A Utility Based Evaluation of Logico-Probabilistic Systems

Paul D. Thorn & Gerhard Schurz

September 30, 2013

Abstract. Systems of logico-probabilistic (LP) reasoning characterize inference from conditional assertions interpreted as expressing high conditional probabilities. In the present article, we investigate four prominent LP systems (namely, systems **O**, **P**, **Z**, and **QC)** by means of computer simulations. The results reported here extend our previous work in this area, and evaluate the four systems in terms of the *expected utility* of the dispositions to act that derive from the conclusions that the systems license. In addition to conforming to the dominant paradigm for assessing the rationality of actions and decisions, our present evaluation complements our previous work, since our previous evaluation may have been too severe in its assessment of inferences to false and uninformative conclusions. In the end, our new results provide additional support for the conclusion that (of the four systems considered) inference by system **Z** offers the best balance of error avoidance and inferential power. Our new results also suggest that improved performance could be achieved by a modest strengthening of system **Z**.

Keywords: Probability logic, Ampliative inference, Scoring rules

1. Introduction

Systems of *logico-probabilistic* (LP) reasoning characterize inference from conditional assertions that are interpreted as expressing high conditional probabilities. In previous work [17], we studied four LP systems (namely, systems **O**, **P**, **Z**, and **QC**; described below), and presented data from computer simulations that illustrated the performance of the four systems. These simulations tested the four systems in terms of their tendency to draw true informative conclusions and avoid drawing false or uninformative conclusions, given accurate information about an environment, where the inferred conclusions take the form of lower probability bounds. In [17], we argued that our simulations support the conclusion that system **Z** provides the best balance of *reward* versus *risk*, in terms of drawing true informative conclusions and avoiding false or uninformative conclusions. A key tool of our evaluation was the *scoring* of individual conclusions licensed by each of the four systems, and one of our essential points was that most of the inferences that are made by system **QC** (the strongest system considered), and not by system **Z** (the second strongest system considered) yield negative scores.

The methods of evaluation employed in [17] were based on certain intuitions which led to three different scoring measures (called ACG, PIR, and sPIR; described in section 3). Of course, intuitions are never sacrosanct, and may be overturned by other considerations, and in retrospect, it appears that reasons can be given for thinking that the scoring rules employed in [17] were biased toward over-punishing inference to false and uninformative conclusions, and may thus have been biased against system **QC** which licenses many inferences, and thereby take greater risks for the sake of drawing true informative conclusions.

In this article, we present data that represents a re-evaluation of systems **O**, **P**, **Z**, and **QC**, according to a new scoring measure, known as "EU", which equates the value of a drawing a conclusion with the expected utility of the actions licensed by that conclusion, and is thereby in line with the dominant paradigm for evaluating the rationality of actions and decisions. We observe that the resulting EU scores earned by system **QC** are significantly better than the scores that it earned by the lights of the ACG and sPIR measures. However, we also note that the mean EU scores for the inferences drawn by system **QC** and not by system **Z** are very low and often negative. Based on these findings, we draw two important conclusions: (1) any LP reasoning system whose EU scores exceed those of system **Z** will presumably result from strengthenings of that system, rather than from weakenings, and (2) such strengthenings will presumably be small, compared to the full strength of system **QC**. We begin our discussion with a brief description of the four LP systems.

2. LP Reasoning: Systems O, P, Z, and QC

We represent the four LP systems using a simple propositional language L, with the usual connectives $\neg$, $\wedge$, $\vee$, and $\supset$ and $\equiv$ (for material implication and material bi-implication), and A, B, C, etc. as meta-logical variables ranging over arbitrary sentences of L. Throughout the article, our interest will be restricted to the extension of L which results from the addition of *simple* uncertain conditionals of the form $\mathrm{A}\Rightarrow\mathrm{B}$. α and β will serve as meta-variables ranging over such simple conditional formulas, while Γ ranges over sets of them. "$\vdash$" is used to denote derivability in classical logic, and "$\perp$" to denote an arbitrary contradiction. The four LP systems that we consider are ordered in terms of the number of inferences they license (**O** ⊂ **P** ⊂ **Z** ⊂ **QC**). We proceed by considering the weakest system first.

2.1. System O

System **O** is of interest because of its close connection to the consequence relation $\Vdash_{\mathbf{s.p.}}$, where "**s.p.**" stands for "strict preservation":

(1) *Strict Preservation:* $A_1{\Rightarrow}B_1,\ldots, A_n{\Rightarrow}B_n \Vdash_{\mathbf{s.p.}} C{\Rightarrow}D$ *iff* for all probability functions P (over L): $P(D|C) \geq \min(\{P(B_i|A_i) : 1{\leq}i{\leq}n\})$.[1]

System **O** was developed by Hawthorne [7] and Hawthorne and Makinson [8] as an inferential calculus for $\Vdash_{\mathbf{s.p.}}$. Throughout the present article, "$\vdash_{\mathbf{O}}$" denotes the syntactical notion of derivability in system **O**.

System O (after Hawthorne):
REF (reflexivity): $\vdash_{\mathbf{O}} A{\Rightarrow}A$
LLE (left logical equivalence): if $\vdash A{\equiv}B$, then $A{\Rightarrow}C \vdash_{\mathbf{O}} B{\Rightarrow}C$
RW (right weakening): if $\vdash B{\supset}C$, then $A{\Rightarrow}B \vdash_{\mathbf{O}} A{\Rightarrow}C$
VCM (very cautious monotony): $A{\Rightarrow}B{\wedge}C \vdash_{\mathbf{O}} A{\wedge}B{\Rightarrow}C$
XOR (exclusive Or): if $\vdash \neg(A{\wedge}B)$, then $A{\Rightarrow}C, B{\Rightarrow}C \vdash_{\mathbf{O}} A{\vee}B{\Rightarrow}C$
WAND (weak And): $A{\Rightarrow}B, A{\wedge}\neg C{\Rightarrow}\bot \vdash_{\mathbf{O}} A{\Rightarrow}B{\wedge}C$

It is easy to see that all of the rules of system **O** are correct with respect to $\Vdash_{\mathbf{s.p.}}$, i.e., $\Gamma \vdash_{\mathbf{O}} A{\Rightarrow}B$ implies $\Gamma \Vdash_{\mathbf{s.p.}} A{\Rightarrow}B$. It was the hope of Hawthorne and Makinson [7] that $\vdash_{\mathbf{O}}$ was also complete with respect to $\Vdash_{\mathbf{s.p.}}$, but, as Paris and Simmonds [13] have shown, this is not the case.

Following [17], we propose a marriage of system **O**, and a rule for inferring lower probability bounds that corresponds to the correctness of system **O** for $\Vdash_{\mathbf{s.p.}}$. To make sense of such inferences, we employ statements of the form "$A{\Rightarrow}_rB$" to express that $P(B|A) \geq r$, and say that system **O** licenses the (valid) inference to $C{\Rightarrow}_{min(\{r_i:1\leq i\leq n\})}D$ from $A_1{\Rightarrow}_{r_1}B_1,\ldots, A_n{\Rightarrow}_{r_n}B_n$, in cases where $A_1{\Rightarrow}B_1,\ldots, A_n{\Rightarrow}B_n \vdash_{\mathbf{O}} C{\Rightarrow}D$.

A remarkable fact about system **O** is its weakness compared to standard systems of conditional logic. According to Segerberg [18], the weakest 'reasonable' system of conditional logic includes REF, LLE, RW, along with the following rule:

(AND): from $A{\Rightarrow}B$ and $A{\Rightarrow}C$ infer $A{\Rightarrow}B{\wedge}C$.[2]

[1] As a convenience, we assume that $P(B|A) = 1$, when $P(A) = 0$.

[2] For an investigation of lattices of systems based on Segerberg's [18] minimal semantics, see [21] and [22].

The inferential power of AND is quite significant. By adding AND to the system **O**, we obtain (in one step) the famous system **P**.[3]

2.2. System P

As we describe below, system **P** represents the confluence of a number of different semantic criteria. However, the feature of system **P** that is of greatest interest here is its connection with the following consequence relation:

(2) *Improbability-Sum Preservation:* $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \Vdash_{\mathbf{i.s.p.}} C \Rightarrow D$ *iff* for all probability functions over L: $I(D|C) \leq \Sigma\{I(B_i|A_i) : 1 \leq i \leq n\}$, where $I(A|B)$ is defined as $1 - P(A|B)$ (cf. [2], [1]).

Adams demonstrated that the following calculus **P** (denoted by $\vdash_{\mathbf{P}}$) is correct and complete for $\Vdash_{\mathbf{i.s.p.}}$.

