) | Ellsberg | % | rnk($\mathcal{A}$) | Ellsberg |
|---|---|---|---|---|---|
| Bet1 | 73,3 | 93,3 | Bet3 | 25,0 | 23,3 |
| Bet2 | 26,7 | 6,7 | Bet4 | 75,0 | 76,7 |

## Results and Discussion

We performed Fisher's exact tests to compare the impact of the different booklets on the response frequencies in the argument ranking tasks and in the Ellsberg tasks. Moreover, we tested influences of the different conditions in the argument rating tasks by analyses of variance. After performing Holm-Bonferroni corrections we did not observe any significant differences. We therefore pooled the data for further analysis ($N = 60$).

**Ellsberg's predictions** The majority of responses in all three types of tasks (i.e., argument ranking, argument rating and Ellsberg task) are consistent with Ellsberg's predictions. Our findings also replicate empirical findings reported in the literature (see, e.g., Becker & Brownson, 1964; Slovic & Tversky, 1974; MacCrimmon & Larsson, 1979). Moreover, our data suggest that classical findings in Ellsberg tasks carry over to (isomorphic) problems formulated in terms of argument strength.

Table 3 shows how the participants ranked the arguments in the argument ranking tasks and how they ranked the bets in the Ellsberg tasks. Bet 1 (resp., argument $\mathcal{A}_1$ supporting Bet 1) is more frequently chosen than Bet 2 (resp., $\mathcal{A}_2$ supporting Bet 2). Likewise, Bet 4 (resp., argument $\mathcal{A}_4$ supporting Bet 4) is more frequently chosen than Bet 3 (resp., $\mathcal{A}_3$ supporting Bet 3).

Moreover, we constructed the underlying preference orders of the argument strengths and the bets from the participants' responses in all the three task types. This allows one to see which choice strategies were most commonly used. In all tasks, strategies consistent with the independence axioms of rational choice were less frequently preferred, as can be seen in Table 4. For constructing the preference orders based on the responses in the argument strength ratings tasks, we made the following assumption: if the strength of an argument $\mathcal{A}_x$ was rated higher than the strength of an argument $\mathcal{A}_y$, then the corresponding Bet $x$ is preferred over Bet $y$. Again, our findings replicate the predictions of Ellsberg and the previous empirical findings (see, e.g., Becker & Brownson, 1964; Slovic & Tversky, 1974; MacCrimmon & Larsson, 1979).

Table 5 shows the mean argument strength rating responses. As predicted by measure $\mathfrak{s}$, the mean argument strength ratings reflect the Ellsberg predictions, i.e.,

Table 4: Percentages of responses consistent with Ellsberg's predictions ($E$), the independence axiom of rational choice ($I$). The preference order $R$ can be interpreted as a reversed version of $E$. "$(x,y) \succ (u,v)$" means "arguments (resp. bets) $x$ and $y$ are preferred over arguments (resp. bets) $u$ and $v$". Preference order responses consistent with $\mathfrak{s}$ are in **bold**.

| Preference Order | Tasks ($N = 60$) | | |
|---|---|---|---|
| | $\mathcal{A}$ Ranking | Ellsberg | $\mathcal{A}$ Rating |
| $(1,4) \succ (2,3)^E$ | **56.67** | **71.67** | **56.10** |
| $(2,3) \succ (1,4)^R$ | 8.33 | 1.67 | 4.88 |
| $(1,3) \succ (2,4)^I$ | 16.67 | 21.67 | 21.95 |
| $(2,4) \succ (1,3)^I$ | 18.33 | 5.00 | 17.07 |

mean rating($\mathcal{A}_1$) > mean rating($\mathcal{A}_2$) and mean rating($\mathcal{A}_4$) > mean rating($\mathcal{A}_3$).

Table 5: Means and standard deviations ($SD$) of the argument strength ratings on a scale from 0 ("extremely weak") to 10 ("extremely strong"; $N = 60$).

| | $\mathcal{A}_1$ | $\mathcal{A}_2$ | $\mathcal{A}_3$ | $\mathcal{A}_4$ |
|---|---|---|---|---|
| Mean | 5,20 | 3,98 | 5,77 | 6,95 |
| $SD$ | 2,64 | 2,58 | 1,74 | 1,87 |

**Consistency among the data** Based on the argument strength ratings, we predicted the participants' choices in the ranking and in the Ellsberg tasks. The data support our predictions: the argument strength rating responses predict the ranking responses in the Ellsberg tasks. The rating responses also predict the responses in the argument strength ranking tasks (see Table 6 and Table 7).