System P (after Adams):
REF, LLE, and RW (as with system **O**)
AND (as above)
CM (cautious monotony): $A \Rightarrow B, A \Rightarrow C \vdash_{\mathbf{P}} A \wedge B \Rightarrow C$
OR: $A \Rightarrow C, B \Rightarrow C \vdash_{\mathbf{P}} A \vee B \Rightarrow C$

In addition to being correct and complete with respect to $\Vdash_{\mathbf{i.s.p.}}$, system **P** is correct and complete for three other semantics, namely: (i) infinitesimally high probability semantics, (ii) *ranked* (or preferentially ordered) *possible world semantics*[4], and (iii) Adams' *yielding* condition (cf. [2], [1], [17]).

Following [17], we propose a marriage of system **P**, and a rule for inferring lower probability bounds that corresponds to the correctness of system **P** for $\Vdash_{\mathbf{i.s.p.}}$. In particular, we say that system **P** licenses the (valid) inference to $C \Rightarrow_{1-\Sigma\{1-r_i : 1 \leq i \leq n\}} D$ from $A_1 \Rightarrow_{r_1} B_1, \ldots, A_n \Rightarrow_{r_n} B_n$, in cases where $A_1 \Rightarrow B_1, \ldots, A_n \Rightarrow B_n \vdash_{\mathbf{P}} C \Rightarrow D$.

2.3. System Z

While system **P** sanctions inference to a larger number of conditionals than system **O**, it still sanctions fewer inferences than one might reasonably accept. For instance, system **P** does not licence inference via *subclass inheritance* based on default assumptions of *irrelevance* (or *independence*). For example, if we know that this animal is a male bird ($B \wedge M$) and that birds can

[3] There are also systems which lie in between **O** and **P**, namely the systems **C** and **CL** (cf. [10], [12], ch. III), for which no known probability semantics exists.

[4] The connection between such ranked world semantics and Spohn's ranking theory of belief is one that has yet to be explored in a systematic way (cf. [19], 15).

normally fly (B⇒F), and nothing else of relevance, then we would intuitively draw the conclusion that this male bird can fly (F). However, B∧M⇒F is not **P**-entailed by B⇒F, because there exist possible probability distributions in which P(F|B∧M) is much smaller than P(F|B). If we do infer B∧M⇒F from B⇒F, in such cases, then we assume, by default, that the additional factor M (in this case the gender of a bird) is *irrelevant* to its ability to fly (or in other words, M and F are assumed to be probabilistically independent given B). A straightforward means of enlarging the set of LP-derivable conditionals, in order to include such default inferences, is to give up the requirement that a reasonable inference be valid for *all possible* probability distributions, and consider only 'normal' probability distributions – in particular those distributions which satisfy the default assumption of irrelevance. An early suggestion for realizing this idea was the *maximum entropy* approach to default inference (see [14], 491-3). By selecting a probability distribution that maximizes entropy, one minimizes probabilistic dependences. Despite having some attractive features, the maximum entropy approach is rather complicated and has some further disadvantages, such as language dependence (cf. [6], 309f; [23]).

System **Z** of Pearl [15] and Goldszmidt and Pearl [5] maintains many of the advantages of the maximum entropy approach, while overcoming its disadvantages.[5] Like the maximum entropy approach, inference in system **Z** proceeds via the construction of a semantic model of the premise conditionals that maximizes probabilistic independences. In system **Z**, this is achieved by maximizing the *degree-of-normality* of the set of possible worlds represented by a ranked model, according to the following definition:

(3) *Definition* (cf. [5], 68, Def. 15; [6], 308f): A ranked model (W, r) is *at least as normal* as a ranked model (W, r*) (with the same world set), in short (W, r) $\geq_n$ (W, r*), *iff* for all w∈W, r(w) ≤ r*(w).[6]

According to theorem 1(i) (below) there exists, for every set of worlds W (which is constructed over the language of the conditional knowledge base Γ), a unique *most normal* ranked model, the so called *z-model.* In order to define the notion of a z-model, we first define the notion of a *z-rank.* A

[5] The approach of [15] and [5] was independently suggested by Lehmann and Magidor [11], who called the system *rational closure.* The approach of Lehmann and Magidor proceeds by constructing a preference relation over the set of all extensions of the P-closure of a given set of premise conditionals Γ which satisfy rational monotony (RM): A⇒B / A∧C⇒B or A⇒¬C. It was then shown that the Z-closure of Γ is the uniquely preferred extension in this set (see [11], p. 29, 31-33, 38-42).

[6] A conditional A⇒B is said to be satisfied by a ranked world model (W, r), if all of the lowest-rank worlds verifying A verify B.

precise statement of this definition requires some additional terminology: A set of conditionals Γ is *P-consistent iff* Γ does not **P**-entail ⊥. A conditional A⇒B is *tolerated* by a set of conditionals Γ *iff* {A∧B}∪{C⊃B: C⇒B ∈ Γ} is consistent.

(4) *Definition* (cf. [15], section 1; [5], 65, fig. 2):
For every (finite) **P**-consistent set of conditionals Γ, the *z-rank* of the elements of Γ is defined by the following *z-algorithm*:
(1) *Initial step:* Set $i = 0$. Set $\Delta = \Gamma$.
(2) *Iterative step:* While Δ is nonempty, let $\Delta_i \subseteq \Delta$ consist of all conditionals $\alpha \in \Delta$ which are tolerated by Δ.
(i) If Δ_i is nonempty, let $\Delta = \Delta - \Delta_i$, and $i = i+1$.
(ii) If Δ_i is empty, let $\Delta_\infty = \Delta$, and $\Delta = \emptyset$.
Output: The *z-partition* $(\Delta_0, \ldots, \Delta_k, \Delta_\infty)$.
The *z-rank of a conditional* α in a **P**-consistent Γ, written "$z_\Gamma(\alpha)$", is defined as the index i of that set Δ_i in the z-partition of Γ in which α occurs.

The preceding definition differs from Pearl and Goldszmidt's corresponding definition in the additional clause (2)(ii), which allows that Γ is ϵ-inconsistent[7] (for details see [17], Def. (9)). The assumption that Γ is **P**-consistent guarantees that there is a z-model for Γ, according to the following definition (where a conditional A⇒B is *verified* by a possible world w *iff* A∧B is true at w, and *falsified* by w *iff* A∧¬B is true at w, and a set of conditionals, Γ, is *falsified* by w *iff* at least one conditional in Γ is falsified by w):

(5) *Definition* (cf. [15], 123-5, Eq. 5, 6, and 10): The *z-model* of a **P**-consistent Γ, (W_Γ, z_Γ), is defined as follows: For each w among the set of logically possible worlds over the language of Γ:
(i) If w falsifies Δ_∞, then $w \notin W_\Gamma$.
(ii) If w does not falsify Δ_∞, then (a) $w \in W_\Gamma$, and (b) $z_\Gamma(w) = 0$, if w does not falsify any $\alpha \in \Gamma$, and $z_\Gamma(w) = \max(\{z_\Gamma(\alpha)\text{: w falsifies } \alpha\})+1$, otherwise.
(iii) The z-rank of an arbitrary formula C relative to (W_Γ, z_Γ) is defined as $z_\Gamma(C) = \min(\{\, z_\Gamma(w) : w \in W_\Gamma \;\&\; \text{w verifies C}\})$, with $\min(\emptyset) := \infty$.
(iv) For all Γ: $\Gamma \vdash_{\mathbf{Z}}$ C⇒D *iff* either (a) Γ is **P**-inconsistent, or (b) C⇒D is satisfied in (W_Γ, z_Γ).

Theorem 1 expresses the crucial property of the z-model:

Theorem 1 (cf. [15], [5]):

[7] A set of conditionals is ϵ-consistent just in case the corresponding conditional probabilities can be simultaneously made arbitrarily close to 1.

(i) For every **P**-consistent Γ there exists a unique most normal ranked model among all ranked models for Γ, and this is the z-model (W_Γ, z_Γ).
(ii) $\Gamma \vdash_{\mathbf{Z}}$ C⇒D (according to (5)(iv)) *iff*
$\{A{\supset}B\text{: } A{\Rightarrow}B \in \Gamma \ \&\ z_\Gamma(A{\Rightarrow}B) \geq z_\Gamma(C)\} \vdash C{\supset}D$.

Proof: For (i), see [5] (67f), since the admission of an ϵ-inconsistent remainder set Δ_∞ doesn't change the uniqueness of the z-model. For (ii), see [17].

Z-entailment validates inference by *default inheritance* (i.e., A⇒B $\vdash_{\mathbf{Z}}$ A∧C⇒B) as well as *default contraposition* (i.e., A⇒B $\vdash_{\mathbf{Z}}$ ¬B⇒¬A). That these inferences hold 'by default' means that they hold under the condition that the conditional knowledge base doesn't contain further conditionals that are ϵ-inconsistent with the conclusions of these inferences. Hence, in contrast to $\vdash_{\mathbf{O}}$ and $\vdash_{\mathbf{P}}$, $\vdash_{\mathbf{Z}}$ is *non-monotonic* over conditionals. For example, A⇒B $\vdash_{\mathbf{Z}}$ A∧C⇒B, but A⇒B, A∧C⇒¬B $\nvdash_{\mathbf{Z}}$ A∧C⇒B. Similarly, ¬B⇒¬A is not **Z**-entailed by A⇒B in the presence of ¬B⇒A.

There is no separate calculus for **Z**-entailment, but Theorem 1(ii) and the z-algorithm, (4), outline a straightforward procedure for deciding **Z**-entailment via propositional satisfiability tests. One significant disadvantage of **Z**-entailment is that (in the absence of further assumptions) it doesn't give us information about probabilistic reliability in the form of almost-tight lower bounds, such as is provided by the improbability sum semantics for system **P**. Nor does system **Z**, on its own, tell us the *minimal* probabilistic default assumptions that are needed to derive particular conclusions. However, it is shown in [17] (based on work in [16]) how to obtain these additional desiderata.

Theorem 2: If $A_1{\Rightarrow}B_1,\ldots, A_n{\Rightarrow}B_n \vdash_{\mathbf{Z}} C{\Rightarrow}D$ holds, then improbability-sum preservation ($I(D|C) \leq \Sigma\{I(B_i|A_i) : 1{\leq}i{\leq}n\}$) holds for all probability functions P that satisfy the default assumptions $P(A_i{\supset}B_i|C) \geq P(B_i|A_i)$, for all $1{\leq}i{\leq}n$.

Proof: See [17], Theorem 2.6.

We proceed here as if the default assumptions specified in Theorem 2 hold, and say that system **Z** licenses the inference to $C{\Rightarrow}_{1-\Sigma\{1-r_i:1\leq i\leq n\}}D$ from $A_1{\Rightarrow}_{r_1}B_1,\ldots, A_n{\Rightarrow}_{r_n}B_n$, in cases where $A_1{\Rightarrow}B_1,\ldots, A_n{\Rightarrow}B_n{\vdash_{\mathbf{Z}}} C{\Rightarrow}D$. As with the evaluations conducted in [17], we concern ourselves with question of whether inference by the proposed rule tends to yield accurate conclusions.

2.4. System QC

Z-entailment is not the strongest (minimally reasonable) inference calculus for 'risky' default inference among uncertain conditionals. An even stronger and extremely simple calculus is *quasi-classical* reasoning. Here one reasons with uncertain conditionals as if they were material implications:

(6) $\Gamma \vdash_{\mathbf{QC}}$ C⇒D *iff* {A⊃B : A⇒B ∈ Γ} ⊢ C⊃D.

Improbability-sum preservation holds for inferences between material conditionals, or more generally, between formulas of propositional logic, as was shown by Suppes [20]. In particular, $\{A_1,\ldots,A_n\} \vdash B$ *iff* it holds for all probability distributions that $I(B) \leq \Sigma\{I(A_i){:}1{\leq}i{\leq}n\}$. Beyond the result of Suppes, it is possible to formulate probabilistic conditions under which **QC**-reasoning approximately satisfies improbability-sum preservation. In particular, it is shown in [17] (Sec. 2.5) that a **QC** inference from a given set of premises is guaranteed to preserve probability in the manner of system **P** *iff* the improbability-sum of the premises is very small, and some decimal powers smaller than the probability of the conclusion's antecedent. Following [17], we proceed as if these conditions hold, and say that system **QC** licenses the inference to $C{\Rightarrow}_{1-\Sigma\{1-r_i:1\leq i\leq n\}}D$ from $A_1{\Rightarrow}_{r_1}B_1,\ldots,$ $A_n{\Rightarrow}_{r_n}B_n$, in cases where $A_1{\Rightarrow}B_1,\ldots,$ $A_n{\Rightarrow}B_n \vdash_{\mathbf{QC}} C{\Rightarrow}D$. The question remains of whether inference by the preceding rule tends to yield accurate conclusions.

3. How to Evaluate Judged Lower Probability Bounds

In line with [17], our interest is in determining which LP system offers the best balance of reward versus risk, and we continue the approach of scoring the individual conclusions licensed by the respective systems. In [17], three scoring measures were considered. The first scoring measure introduced was called the *advantage-compared-to-guessing* measure:

(7) The *advantage-compared-to-guessing* (ACG) score for derived conditionals: $\mathrm{Score}_{ACG}(C{\Rightarrow}_r D,\, P) := 1/3 - |r - P(D|C)|$.

The idea behind this measure derives from the fact that the mean difference between two random choices of two real values r and s from the unit interval is (provably) 1/3. Based on this fact, we assessed each system by counting a judged lower probability bound that differs from the true (actual) probability by more than one-third *negatively*, and by counting a judged lower probability bound that differs from the true probability by less than one-third *positively.* The ACG measure scores judged lower probability bounds by a simple *linear* measure of their distance from the true probabilities.

Within [17], we acknowledged that the ACG measure does not provide a fully adequate means of evaluating LP systems, since it sometimes punishes a true informative judgment, in comparison with making no judgment at all. This will occur, for example, if the conclusion $\mathrm{C}\Rightarrow_{0.6}\mathrm{D}$ is inferred, when $\mathrm{P}(\mathrm{D}|\mathrm{C}) = 0.95$. In order to take a broader view of the advantages and disadvantages of reasoning in accordance with the four systems, we considered two additional scoring rules:

(8) The *price-is-right* (PIR) score for derived conditionals:
$\mathrm{Score}_{PIR}(\mathrm{C}\Rightarrow_r\mathrm{D}, \mathrm{P}) := r$, if $r \leq \mathrm{P}(\mathrm{D}|\mathrm{C})$,
$:= 0$, otherwise.

(9) The *subtle-price-is-right* (sPIR) score for derived conditionals:
$\mathrm{Score}_{sPIR}(\mathrm{C}\Rightarrow_r\mathrm{D}, \mathrm{P}) := r$, if $r \leq \mathrm{P}(\mathrm{D}|\mathrm{C})$,
$:= \mathrm{P}(\mathrm{D}|\mathrm{C}) - r$, otherwise.

The PIR measure rewards true inferred lower bounds, but it does not punish false inferred lower bounds (i.e., cases where $r > \mathrm{P}(\mathrm{D}|\mathrm{C})$). We acknowledged this defect of the PIR score, and introduced the sPIR scoring measure in order to address the defect: The sPIR measure rewards true inferred lower bounds (proportional to their closeness to the true probability), and punishes false inferred lower bounds (proportional to their distance from the true probability). A possible problem with the sPIR measure is that sometimes punishes inference to highly informative, though incorrect, lower probability bounds. This will occur, for example, if the conclusion $\mathrm{C}\Rightarrow_{0.96}\mathrm{D}$ is inferred, when $\mathrm{P}(\mathrm{D}|\mathrm{C}) = 0.95$.

Our evaluations in [17] were based primarily on the ACG and sPIR scoring measures. But in retrospect, reasons can be given for thinking that the two measures are biased toward over-punishing inference to false and uninformative conclusions, and thus biased against system **QC**. Moreover, although the ACG and sPIR measures are based on reasonable intuitions, it is not precisely clear how these measures correlate with the expected utility of decision making based on the conclusions inferred by the respective LP systems. In order to address this concern, we here develop a scoring rule that measures the expected utility of decision making based on the inferences licensed by respective LP systems. As it turns out, the resulting measure is also relatively permissive in its evaluation of inference to false and uninformative conclusions. So in contrast to the ACG and sPIR measures, evaluation via the new measure provides test conditions that are very favorable to system **QC**.

The behavioral import of a probability judgment is naturally explicated in terms of the actions that the judgment licenses for agents who are prudent. A relatively simple explication of the actions licensed by respective probability estimates is cashed out in terms of betting behavior, in particular, in terms of the price that an agent should be willing to pay for a wager on a proposition B, which is assumed to pay \$1 if B is true, and \$0 otherwise. According to this explication, an agent who assigns probability r to B: (i) should be willing to pay \$$s$ for a wager on B so long as s $< r$, (ii) should be indifferent to paying \$$r$ for such a wager if $r = s$, and (iii) should be unwilling to pay \$$s$ for such a wager, if $s > r$. (\$$s$ is called the "stake" of the bet.) This framework is easily extended in order to explicate judgments about conditional probabilities in terms of conditional wagers, by assuming that a conditional wager is called off, and the stake returned, if the antecedent proposition is false.