As some participants had rated the arguments for the bets equally strong, no predictions could be derived in these cases. When taking into account only those cases, in which making predictions was possible, the responses of roughly 3/4 of the participants were consistent with their responses in the ranking tasks. In the argument strength ranking tasks, 77.3 % of the participants chose as predicted between the first two bets and 75.0 % chose as predicted between the second two bets. For the Ellsberg tasks, we observed similarly high percentages (i.e., 75.0 % and 70.8 % of the participants, for the first and the second bet rankings, respectively). This is again experimental support for the psychological plausibility of measure $\mathfrak{s}$.

Finally, we discuss qualitative data taken from structured interviews on folk psychological conceptions on what argument strength means.

**Interview results** After the participants completed the paper and pencil tasks, we collected folk psychological conceptions on what "argument strength" (*argumentin vahvuus*) means by structured interviews. We asked the participants how they would define argument strength in their own words. Participants who had received the *persuasive* booklets, we hypothesized, mentioned persuasive aspects (like how *convincing* arguments are) more frequently than those of the epistemic condition. Moreover, participants who had received the *epistemic* booklets focused more on epistemic aspects (like truth and knowledge) than those of the persuasive condition. However, the interview responses do not confirm these hypotheses.

Table 6: Predictions of bet rankings in Ellsberg tasks based on responses in the argument strength rating tasks ($N = 60$).

| % | Ranking | |
|---|---|---|
| | Bet 1 vs. Bet 2 | Bet 3 vs. Bet 4 |
| Chose as predicted | 55.00 | 56.67 |
| Did not choose as predicted | 18.33 | 23.33 |
| No prediction made | 26.67 | 20.00 |

Table 7: Predictions of argument strength rankings based on the responses in argument strength rating tasks ($N = 60$).

| % | Ranking | |
|---|---|---|
| | $\mathcal{A}_1$ vs. $\mathcal{A}_2$ | $\mathcal{A}_3$ vs. $\mathcal{A}_4$ |
| Chose as predicted | 56.67 | 60.00 |
| Did not choose as predicted | 16.67 | 20.00 |
| No prediction made | 26.67 | 20.00 |

The responses to the interview question concerning the meaning of "argument strength" reflected features of our measure $\mathfrak{s}$. Specifically, the *location* of the coherent conclusion probability interval was referred to by almost all of the participants. For many participants the location seemed to be more important than the *precision* of the coherent conclusion probability interval. They had, for example, focused solely on the lower probability bound of the interval and ignored the upper bound or responded based on the mean value of the interval.

However, a few participants also referred to the *precision* of the coherent conclusion probability interval by sentences like:

> "The size of this gap between 33 [%] and 100 [%] is so big that it increases the uncertainty." *(Epävarmuutta lisää se, että väli 33:n ja 100:n välillä on niin suuri)*

Some participants also talked about the truth or correctness of the probability bounds of the conclusion. For them, the arguments were strong, when the probabilities in the conclusions were *correct*, almost regardless of the values in them.

The interview responses provide folk psychological evidence for using location and precision of conclusion probability intervals for evaluating the strength of uncertain argu-

ments. Location and precision are the key ingredients of our argument strength measure $\mathfrak{s}$.

Finally, we note that **Bet 1** is usually not compared directly against **Bet 3** in the traditional Ellsberg task. The corresponding $\mathfrak{s}(\mathcal{A}_1)$ is a bit higher compared to $\mathfrak{s}(\mathcal{A}_3)$: epistemically, this makes sense since the conclusion of $\mathcal{A}_3$ is highly imprecise while the conclusion of $\mathcal{A}_1$ is perfectly precise (see Table 1 above). However, it seems plausible to assume that people would prefer **Bet 3** over **Bet 1**. To accommodate $\mathfrak{s}$ for this prediction, one could reduce the impact of the *precision* by adding suitable weights to the definition of $\mathfrak{s}$.