We can also extend the preceding account in order to explicate prudent betting behavior based on judged greatest lower probability bounds, and judged greatest lower conditional probability bounds. For the sake of simplicity, we assume that a prudent agent will reject any wager (or conditional wager) about which she is indifferent.[8] Then we have the following explication of the behavioral import of judged greatest lower conditional probability bounds: If r is the greatest lower probability bound that a given agent accepts for B given A, then (if she is prudent and has sufficient capital) she will purchase all wagers on B, conditional on A, at price \$$s$, so long as $s < r$, and refuse to accept such wagers for $s \geq r$.[9] Given these conditions, we can compute the *expected utility* of accepting a greatest lower probability bound r on P(B|A) in the case where a respective agent is offered only a single opportunity to wager on B conditional on A, at price \$$s$.[10]

Assuming that an agent accepts and rejects wagers in the manner described, we can also determine the expected utility of accepting a particular greatest lower probability bound on B conditional on A, independent from a specific stake s, by assuming an environment in which the agent is offered a single opportunity to purchase a wager on B conditional on A with a stake s, where s is determined at random, according to a uniform probability dis-

[8] This assumption makes no difference to the scores agents receive by the EU scoring measure (below), since the model upon which this measure is based assigns zero probability to an agent being offered a wager at a price that is identical to her greatest lower probability bound for the proposition in question.

[9] Since an agent is offered only a single opportunity to wager within the model introduced in the following paragraph, the kind of case contemplated in [4] cannot arise.

[10] We here adopt the assumption that an agent's utility function for wealth is linear, and thus that the expected utility of judged probabilities, within the model, is identical to the expected changes in wealth consequent to those judged probabilities. Alternatively, the present account can be recast so that the currency of the wagers considered is measured in units of utility.

tribution over the interval [0, 1].[11] Given this assumption, we arrive at the following scoring rule, which measures the expected utility of accepting the greatest lower bound r for P(B|A):

(11) The *expected utility* (EU) score for derived conditionals:
$\text{Score}_{EU}(\text{A}{\Rightarrow_r}\text{B, P}) := (\text{P(B|A)}^2 - (\text{P(B|A)} - r)^2)\cdot\text{P(A)}/2.$

The following features of the EU measure are of interest:
(i) The EU measure punishes uninformativeness, but to a lesser degree than the ACG measure, because the absolute difference between P(B|A) and r is squared, which makes it smaller (assuming $|\text{P(B|A)} - r|$ is less than one).
(ii) The EU measure rewards high inferred lower bounds to a higher degree than low inferred lower bounds of comparable accuracy, because the expected gain of the former ones is much greater than that of the latter ones.
(iii) The EU score earned for an inference is discounted as a function of the probability of the antecedent condition, which is reasonable, since the more likely it is that a wager is called off, the less the impact of purchasing the wager, in terms of possible gains and losses.
(iv) The EU measure does not punish falsely inferred lower bounds, because deviating upwards and downwards from the true probability P(B|A) is punished to the same degree $(\text{P(B|A)} - r)^2\cdot\text{P(A)}/2$. The EU measure shares this feature with the ACG measure.

In the following section we review the simulation results of [17], and present the results of new simulations which evaluate the four LP reasoning systems by means of the EU scoring rule.

4. The Simulations

Our simulations were identical to the ones described in [17] save that we scored each of the systems according to the EU measure, in addition to replicating the results for the three other measures.[12] Following [17], we assumed a simple language with four two-valued variables: a, b, c, and d. We likewise assumed a probability distribution over the sixteen possible worlds: ±a∧±b∧±c∧±d (where “±” connotes a negated or unnegated variable). For all of our simulations, we generated a probability distribution

[11] Alternatively, one may consider the case where the agent is offered repeated opportunities to wager on A conditional on B (where the stake s varies according to a uniform distribution on [0, 1]), and consider the average amount earned by an agent who is disposed to accept and reject wagers in the described manner.

[12] The simulations in [17] were programmed in Visual Basic .NET 2010. We adapted that code in order to run the simulations described here.

over these worlds by setting the values of the following fifteen independently variable probabilities: P(a), P(b|a), P(b|a), P(c|a∧b), P(c|a∧b), P(c|a∧b), P(c|a∧b), P(d|a∧b∧c), P(d|a∧b∧c), P(d|a∧b∧c), P(d|a∧b∧c), P(d|a∧b∧c), P(d|a∧b∧c), P(d|a∧b∧c), and P(d|a∧b∧c). In all cases, the values for these probabilities were determined at random, setting each of the fifteen values independently, according to a uniform probability distribution on the unit interval.[13]

For all simulations, we assumed that a small number of conditionals, so-called *base conditionals*, together with their associated probabilities, were known to the reasoning systems. We then allowed each LP reasoning system to infer, from the base conditionals, so-called *derived conditionals* C$\Rightarrow_r$D, which follow according to the respective system. For systems **P**, **Z**, and **QC**, the value r, for each derived conditional, was set to be one minus the sum of the improbabilities of the base conditionals needed in deriving the conclusion. For system **O**, r was set to be the probability value of the least probable base conditional needed for the derivation of C⇒D in **O**.

To manage the search space in assessing the four LP systems, we restricted our attention to conditionals whose antecedent and consequent consist in conjunctions of literals, i.e., of formulas of the form $\pm x$ (for $x \in \{$a, b, c, d$\}$). We also assumed that no propositional atom appears twice in a potential base or derived conditional. These restrictions effectively limited the language under consideration to 464 conditionals (cf. [17]). We call the language composed of this set of 464 conditionals L_4.

For the systems **P**, **Z**, and **QC**, our program tested whether a given conditional follows from a given set of base conditionals via a series of propositional satisfiability tests, by an implementation of resolution/refutation theorem proving. In the case of system **O**, where it is impossible to test for implications via propositional satsifiability checks, our algorithm was different, and exploited the restrictions that were imposed on the set of potential base and derived conditionals.

5. Results

Given a probability distribution P, our program selected a small set of base conditionals at random from among the set of those conditionals whose conditional probability met or exceeded a fixed value s. We call the value s

[13] We used the RandomClass constructor that is built into the .NET Framework in order to generate these values. The constructor generates pseudo random numbers according to algorithm based on Donald E. Knuth's subtractive random number generator algorithm [9], with a time-dependent seed value which is determined by the system clock.

the *minimum probability of base conditionals.* We varied the number of base conditionals b ($b = 1$, 2, 3, or 4), along with the value s ($s = 0.5$, 0.6, 0.7, 0.8, 0.9, or 0.9999). For each combination of values for b and s, we ran one thousand simulations (so for each combination we generated a probability function and a set of base conditionals one thousand times).[14] For each simulation, our program tabulated the *total* of the ACG, PIR, sPIR, and EU scores accumulated by each system for the set of conditionals (and associated bounds) derived by that system.

Our results for the ACG, PIR, and sPIR scoring measures were very similar to the results reported in [17]. A representative selection of the results are presented in Figures 1 and 2, which chart the average total AvG and sPIR scores *per simulation* (including standard error bars), for varied values of s (the minimum probability for base conditionals), where the number of base conditionals is held fixed at four.[15]

The poor performance of system **QC**, as measured by the ACG and sPIR scoring rules, is best grasped by noting that the set of **QC** inferences includes the set of system **Z** inferences. It follows that the aggregate score earned for the inferences among **QC**−**Z** is equal to the result of subtracting the aggregate score earned by system **Z** from the aggregate score earned by system **QC**. As with the simulations described in [17], the average aggregate score earned for the inferences among **QC**−**Z** was negative for both the ACG and sPIR measures, for each combination of b (number of base conditionals) and s (minimum probability of base conditionals) that we considered. That systems **QC**'s sPIR scores are so much worse than that of systems **O** and **P** stems from the fact that most inferences in **QC**−**Z** are erroneous (and sPIR strongly punishes mistakes). Similarly, **QC**'s ACG scores are much worse than those of systems **O** and **P** for minimum probabilities greater than 0.75, which means that **QC**−**Z** inferences based on premises of moderate probability are very uninformative.

The results of [17] strongly suggest that inferences made by system **Z** are of good quality, and that the additional inferences made by system **QC** (the

[14] Though the standard error rates for the mean values reported below vary greatly, we judged (based on trial simulations) that 1000 simulations would be adequate to achieve sufficiently accurate results (balancing concerns about computational feasibility). Our judgment is vindicated, for the most part, as the reported standard error rates are in most cases small or relatively small in comparison to the differences in performance of the four systems.

[15] For all figures, we used the spline interpolater that is built into the .NET Chart Control in order to fill in the values not gathered through our simulations. The default line tension value (0.5) was used for all figures.

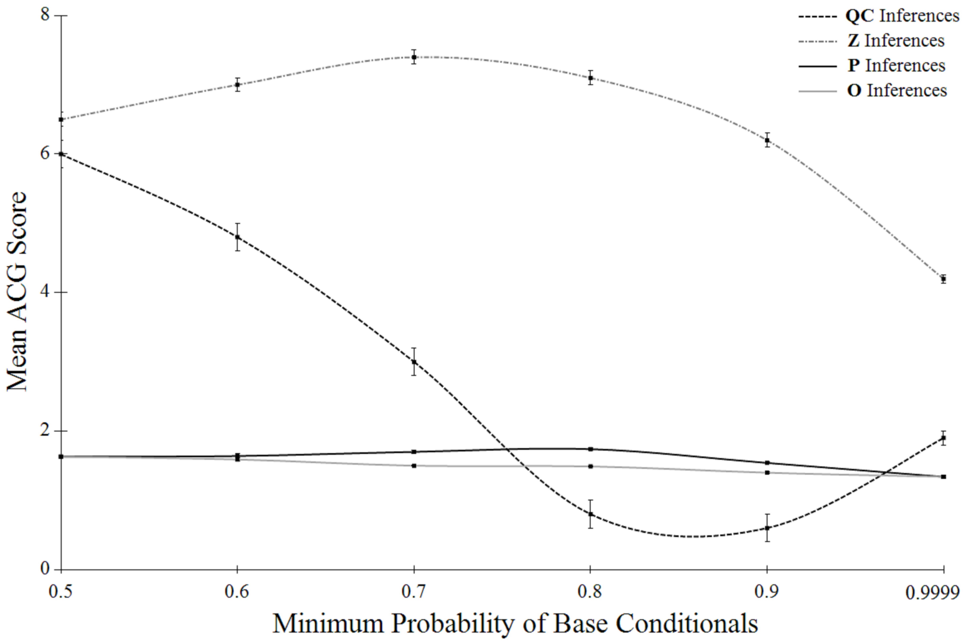


Figure 1. Mean total ACG scores per simulation, with four base conditionals

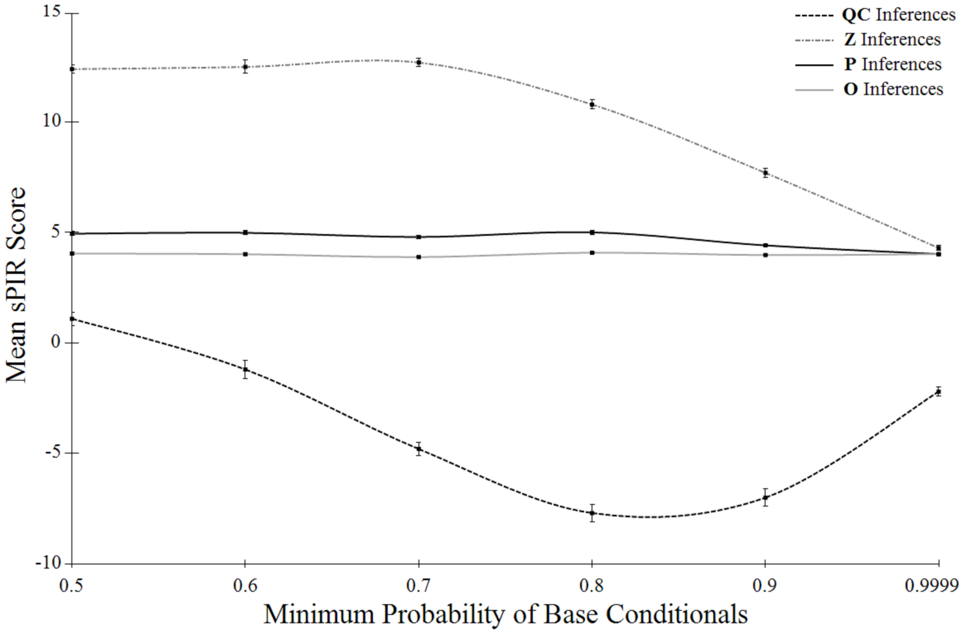


Figure 2. Mean total sPIR scores per simulation, with four base conditionals

QC−**Z** inferences) are of rather poor quality. However, the scoring rules applied within [17] were quite tough, and perhaps biased against system **QC**. In order to address this concern, we now present results that illustrate the performance of system **QC** by the light of the EU scoring rule.

The EU scores for the four LP systems for varied values of b (the number of base conditionals) and s (the minimum probability for base conditionals) are summarized within Table 1, which reports respective mean total EU scores for a single simulation, along with standard error rates for the mean values "±S.E.M."[16]

# of Base Conditionals	Their Minimum Probability	Mean EU Scores (±S.E.M.)			
		O Inferences	**P** Inferences	**Z** Inferences	**QC** Inferences
One	0.5	0.154 ±0.006	0.154 ±0.006	0.71 ±0.02	0.77 ±0.02
	0.6	0.175 ±0.008	0.175 ±0.008	0.76 ±0.02	0.80 ±0.02
	0.7	0.191 ±0.008	0.191 ±0.008	0.81 ±0.02	0.82 ±0.02
	0.8	0.180 ±0.009	0.180 ±0.009	0.79 ±0.02	0.79 ±0.02
	0.9	0.178 ±0.005	0.178 ±0.005	0.73 ±0.02	0.73 ±0.02
	0.9999	0.136 ±0.003	0.136 ±0.003	0.230 ±0.008	0.230 ±0.008
Two	0.5	0.33 ±0.01	0.37 ±0.01	1.52 ±0.03	1.71 ±0.03
	0.6	0.37 ±0.01	0.41 ±0.02	1.66 ±0.03	1.76 ±0.03
	0.7	0.36 ±0.01	0.39 ±0.01	1.65 ±0.03	1.69 ±0.03
	0.8	0.38 ±0.01	0.41 ±0.02	1.67 ±0.03	1.65 ±0.03
	0.9	0.35 ±0.009	0.37 ±0.01	1.38 ±0.03	1.36 ±0.03
	0.9999	0.27 ±0.006	0.27 ±0.006	0.44 ±0.01	0.44 ±0.01
Three	0.5	0.48 ±0.01	0.54 ±0.01	2.34 ±0.03	2.53 ±0.04
	0.6	0.49 ±0.01	0.61 ±0.02	2.47 ±0.04	2.64 ±0.04
	0.7	0.52 ±0.01	0.61 ±0.02	2.53 ±0.04	2.57 ±0.04
	0.8	0.53 ±0.01	0.64 ±0.02	2.45 ±0.04	2.42 ±0.04
	0.9	0.52 ±0.01	0.59 ±0.02	1.95 ±0.04	1.92 ±0.04
	0.9999	0.403 ±0.009	0.403 ±0.009	0.64 ±0.02	0.64 ±0.02
Four	0.5	0.69 ±0.02	0.90 ±0.03	3.15 ±0.05	3.56 ±0.05
	0.6	0.69 ±0.01	0.92 ±0.03	3.44 ±0.05	3.56 ±0.05
	0.7	0.69 ±0.01	0.90 ±0.03	3.56 ±0.04	3.36 ±0.04
	0.8	0.74 ±0.01	0.97 ±0.03	3.25 ±0.05	3.20 ±0.05
	0.9	0.65 ±0.01	0.74 ±0.01	2.35 ±0.04	2.30 ±0.04
	0.9999	0.57 ±0.01	0.57 ±0.01	0.97 ±0.02	0.97 ±0.02

Table 1. Mean total EU scores per simulation for **O**, **P**, **Z**, and **QC** inferences

[16] We adopt the convention for reporting significant digits proposed in [3], with the following exceptions: For Table 2, we do not report standard errors smaller than 0.001, and report mean values to the third place beyond the decimal, in such cases. For Table 3, we do not report standard errors smaller than 0.0001, and report mean values to the fourth place beyond the decimal, in such cases.

Examining Table 1, we see that the EU scores for systems **Z** and **QC** invariably exceed the scores for systems **O** and **P** by a considerable margin. In cases where the minimum probability for base conditionals is low, we also see that scores for system **QC** tend to exceed the scores for system **Z**. Representative behavior of the four systems, as tracked by the EU measure, is illustrated by Figure 3, which presents the total average EU scores for the four systems, in the case where the number of base conditionals is held fixed at four, and the minimum probability for base conditionals is varied.