## Concluding Remarks

Based on the *location* and the *precision* of the conclusion's probability interval, we proposed a formal measure of argument strength $\mathfrak{s}$ and showed how $\mathfrak{s}$ predicts responses in Ellsberg tasks. Specifically, we framed choices among bets in terms of probability logical argument forms. Our data support the hypothesis that Ellsbergs predictions can be justified by argument strength rankings and argument strength ratings.

Since the proposed measure exploits tools available in coherence-based probability logic and since it is based on a deductive consequence relation, it allows for dealing with arguments involving conditionals. The proposed measure has many plausible consequences, which calls for future formal-normative and experimental research for modeling also other argument types, like the conditional syllogisms.

Understanding argument strength is important for theories about reasoning and argumentation in general. Our paper sheds formal and experimental light on what argument strength can mean.

**Acknowledgments** DFG project PF740/2-2 (SPP1516).

## References

Becker, S. W., & Brownson, F. O. (1964). What price ambiguity? or the role of ambiguity in decision-making. *Journal of Political Economy*, *72*(1), 62–73.

Briggs, R. (2016). Normative theories of rational choice: Expected utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). `http://tinyurl.com/hjwmajw`.

Coletti, G., & Scozzafava, R. (2002). *Probabilistic logic in a coherent setting*. Dordrecht: Kluwer.

Crupi, V., Tentori, K., & Gonzales, M. (2007). On Bayesian measures of confirmation. *Philosophy of Science*, *74*, 229-252.

de Finetti, B. (1970/1974). *Theory of probability* (Vols. 1, 2). Chichester: John Wiley & Sons.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford: OUP.

Gilio, A., Over, D. E., Pfeifer, N., & Sanfilippo, G. (2017). Centering and compound conditionals under coherence. In M. B. Ferraro et al. (Eds.), *Soft methods for data science* (pp. 253–260). Berlin, Heidelberg: Springer.

Gilio, A., Pfeifer, N., & Sanfilippo, G. (2016). Transitivity in coherence-based probability logic. *Journal of Applied Logic*, *14*, 46–64.

Haenni, R. (2009). Probabilistic argumentation. *Journal of Applied Logic*, 155–176.

Hahn, U., & Oaksford, M. (2006). A normative theory of argument strength. *Informal Logic*, *26*, 1-22.

MacCrimmon, K. R., & Larsson, S. (1979). Utility theory: Axioms versus 'paradoxes'. In M. Allais & O. Hagen (Eds.), *Expected utility and the Allais paradox* (Vol. 1979, pp. 333–409). Dordrecht: Reidel.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Oaksford, M., & Hahn, U. (2007). Induction, deduction, and argument strength in human reasoning and argumentation. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 269–301). Cambridge: Cambridge University Press.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200.

Over, D. E., & Cruz, N. (in press). Probabilistic accounts of conditional reasoning. In L. Macchi, M. Bagassi, & R. Vialem (Eds.), *International handbook of thinking and reasoning.* Hove Sussex: Psychology Press.

Pfeifer, N. (2007). Rational argumentation under uncertainty. In G. Kreuzbauer, N. Gratzl, & E. Hiebl (Eds.), *Persuasion und Wissenschaft* (pp. 181–191). Wien: LIT Verlag.

Pfeifer, N. (2013a). The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning*, *19*(3–4), 329–345.

Pfeifer, N. (2013b). On argument strength. In F. Zenker (Ed.), *Bayesian argumentation. The practical side of probability* (pp. 185–193). Dordrecht: Synthese Library (Springer).

Pfeifer, N. (2014). Reasoning about uncertain conditionals. *Studia Logica*, *102*(4), 849-866.

Pfeifer, N., & Kleiter, G. D. (2006). Inference in conditional probability logic. *Kybernetika*, *42*, 391-404.

Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*(2), 206–217.

Prakken, H., & Vreeswijk, G. (2002). Logic for defeasible argumentation. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (2nd ed., Vol. 4, pp. 219–318). Dordrecht: Kluwer.

Sanfilippo, G., Pfeifer, N., & Gilio, A. (2017). *A generalized probabilistic version of modus ponens.* `https://arxiv.org/abs/1705.00385`.

Slovic, P., & Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science*, *19*(6), 368–373.

Zenker, F. (Ed.). (2013). *Bayesian argumentation: The practical side of probability*. Dordrecht: Synthese Library (Springer).