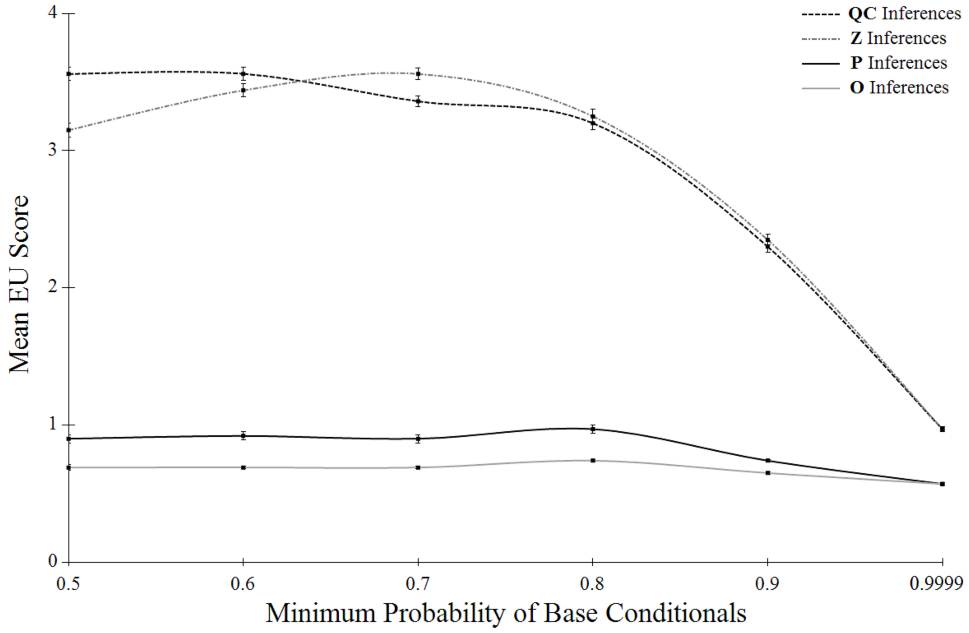
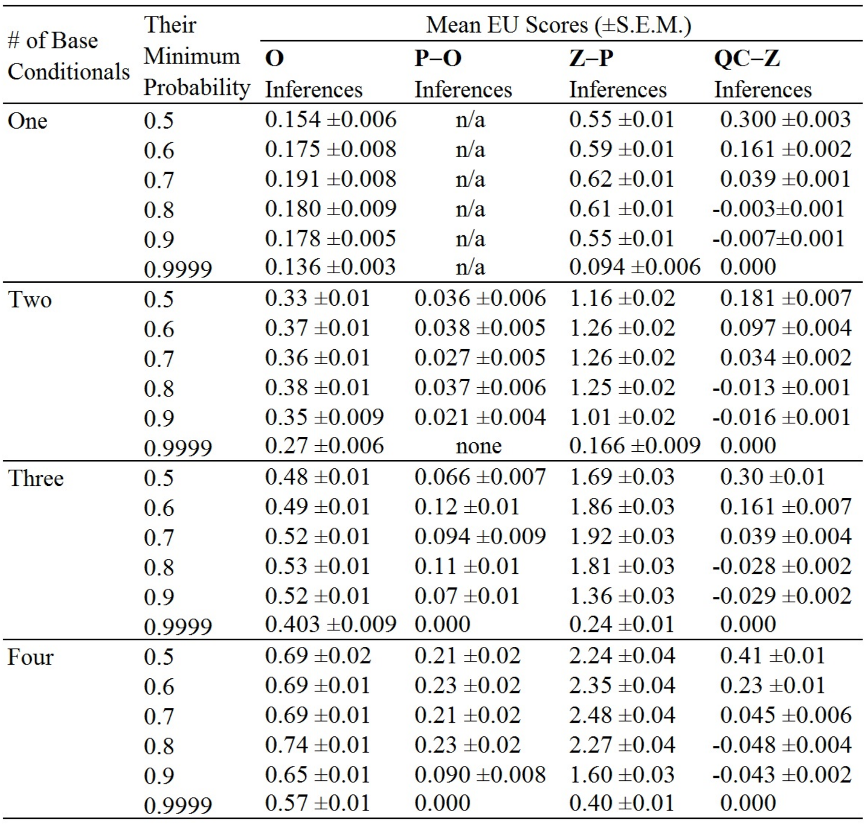


Figure 3. Mean total EU scores per simulation, with four base conditionals

While the results *appear* favorable to system **QC**, further analysis supports a different conclusion. Recalling the fact that the four systems can be ordered in terms of the number of inferences they license, it is instructive to consider the scores earned for the 'new' inferences licensed by each system as one proceeds from system **O** to system **QC**, i.e., the inferences licensed by system **O**, the inferences licensed by system **P** that are not licensed by system **O** (**P**−**O**), etc. The results are recorded on Table 2.

Table 2 presents a more balanced picture of the performance of the four systems, since it records the score earned by each system for the new inferences that the system adds to the inferences that are already licensed by its more conservative predecessor. Figure 4 illustrates the general pattern, and



# of Base Conditionals	Their Minimum Probability	Mean EU Scores (±S.E.M.)			
		O Inferences	**P–O** Inferences	**Z–P** Inferences	**QC–Z** Inferences
One	0.5	0.154 ±0.006	n/a	0.55 ±0.01	0.300 ±0.003
	0.6	0.175 ±0.008	n/a	0.59 ±0.01	0.161 ±0.002
	0.7	0.191 ±0.008	n/a	0.62 ±0.01	0.039 ±0.001
	0.8	0.180 ±0.009	n/a	0.61 ±0.01	-0.003±0.001
	0.9	0.178 ±0.005	n/a	0.55 ±0.01	-0.007±0.001
	0.9999	0.136 ±0.003	n/a	0.094 ±0.006	0.000
Two	0.5	0.33 ±0.01	0.036 ±0.006	1.16 ±0.02	0.181 ±0.007
	0.6	0.37 ±0.01	0.038 ±0.005	1.26 ±0.02	0.097 ±0.004
	0.7	0.36 ±0.01	0.027 ±0.005	1.26 ±0.02	0.034 ±0.002
	0.8	0.38 ±0.01	0.037 ±0.006	1.25 ±0.02	-0.013 ±0.001
	0.9	0.35 ±0.009	0.021 ±0.004	1.01 ±0.02	-0.016 ±0.001
	0.9999	0.27 ±0.006	none	0.166 ±0.009	0.000
Three	0.5	0.48 ±0.01	0.066 ±0.007	1.69 ±0.03	0.30 ±0.01
	0.6	0.49 ±0.01	0.12 ±0.01	1.86 ±0.03	0.161 ±0.007
	0.7	0.52 ±0.01	0.094 ±0.009	1.92 ±0.03	0.039 ±0.004
	0.8	0.53 ±0.01	0.11 ±0.01	1.81 ±0.03	-0.028 ±0.002
	0.9	0.52 ±0.01	0.07 ±0.01	1.36 ±0.03	-0.029 ±0.002
	0.9999	0.403 ±0.009	0.000	0.24 ±0.01	0.000
Four	0.5	0.69 ±0.02	0.21 ±0.02	2.24 ±0.04	0.41 ±0.01
	0.6	0.69 ±0.01	0.23 ±0.02	2.35 ±0.04	0.23 ±0.01
	0.7	0.69 ±0.01	0.21 ±0.02	2.48 ±0.04	0.045 ±0.006
	0.8	0.74 ±0.01	0.23 ±0.02	2.27 ±0.04	-0.048 ±0.004
	0.9	0.65 ±0.01	0.090 ±0.008	1.60 ±0.03	-0.043 ±0.002
	0.9999	0.57 ±0.01	0.000	0.40 ±0.01	0.000

Table 2. Mean total EU scores per simulation for **O**, **P−O**, **Z−P**, and **QC−Z** inferences

presents the data from Table 2 for the case where the number of base conditionals is fixed at four, and the minimum probability for base conditionals is varied.

While system **QC** is the only system that tends to permit *new* inferences that earn negative scores, in some cases, it appears, on balance, that **QC−Z** inferences tend to earn positive scores. However, a closer analysis reveals that **QC−Z** inferences are of low quality. To see why this is so, consider Table 3, which displays the average score earned *per inference*, among inferences in the categories **O**, **P−O**, **Z−P**, and **QC−Z**.[17]

In presenting Table 3, our concern is to offer a reasonable assessment

[17] Note that the reported average EU scores for **P−O** inferences from 3 and 4 base conditionals, and minimum probability 0.9999, are each based on a single inference, and are thus not good measures of the typical EU score earned by **P−O** inferences in such cases.

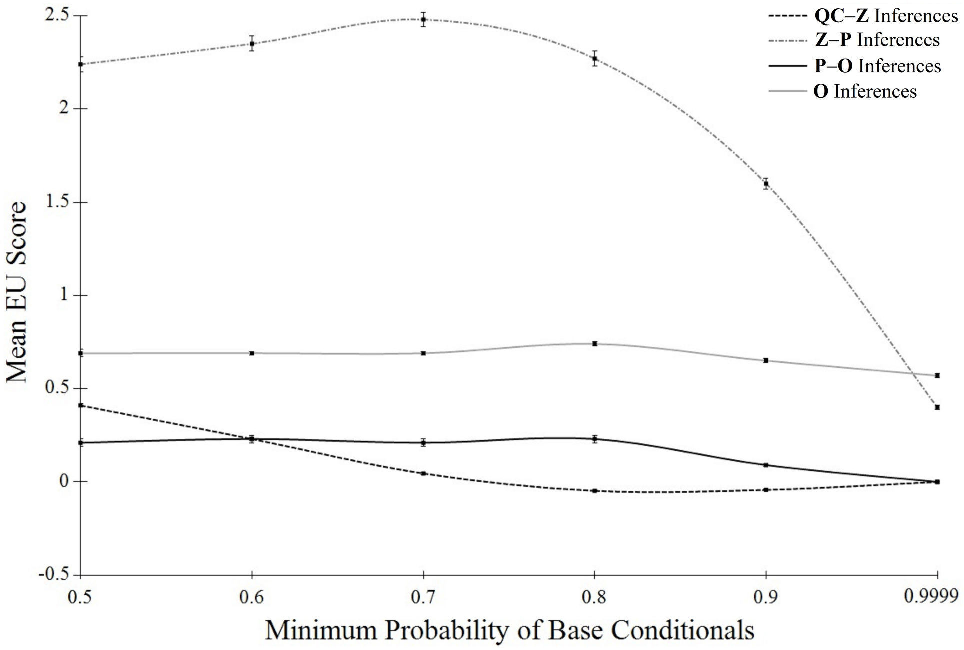
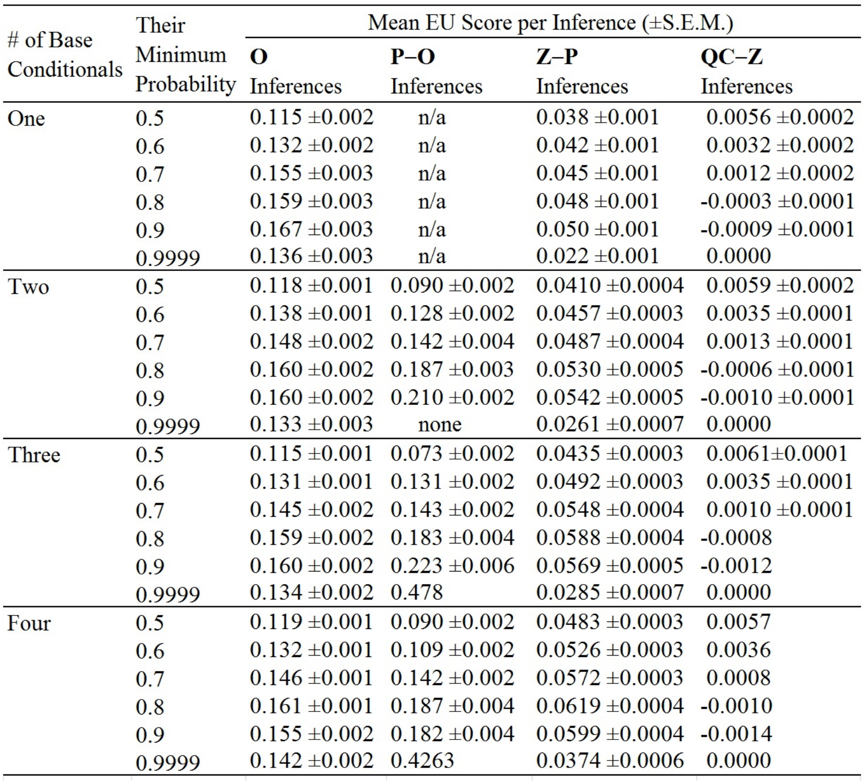


Figure 4. Mean total EU scores per simulation, with four base conditionals

of the value of the conclusions drawn by system **QC**. What Table 3 illustrates is that the inferences made by system **QC** tend to generate some value (according to the EU measure) in many situations. However, a reasonable assessment of the value of the conclusions drawn by system **QC** recommends that we consider the quality of **QC**−**Z** inferences in comparison with non-evidential methods of probability judgment, in particular, methods of assigning lower probability bounds to the elements of L_4 without exploiting the information that was supplied the four LP systems (in the form of base conditionals).

We considered three non-evidential methods of setting lower probability bounds. The first non-evidential method that we considered (ne-method one) assigned the lower bound 1/2 to each of the conditionals in L_4. The next non-evidential method that we considered (ne-method two) was slightly more subtle, and assigned the lower bound 10/29 to each conditional in L_4, based on the fact that the average probability of the elements of L_4 within our simulations is 10/29. The final non-evidential method that we considered (ne-method three) was the most subtle and assigned the lower bound 1/2 to conditionals with a single conjunct in their consequent, 1/4 to conditionals

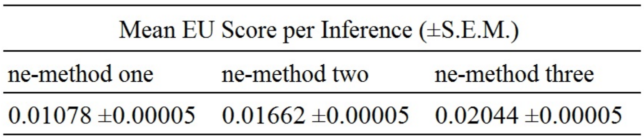


# of Base Conditionals	Their Minimum Probability	Mean EU Score per Inference (±S.E.M.)			
		O Inferences	**P–O** Inferences	**Z–P** Inferences	**QC–Z** Inferences
One	0.5	0.115 ±0.002	n/a	0.038 ±0.001	0.0056 ±0.0002
	0.6	0.132 ±0.002	n/a	0.042 ±0.001	0.0032 ±0.0002
	0.7	0.155 ±0.003	n/a	0.045 ±0.001	0.0012 ±0.0002
	0.8	0.159 ±0.003	n/a	0.048 ±0.001	-0.0003 ±0.0001
	0.9	0.167 ±0.003	n/a	0.050 ±0.001	-0.0009 ±0.0001
	0.9999	0.136 ±0.003	n/a	0.022 ±0.001	0.0000
Two	0.5	0.118 ±0.001	0.090 ±0.002	0.0410 ±0.0004	0.0059 ±0.0002
	0.6	0.138 ±0.001	0.128 ±0.002	0.0457 ±0.0003	0.0035 ±0.0001
	0.7	0.148 ±0.002	0.142 ±0.004	0.0487 ±0.0004	0.0013 ±0.0001
	0.8	0.160 ±0.002	0.187 ±0.003	0.0530 ±0.0005	-0.0006 ±0.0001
	0.9	0.160 ±0.002	0.210 ±0.002	0.0542 ±0.0005	-0.0010 ±0.0001
	0.9999	0.133 ±0.003	none	0.0261 ±0.0007	0.0000
Three	0.5	0.115 ±0.001	0.073 ±0.002	0.0435 ±0.0003	0.0061±0.0001
	0.6	0.131 ±0.001	0.131 ±0.002	0.0492 ±0.0003	0.0035 ±0.0001
	0.7	0.145 ±0.002	0.143 ±0.002	0.0548 ±0.0004	0.0010 ±0.0001
	0.8	0.159 ±0.002	0.183 ±0.004	0.0588 ±0.0004	-0.0008
	0.9	0.160 ±0.002	0.223 ±0.006	0.0569 ±0.0005	-0.0012
	0.9999	0.134 ±0.002	0.478	0.0285 ±0.0007	0.0000
Four	0.5	0.119 ±0.001	0.090 ±0.002	0.0483 ±0.0003	0.0057
	0.6	0.132 ±0.001	0.109 ±0.002	0.0526 ±0.0003	0.0036
	0.7	0.146 ±0.001	0.142 ±0.002	0.0572 ±0.0003	0.0008
	0.8	0.161 ±0.001	0.187 ±0.004	0.0619 ±0.0004	-0.0010
	0.9	0.155 ±0.002	0.182 ±0.004	0.0599 ±0.0004	-0.0014
	0.9999	0.142 ±0.002	0.4263	0.0374 ±0.0006	0.0000

Table 3. EU scores per inference

with pair of conjuncts in their consequent, and 1/8 to conditionals with three conjuncts in their consequent. Within our simulations, these values (1/2, 1/4, and 1/8) are the average probabilities for conditionals with the corresponding number of conjuncts in their consequents. The scores received by the three non-evidential methods are, of course, independent of variations in the number of base conditionals provided to the LP systems (and of their associated minimum probabilities). The average EU score earned per inference for the three non-evidential methods over one thousand simulations are presented in Table 4.

As one can see from the data represented in Table 4, the average EU score earned per inference by each of the non-evidential methods exceeds the scores earned by **QC**$-$**Z** inferences (but not those earned by **Z**$-$**P** inferences). We take this to show: (i) that the positive EU scores earned by **QC**$-$**Z** inferences result from the tendency of the EU measure to reward contentful judgments, in general, and (ii) that the positive scores earned by **QC**$-$**Z**



Mean EU Score per Inference (±S.E.M.)		
ne-method one	ne-method two	ne-method three
0.01078 ±0.00005	0.01662 ±0.00005	0.02044 ±0.00005

Table 4. EU scores per inference for non-evidential methods

inferences do not reflect a significant capacity of **QC**−**Z** inferences to exploit information about an environment to draw reasonable conclusions about that environment.

6. Conclusion

Our concern in the present article was to evaluate four well known LP reasoning systems in terms of their tendency to draw true informative conclusions and avoid drawing false or uninformative conclusions. It is known that inference via systems **O** and **P** is *correct* with respect to strict premise probability preservation, and improbability-sum preservation, respectively. Due to such validity results, it is clear that it is reasonable to accept the conclusions that are licensed by systems **O** and **P** (coupled with the lower probability bounds that one may validly infer via those systems). So we think that the central issue is whether it is reasonable to go beyond inference by systems **O** and **P**, and reason by system **Z** or **QC** (and assign corresponding lower probability bounds upon the default assumption that the improbability of an inferred conditional is not greater than the sum of the improbabilities of the premises required for the inference). Within [17], it is argued that it is reasonable to reason in accordance with system **Z** (and assign corresponding lower probability bounds), but it is not reasonable to reason in accordance with system **QC**. As grounds for this conclusion, we appealed to the fact that (on average) the ACG and sPIR scores earned for **QC**−**Z** inferences are negative.

Since it is possible to raise concerns about the adequacy of the ACG and sPIR scoring measures, we introduced the EU measure which precisely reflects the expected value of rational betting in accordance with inferred lower probability bounds. As with the results presented in [17], the results presented here support the conclusion that system **Z** offers the best balance of error avoidance and inferential power. In contrast to the results presented in [17], we observe that the additional risks taken by system **QC** are not severely penalized by the EU scoring measure (as reflected in the fact that **QC**−**Z** inferences tend to receive positive EU scores in many cases). Nevertheless, the results presented in Table 4 show that the tendency of **QC**−**Z** inferences to earn positive EU scores (in some cases) does not derive from the *quality* of **QC**−**Z** inferences, but only from the tendency of the EU measure to reward contentful judgments, in general. Indeed, unlike **Z**−**P** inferences, **QC**−**Z** inferences tend to earn average EU scores that are lower than various non-evidential methods of setting lower probability bounds. Nevertheless, since we know that many **QC**−**Z** inferences are both false and highly uninformative, the fact that **QC**−**Z** inferences received slightly positive EU scores on average tells us that many other **QC**−**Z** inferences achieved significant positive EU scores. This consideration supports the following conclusion: If there are LP reasoning systems whose typical EU scores are greater than that of system **Z**, then these systems are presumably to be found in the class of reasoning systems which are stronger than **Z** but weaker than **QC**.

Acknowledgements. Work for this paper was supported by DFG-Grant SCHU1566/5-1 as part of the LogiCCC project *The Logic of Causal and Probabilistic Reasoning* (LcpR), and by DFG-Grant SCHU1566/9-1 as part of the SPP 1516 priority program *New Frameworks of Rationality.* For valuable comments, we are indebted to David Makinson, Hannes Leitgeb, James Hawthorne, Angelo Gilio, Donald Bamer, Gerhard Brewka, Wolfgang Spohn, Gernot Kleiter, Matthias Unterhuber, and in memoriam to Ernest Adams.

References

[1] Adams, Ernest W., *The Logic of Conditionals*, Dordrecht, Reidel, 1975.

[2] Adams, Ernest W., 'A Note on Comparing Probabilistic and Modal Logics of Conditionals', *Theoria*, 43 (1977), 186–194.

[3] Bindel, David, and Jonathan Goodman, *Principles of Scientific Computing*, Manuscript, 2009.

[4] Elga, Adam, 'Subjective Probabilities Should be Sharp', *Philosophers' Imprint*, 10 (2010), 1–11.

[5] Goldszmidt, Moises, and Judea Pearl, 'Qualitative Probabilities for Default Reasoning, Belief Revision and Causal Modeling', *Artificial Intelligence*, 84 (1996), 57–112.

[6] Halpern, Joseph Y., *Reasoning about Uncertainty*, MIT Press, Cambridge, Massachusetts, 2003.

[7] Hawthorne, James, 'On the Logic of Non-Monotonic Conditionals and Conditional Probabilities', *Journal of Philosophical Logic*, 25 (1996), 185–218.

[8] Hawthorne, James, and David Makinson, 'The Quantitative / Qualitative Watershed for Rules of Uncertain Inference', *Studia Logica*, 86 (2007), 247–297.

[9] Knuth, Donald E., *The Art of Computer Programming, volume 2: Seminumerical Algorithms*, Addison-Wesley, Reading, MA, 1981.

[10] Kraus, Sarit, Daniel Lehmann, and Menachem Magidor, 'Nonmonotonic Reasoning, Preferential Models and Cumulative Logics', *Artificial Intelligence*, 44 (1990), 167–207.

[11] Lehmann, Daniel, and Menachem Magidor, 'What Does a Conditional Knowledge Base Entail?', *Artificial Intelligence*, 55 (1992), 1–60.

[12] Leitgeb, Hannes, *Inference on the Low Level*, Kluwer, Dordrecht, 2004.

[13] Paris, J. B., and R. Simmonds, 'O is not Enough', *Review of Symbolic Logic*, 2 (2009), 298–309.

[14] Pearl, Judea, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, Santa Mateo, California, 1988.

[15] Pearl, Judea, 'System Z', in *Proceedings of Theoretical Aspects of Reasoning about Knowledge*, Santa Mateo, California, 1990, pp. 21–135.

[16] Schurz, Gerhard, 'Probabilistic Default Reasoning Based on Relevance and Irrelevance Assumptions', in Dov M. Gabbay, Rudolf Kruse, Andreas Nonnengart, and Hans Jurgen Ohlbach, (eds.), *Qualitative and Quantitative Practical Reasoning*, no. 1244 in LNAI, Springer, Berlin, 1997, pp. 536–553.

[17] Schurz, Gerhard, and Paul D. Thorn, 'Reward versus Risk in Uncertain Inference: Theorems and Simulations', *Review of Symbolic Logic*, 5 (2012), 574–612.

[18] Segerberg, Krister, 'Notes on Conditional Logic', *Studia Logica*, 48 (1989), 157–168.

[19] Spohn, Wolfgang, *The Laws of Belief: Ranking Theory and Its Philosophical Applications*, Oxford University Press, 2012.

[20] Suppes, Patrick, 'Probabilistic Inference and the Concept of Total Evidence', in Jaakko Hintikka, and Patrick Suppes, (eds.), *Aspects of Inductive Logic*, North-Holland Publ. Comp., Amsterdam, 1966, pp. 49–65.

[21] Unterhuber, Matthias, *Possible Worlds Semantics for Indicative and Counterfactual Conditionals? A Formal-Philosophical Inquiry into Chellas-Segerberg Semantics*, Ontos Verlag (Logos Series), Frankfurt, 2013.

[22] Unterhuber, Matthias, and Gerhard Schurz, 'Completeness and Correspondence in Chellas-Segerberg Semantics', *Studia Logica*, this volume (forthcoming).

[23] Williamson, Jon, 'Motivating Objective Bayesianism: from Empirical Constraints to Objective Probabilities', in William Leonard Harper, and Gregory Wheeler, (eds.), *Probability and Inference: Essays in Honor of Henry E. Kyburg Jr.*, College Publications, London, 2007, pp. 155–183.

Paul D. Thorn & Gerhard Schurz
Duesseldorf Center for Logic and Philosophy of Science
Department of Philosophy
University of Duesseldorf
Universitaetsstr. 1
40215 Duesseldorf, Germany
thorn@phil.hhu.de, gerhard.schurz@phil.hhu.de